# TV IN THE AGE OF THE INTERNET:

# INFORMATION QUALITY OF SCIENCE

# FICTION TV FANSITES

Jonathan D. Warren

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements for the degree

Doctor of Philosophy in the School of Library and Information Science,

Indiana University

February 2011

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____

John C. Paolillo, Ph.D.

_____

Howard Rosenbaum, Ph.D.

_____

Ying Ding, Ph.D.

_____

Karen Kafadar, Ph.D.

November 18, 2010

# Acknowledgments

Heartfelt thanks to the following people, for their support throughout this work: *my advisors* (John Paolillo, Howard Rosenbaum, Ying Ding, and Karen Kafadar), *administrators and staff at SLIS* (Blaise Cronin, Elin Jacob, Arlene Merkel, Rhonda Spencer, Mary Kennedy, Sarah Burton, and Jill Clancy), *administrators and editors of the sites studied*, *friends* (Virginia Dearborn, Julia Haskin, Debbie Light, Lai Ma, and Dustin Moore), and *family* (Cindy Mitchell, Mark Pasquariello, Jeanne and Bill Mitchell, Jennifer Mitchell, Dan Mitchell, and Lee Mitchell).

Jonathan D. Warren

# TV IN THE AGE OF THE INTERNET: INFORMATION QUALITY OF SCIENCE FICTION TV FANSITES

Communally created Web 2.0 content on the Internet has begun to compete with information provided by traditional gatekeeper institutions, such as academic journals, medical professionals, and large corporations. On the one hand, such gatekeepers need to understand the nature of this competition, as well as to try to ensure that the general public are not endangered by poor quality information. On the other hand, advocates of free and universal access to basic social services have argued that communal efforts can provide as good or better-quality versions of commonly needed resources. This dissertation arises from these needs to understand the nature and quality of information being produced on such websites. Website-oriented information quality (IQ) literature spans at least 15 different academic fields, a survey of which identified two types of IQ: perceptual and artifactual fitness-related, and representational accuracy and completeness-related. The current project studied websites in terms of all of these, except perceptual fitness.

This study may be the only of its kind to have targeted fansites: websites made by fans of a mass media franchise. Despite the Internet's becoming a primary means by which millions of people consume and co-produce their entertainment, little academic attention has been paid to the IQ of sites about the mass media. For this study, the four central non-studio-affiliated sites about a highly popular and fan-engaging science fiction television franchise, Stargate, were chosen, and their IQ examined across sites having different

sizes as well as editorial and business models. As exhaustive of samples as possible were collected from each site. Based on 21 relevant variables from the IQ literature, four qualitative and 17 exploratory statistical analyses were conducted. Key findings include: five possibly new IQ criteria; smaller sites concerned more with pleasing connoisseuring fans than the general public; larger sites being targeted towards older users; professional editors serving their own interests more than users'; wikis' greater user freedom attracting more invested and balanced writers; for-profit sites being more imposing upon, and less protecting of, users than non-profit sites; and the emergence of common writing styles, themes, data fields, advertisement types, linking strategies, and page types.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Background

Communally created, Web 2.0 content on the Internet is becoming an important source of information for many people – for undergraduate term papers, general medical and philosophical questions, leisure activities, product choices, and so on. For everyday information needs that do not require perfect quality or a diploma, this makes a good deal of pragmatic sense. Marketing literature is often biased in favor of a particular company, and academic or other expert opinions are often hidden behind gatekeeper institutions charging expensive consultation, subscription, or tuition fees.

On the current version of the Internet, communal sharing of information about a particular topic or phenomenon often manifests as a "fansite." In general, a person can be said to be a "fan" of any object, person, topic, or event about which they have a strong personal interest, attachment, or identification, though that interest is not usually so fanatical as to involve violating basic social norms of behavior (Thorne and Bruner, 2006;

Wann, 1997). Fan behavior and artifacts are collectively referred to as "fandom" (Fiske and Lewis, 1992). The phenomenon of fansites descends from print "fanzines," which, before the Internet, were magazines produced and circulated by/for fans of a particular topic (Moskowitz, 1990). Much like fanzines, current fansites are perhaps most characterized by their rejection of "official" (i.e., institutional or corporate) websites, on the grounds that the self-serving agendas of the organizations behind such sites can bias the sites' content and policies (AmberlightHCI, 2008). However, current fansites also facilitate mass communication and collaboration in ways not feasible in print, making them more resemblant of virtual communities, knowledge bases, or digital repositories than of magazines.

Following in the mold of traditional mass media studies, cultural studies research has routinely tried to classify online fan activities into mutually exclusive production vs. consumption behaviors. However, many scholars, from a variety of social science fields, have begun to argue how problematic this can be for fan activity on the Web (McKee, 2004). For example, critical theorists have emphasized that, by writing their own descriptions and critiques of phenomena, users can collectively resist the messages that institutions would impose upon them (Costello and Moore, 2007). Communication scholars have noted that computer-mediated communication (CMC) technologies on the Internet offer users and professional producers unprecedented access to each other (Theberge, 2005). The emerging consensus seems to be that, rather than a stark dichotomy, consumption on the Web is often active and productive. Fans produce both *interpretive descriptions* (e.g., plot summaries and critiques, annotated transcripts, and glossaries), for the purpose of helping both themselves and others find and understand the original content, as well as *original extensions* (e.g., fan fiction, original screenplays, and artwork) for the purpose of developing the

studio's content in new directions.

Fansites also manifest a spectrum of editorial models. At one extreme, sites such as Wikipedia.org, which is an open-source community project, allow almost anyone to contribute almost anything, with community members themselves policing content for compliance with communally agreed-upon rules of behavior and standards of quality (Wikipedia, 2009a). At the other extreme, sites such as IMDb.com, which is owned by Amazon.com, devote a large portion of their paid staff to editing user contributions (IMDb, 2009e). The reason for this clear divide ultimately comes from ideological differences in how open-source communities and corporations view copyright (Elliott, 2001; Moore, 2001; Yamamoto, 2000; Tussey, 2000). In the open-source case, users typically retain copyright on their contributions, allowing them to own and protect their intellectual property (Wikipedia, 2009h). In the corporate case, users typically forfeit their copyright to the company, which may do with the content as it pleases (IMDb, 2009d). As has long been the case with content submitted to newspapers, user reward in the corporate case usually derives from the public recognition of having something similar to one's writing appear on a popular (web)page, hopefully with one's name attached. Furthermore, these editorial differences mirror somewhat the sites' choices of business models, which will be discussed for the specific sites under study in the next section.

This dissertation argues that, since many users of the Internet are coming to use communally created fansites *rather than* academic or industrial literature for information about the world, and since much of the content on these fansites is both descriptive and analytical (i.e., quasi-scholarly) in nature, the structure and quality of those communal information sources need to be assessed, in comparison to traditional scholarly materials. This docu-

ment constitutes a mixed-methods study of the information quality of several large fansites on a popular cultural phenomenon, which employ a variety of editorial and business models.

## 1.2 Context

Often called "Hollywood North" (Gasher, 2002), Vancouver, British Columbia, Canada is the third most productive city for the world's film industries, behind Los Angeles and New York City, and is the second most productive for television (BCFilmCommission, 2008; HollywoodNorthFilmNet, 2008). Vancouver has been the home of over 200 distinct science fiction television series, including United States successes such as MacGyver and The X-Files, and has helped give rise to such internationally famous science fiction actors as Dan Aykroyd (from Ghostbusters), James Doohan and William Shatner (Star Trek), and Keanu Reeves (The Matrix).

The social phenomenon under study for this project will be fansites pertaining to the popular Stargate science fiction television and film franchise – including the SG-1, Atlantis, and Universe series, as well as several feature and direct-to-DVD films – which has been a centerpiece of the Vancouver science fiction television and film community since 1997. This phenomenon was chosen for several reasons. First, due to its popularity, a considerable amount of both academic and amateur literature have been generated describing and analyzing it. However, the franchise is not so old that academic literature about it has become tedious or nostalgic. Yet, the franchise is old enough that a handful of fansites have become dominant, and those sites each have chosen either a more editorially or community

controlled editorial model. Finally, both the historical socio-cultural contexts and narrative techniques of international co-productions in Vancouver (including Stargate), as well as of Anglo-American science fiction more generally, have been academically well-studied. The cultural magnitude of the topic, the public availability of fansite data, as well as the researcher's own experience with the franchise and these fansites, make it a viable and generalizable context in which to study the information quality of fansites that use different editorial models.

The specific sites under study will be the Stargate-related pages of the following websites: Wikipedia.org, IMDb.com, Wikia.com, and GateWorld.net. This set of sites is not at all arbitrary, as explained in §3.2. Briefly, these four sites appear to represent the core large fansites about this media franchise. This was determined by consulting the webpages of science fiction associations, by searching the Web with several popular search engines and comparing the rankings of the results, and by consulting the Open Directory Project (DMOZ.org). Also, besides being the highest-ranked, the Stargate pages on these four sites frequently link to one of the other three sites. By contrast, small fansites, of which there are many, were not considered, because this study hopes to generalize to other large Web 2.0 fansites. Furthermore, this group of four sites well-represents a variety of editorial and business models. Finally, because each of these sites is large, a relatively small group of them is required, in order to insure the project's feasibility. Consequently, Human Subjects Committee approval to study the public digital artifacts of these and similar sites has been obtained.

The organizational nature of these websites is generally as follows. Wikipedia and IMDb are large communally and editorially controlled sites, respectively, where fans con-

tribute information about much more than just Stargate or Vancouver (Wikipedia, 2009e; IMDb, 2009i). Wikia is a private company that shares founding/charter members (Jimmy Wales and Angela Beesley), as well as the MediaWiki software, with the Wikimedia Foundation, which is the non-profit organization behind Wikipedia. A "wiki farming" company (Pink, 2005), funded by advertisement revenue and venture capital investments (Hinman, 2006; Primack, 2007), Wikia selectively assimilates large wiki communities from elsewhere on the Web, raising those communities' search engine and directory rankings (Wikia's Alexa rank is 203, as of June 2010, a high rank), while maintaining the Wikipedia-like policies of free editing and open-source content licenses (Wikia, 2009a, 2009e). Finally, GateWorld.net is perhaps the most dominant standalone fansite about Stargate on the Web. Run by a small group of dedicated volunteers, who edit all contributed content and have befriended the franchise's producers and cast, the site attempts to be the most timely and authoritative word on anything to do with Stargate. None of these sites are sponsored by either Stargate's parent companies or the international television networks where it is aired (i.e., MGM, the SciFi/Syfy Channel, Sky One in the UK, etc.), and this research project was conducted without any conflicts of interest.

## 1.3 Approach

The current project is information scientific. Focus will be given to features and themes in the webpages themselves, as an investigation into the nature of information that occurs in this context. The value of this is in providing an initial characterization of the information available on this type of site, which can aid both future research and organizational

utilization of this type of information. It should be noted that this phenomenon could be approached from a variety of different research/disciplinary perspectives. Mass media scholars could study how the sites represent the Vancouver film industry; narrative studies scholars could compare the writings of fans against the literature on the narrative techniques used in Stargate and other sci-fi franchises; business researchers might analyze the sites' marketing techniques; and fandom researchers could study how the sites facilitate fan events and discussions, as well as how they allow non-celebrities to interact with famous or otherwise powerful members of the mass media. Some of these could be considered contextually relevant measures of the websites' information quality, and some could not. Though the current study will identify specific issues that such researchers might find of interest (see §6.2), its focus will remain more interdisciplinary and meta-analytic.

The most appropriate information science literature to this project is perhaps the branch of "information quality" (IQ) literature that studies websites. In that literature, there exist two broad research programs: the representation program, and the fitness program.

Researchers doing the *representation*-based program (e.g., Frické and Fallis, 2004) study relatively objective notions of IQ – whether in some context or not – which can be investigated by expert researchers and documents alone (e.g., accuracy, completeness, and bibliometric features), often using quantitative content analytic and statistical methods to explore the structural qualities of an entire corpus (i.e., artifact collection). Outside of library and information science (LIS), philosophers (Fallis, 2008), computer scientists (Hu et al., 2007), and historians (Rosenzweig, 2006) most often take this approach. From a representation mindset, the literature primarily speaks of factors related to either *accuracy* or *completeness*. In this context, accuracy refers to the currentness, objectivity, empiricality,

etc. of an artifact. Completeness refers to an artifact's depth and breadth of descriptive and analytical coverage of some topic.

On the other hand, researchers in business (Barnes and Vidgen, 2006), informatics (Collins, 2006), and communication (Dutta-Bergman, 2004) often take a more *fitness*-based, HCI-oriented approach, conducting studies on users' perceptions of documents' credibility, trustworthiness, usability, value, or the like with respect to some goal or purpose. Such studies are more often based on questionnaires, laboratory experiments, and interviews, and are most concerned with their research leading to improved interface implementations. In addition to perceptual fitness, one can also consider the *artifactual fitness* of informational objects. By taking artifacts rather than users as the unit of analysis, one could examine how an artifact was implemented, and imagine what potentials it has to be fit for use in any given context. For example, one could evaluate the artifact's accessibility features for users (e.g., its cost, speed, language, reading level, or personalizability) or for vendors (e.g., its distributability or maintainability). LIS and medicine appear to be the two fields in which researchers, though usually not the same researchers, have significantly engaged in both of these programs. By comparison with the other fields, LIS studies are considerably more concerned with bibliometric issues, and are more often pragmatic and domain/implementation-oriented than are philosophy studies.

The current project assesses fansites in terms of both representation- and artifactual fitness-based IQ, but not perceptual fitness. The perceptual fitness program is not appropriate, because this study is seeking to academically evaluate a corpus of websites itself, not user perceptions of that corpus. Also, unlike on fansites, because empirical academic literature usually *does not* include fictional extensions of the phenomenon under study, as

one finds in original fan fiction and screenplays, only those fansite materials that *describe, interpret, or critique* the Stargate franchise will be investigated. Conveniently, on the fan-sites under study, the originative/extending fan content is separated from the descriptive and analytical content, often given separate sub-sections of the site. This suggests that the sites' users/editors find the descriptive::originative distinction to be natural as well.

Additionally, the reader should be aware that some ambiguity exists in the literature's use of the terms 'data quality' and 'information quality.' Philosophically, the complexity of this ambiguity resembles the complexity of the debate around what constitutes data vs. information, which is beyond the scope of this dissertation. However, there exist clear pragmatic affiliations between researchers who use the phrase 'data quality' and who study computer scientific issues surrounding the *internal consistency* of large enterprise databases. Internal consistency here refers to reconciling all of a database's records, so that no two records are in conflict (e.g., the record of a person's age not equaling the current date minus a record of the person's birth date). For example, the newly founded Journal of Data and Information Quality is related to Richard Wang's Total Data Quality Management program at MIT, and the associated International Conference on Information Quality. The primary scope of that journal includes: data provenance, data cleaning, data integration, record linkage, enterprise architecture deployment, data privacy and protection, ensuring data integrity over time, and the roles of these topics in an enterprise context (JDIQ, 2010). Although the majority of this systems-engineering literature is not relevant to the current project, all data quality resources that could be found were considered for this dissertation's literature review (e.g., Fisher et al., 2009; Heinrich et al., 2009; Moustakides and Verykios, 2009; Su et al., 2009), and a few contained IQ variables and measures worthy of

inclusion. For example, studies of page sectioning as well as themes, page types, and data fields shared across sites and their sections could be considered studies of internal consistency. Internal consistency IQ will not be given its own section of this study, because, in this context, the relevant variables more naturally fit within representational IQ.

## 1.4 Research questions

Like most IQ literature, the current study is a combination of evaluating, in the current context, IQ criteria identified by previous studies as being potentially relevant for any website, plus inductively identifying criteria that are important for this context. All of the criteria inductively identified throughout this study emerged through iterative hermeneutic interpretations and content analyses of webpages, during the course of evaluating general IQ criteria prescribed by the literature.

Hence, the three relevant research programs identified in the previous sub-section correspond to the current study's three general research questions. Literature on each research program identified general criteria for evaluating IQ in that sense. At the end of each section of the literature review chapter (see §§2.3.5, 2.4.10, and 2.5.7), these criteria have been framed as questions, and are the senses in which the three general research questions are answered by this dissertation. For this reason, those questions will be called "research sub-questions" throughout this work. Here are the research questions and sub-questions presented together:

First, how artifactually fit-for-use are these websites for the general public? Relevant criteria from the IQ literature for answering this question in this context include: how

simple are the sites' conceptual architectures, what forms of media are available on the sites, what are the occurrences of broken hyperlinks on the sites' pages, to what extent do the sites' pages comply with international markup language standards and accessibility guidelines, what are the formulaic reading levels of the pages' texts, and what are the visual stylistic techniques used by each site?

Second, regarding these sites' capability to represent some phenomenon, what indicators of accuracy are present on each site? Relevant criteria from the literature include: how timely are the webpages; what types of sources, authors, and evidence are cited on pages; what forms of original research are evident on pages; do normative descriptions, interpretations, and data fields exist across the sites on the same topics; what can be inferred from the sites about their vendors and target audiences; how do ratings given by fans compare with those by editors, and do either correlate with other criteria; how do inlink and PageRank counts to individual pages compare with each other, and do either correlate with other criteria; and to where are hyperlinks going out of these sites usually destined, and from where do inlinks usually originate?

Third, regarding these sites' capability to represent some phenomenon, what indicators of completeness are present on each site? Relevant criteria from the literature include: what are the details of the sites' ownership, funding, affiliations, primary purposes/interests, organizational types, and locations; why do the sites cite each other; to what extent are pages with lengthy texts or many authors more thorough, inconcise, or ineloquent; and what differences in copyright and disclaimer statements exist between the sites?

An ambitious project, 21 separate qualitative and quantitative analyses were conducted to address these issues. The result is a large-scale, multi-faceted, and quite objective com-

parison of these sites' aptitudes for academic information quality.

## 1.5 Outline of the dissertation

The dissertation is structured as follows. Chapter 2 critically surveys the relevant literature, identifying the research sub-questions mentioned in the previous section. Chapter 3 surveys the rationale for collecting certain data in this context, the means by which it was collected, the ways in which information quality variables were operationalized, and the analytical methods employed by both the literature and this study. Chapter 4 presents the results of the various analyses, in the same order as the issues are presented in the literature review chapter. Chapter 5 offers discussion of the various findings and their implications for LIS and related fields, grouping them according to whether they relate to either: site size, editorial model, business model and mission, individual sites, or all of the sites. Finally, chapter 6 summarizes the dissertation's main results, contributions, and limitations, closing with thoughts on worthwhile future research in this area.

# Chapter 2

# Literature

## 2.1 Introduction

As with many abstract terms in LIS, no clear consensus exists on a definition of information quality (IQ). Hundreds of authors have used whatever criteria they prefer, or whichever arguably seems most appropriate to their research context and field. In total, 65 unique *categories* of IQ-related variables were encountered in the literature; several hundred unique variables were encountered, often the same or similar terms differently worded. In accordance with the research project described in the Introduction (§1.3), the following review covers only those variables that take digital artifacts (usually webpages) as their unit of analysis, not user perceptions of artifact quality. This narrows the number of variable categories to 37. Henceforth, these categories of variables are merely referred to as "variables."

IQ studies nearly always study multiple IQ factors, often more than 10. As with the present project, those IQ studies that regard new or underexplored phenomena often include more factors, and survey them mostly descriptively in some research context, rather than

inferring or testing the properties of just a few factors. The current project has done two things to make the large number of variables more manageable.

First, a content analysis was conducted on every research source (article, book, web-page, etc.) read for this review, in which each source was coded for the field to which it belonged, the methodologies it used, the IQ factors it studied, and the units of analysis it employed. (Each source's data and findings were also recorded in a qualitative way.) These codes were subjected to a principal components analysis, which found combinations of correlated fields, methods, factors, and units of analysis across studies. That analysis showed the distinction described in the introduction (§1.3) between fitness-based and representational studies. The analysis also confirmed a division in the fitness literature between studies that take either user perceptions or artifacts as their units of analysis, as well as the distinction often made in the representational literature between accuracy and completeness. Finally, the analysis showed the relative frequency with which each type of literature has studied different IQ factors, as well as which authors have published the most in each area. From these findings, the sections and order of variables and authors presented in this chapter were determined.

Second, the methods chapter (§3.3) outlines a pragmatic way in which the many questions generated in the literature review can be answered in groups. By identifying which questions can be answered using the same few data samples in this research context, those research tasks (data collection, coding, etc.) that should answer the largest number of most interesting questions can be prioritized, and the overall number of research tasks minimized.

This chapter is organized as follows. Sections are dedicated to the findings of a recent

dissertation by Stvilia (2006) on measuring IQ, and to each of the three IQ types identified in the literature that are relevant to the current project, namely: artifactual fitness, representational accuracy, and representational completeness. Each sub-section presents the variables mentioned by the IQ literature when discussing that IQ type. For example, artifact length is discussed as an indicator of completeness. The variables in each sub-section are ordered in terms of those that are most-to-least commonly mentioned in the IQ literature, with respect to that IQ type. Because artifact length is more commonly associated with completeness than is the presence of a copyright statement, length is presented before copyright in the representational completeness sub-section of this review. At the beginning of each sub-section, for the sake of brevity and focus, those variables that are either inapplicable or trivial to study in this research context are only briefly glossed, and not discussed at length. For each remaining variable, the academic sources are presented in terms of those that are most-to-least affiliated with LIS. Ending the discussion of each variable, research sub-questions are given that relate that variable to the research context under study. Ending each sub-section, the research sub-questions identified throughout that section are summarized. Discussions of the research methods used in the literature in this chapter, as well as of the operationalization of each variable in this research context, are given in the next chapter (3).

Table 2.1 summarizes how the IQ factors in this review have been categorized, using the order in which they occur in the review.

Table 2.1: IQ factor categorizations

| **Artifactual fitness** | **Representational accuracy** | **Representational completeness** |
|---|---|---|
| accessibility | currency | author agenda |
| maintainability | citing sources | citing similar sources |
| user feedback | identifying authors | statements of benefits and risks |
| searchability | peer review | length |
| language and readability | objectivity | collaborative filtering |
| cost | empiricality | recommendations |
| availability | consistency | critical analyses |
| personalizability | association-affiliated buttons | original research |
| portability | advertisements | number of authors |
| distributability | inlinks | descriptive synopses |
| standards compliance | PageRank | copyright statements and disclaimers |
| interoperability | traceability | |
| | rate of change | |
| | number of edits | |

## 2.2 Stvilia's "Measuring information quality"

Perhaps the most comparable work to the current study is Stvilia (2006), a recent dissertation on the measurement of the quality of ostensibly any type of information, which focused most on assessing IQ in manufacturing/production and other collaborative content creation (i.e., "quality assurance") environments. This section critically reviews that dissertation's theories and methods.

Like the current dissertation, the IQ variables discussed in that dissertation included only those that are either representation- or artifactual fitness-related, and not perceptual fitness-related (though these distinctions were left un-named). Unlike the current dissertation, the approach taken to operationalizing the variables in context embraced information retrieval and heuristic machine learning methods more than statistically principled methods (cf. §3.4).

Three types of IQ criteria were identified in Stvilia (2006), namely: intrinsic, contextual, and reputational ("reputation" here refers to network graph-based authority, not perceived reputation; pp. 61-80). Intrinsic IQ refers to criteria that can be measured objectively about artifacts, out of context, including: accuracy (i.e., factual validity), cohesiveness (i.e., on-topic-ness), cognitive complexity (e.g., readability), semantic consistency (same values for same concepts), structural consistency (similar things are represented similarly), currency (age), informativeness (information-theoretic redundancy), naturalness (amount of typification/templating), and completeness (granularity or precision). Contextual IQ refers to many of the same criteria as in the intrinsic category, but evaluated with respect to the information's quality in only that context, the repeated variables being: factual accuracy, cognitive complexity, naturalness, informativeness, completeness, and semantic and structural consistency. Additionally, accessibility (speed, ease of obtaining), relevance (aboutness), security (protecting the information from harm), verifiability (provability in context), and volatility (time that the information remains valid) were included in the contextual IQ category, though several of these possess a-contextual measures as well. Finally, reputational IQ refers to the information's authority or position within some kind of reputational social system or network. The only variable in this category, authority, was measured via

PageRank, HITS, inlinks, and the like.

The distinction between measuring mostly the same IQ variables using the same measures either contextually or a-contextually can have value, such as when attempting to compare the IQ in some context either with different contexts or across an ecosystem of contexts. However, when several contexts within a common genre of information artifacts are being assessed, the distinction seems more related to a necessary research practice than to semantically mutually exclusive classification of the IQ criteria. That is, the definitions of factual accuracy, naturalness, or completeness do not change depending on research context.

Finally, Stvilia employed exploratory factor analysis (EFA) to find linear dimensions of IQ variables on Wikipedia (pp. 182-183). (For reasons behind the current dissertation's use of principal components analysis (PCA) instead EFA, which is a related type of latent variable model, see §3.4.4. The reader should also note that a later work by Stvilia et al. (2009) switched to PCA.) In order to create IQ metrics specific to Wikipedia, each dimensions' coefficients were mathematically equated with the IQ concept that the variables loading most highly on each dimension typically measure. For example, authority = some coefficient * a page's count of unique editors + some coefficient * a page's count of edits + other coefficients * variables. This could be reasonable, if the coefficients signs, which form a spectrum, are preserved; only one negative sign was present in the equations, though this need not necessarily indicate an error. However, rather than employing statistical re-expression techniques to normalize the variables before EFA, machine learning clustering and supervised learning algorithms were used, to attempt to account for the non-normality of both the variables and EFA 'metrics' post hoc. This is inadvisable, because nearly all

clustering and supervised learning methods were developed with Gaussian-like clusters in mind, meaning that such methods will not account for non-normality (Hastie et al., 2005). Such a mixture of heuristics and statistics makes dubious the EFA results.

For a presentation of how all of the variables in Stvilia (2006) were operationalized in the current project's context, see §3.3. Finally, for a discussion of how the current study's results compare with the Wikipedia case study in that dissertation, see §5.7.

## 2.3 Artifactual fitness IQ

From most-to-least frequent, the IQ literature includes the following 12 affectation-related IQ variables: accessibility, maintainability, user feedback, searchability, language and readability, cost, availability, personalizability, portability, distributability, standards compliance, and interoperability. Only the most non-trivial and applicable variables to the current research context are discussed at length, with the trivial or inapplicable variables summarized briefly at the end of this section.

### 2.3.1 Accessibility

Accessibility has four meanings in the literature, the most common of which is *simplicity or efficiency*. The following authors use the term in a cursory way, because they are focused on broad summaries of criteria in the IQ literature. Calero et al. (2004), a survey of computer science (CS) IQ literature, refers only to "improving web information access and use" (p. 148). Knight and Burn (2005) contains a very good survey of the IQ literature, including authors who mention accessibility in all of the following senses. Halaris et al. (2007)

also summarizes variables included in the multitude of service- and information-quality instruments, noting that "accessibility is a general term used to describe the degree to which a system is usable by as many people as possible without modification" (p. 385).

More specifically, many authors speak of simplicity and efficiency in terms of navigation, search interfaces, sitemaps, documentation, and "what's new" sections of websites, including: Diering and Palmer (2001), Stausberg and Fuchs (2000), and Sandvik (1999). Much of this literature is from medical researchers, who do not reflect more deeply on what exactly makes navigation or a search interface seem simple and efficient. Of these, Breul et al. (1999) is the only to list all of the website features just given. For the current project, this literature suggests that one should compare these forms and features across the sites under study.

In this literature, one also can notice that a flourish of medical papers was published in the late 1990s and early 2000s, when the Web was entering mainstream culture. Many of these articles have a similar form, namely: a subject specialist (e.g., a radiologist) wants to know what kind of information is available on the Web about their specialty. So, they find a few general IQ or subject-specific content criteria that seem important (in their expert opinion); they conduct a basic Web search ("because that is what their patients would do," as they commonly say) and content analysis of the results; and they report how well the sites they found meet those criteria. Because the current project is not about a medical topic, the findings of such studies will not be presented, though the criteria they use will be.

The second most common way of characterizing accessibility in the IQ literature is to speak generally about the *digital divide*, usually advocating the position that digital com-

puting resources should become easy, free, and ubiquitous for everyone. Eysenbach and Jadad (2001) is a low-tech summary of the issues involved in providing healthcare information to consumers, including: how to keep the consumer "thirsty" for information, making large databases (e.g., MedLine) available to the general public, improving literacy about health topics, increasing Internet access, ensuring that information is well-represented and well-organized, and giving consumers control over their information. Moran and Oliver (2007) speaks of how difficult it would be to regulate or deny access to sites of low quality on the Net. Clement et al. (2002) puts forth system design recommendations, hoping to facilitate "instant patient access to information regarding their condition at any time ...[and] it allows patients to play a more active part in their own healthcare" (Conclusion, para. 1), sentiments also advocated by Moran and Oliver (2007). Cline and Haynes (2001) is a literature review of health information seeking on the Web. It advocates improving health information access by focusing on improving three modes of access: webpages, support groups, and consultations with professionals. Although many others have expressed similar views (e.g., Coulter et al., 2006; Burkell, 2004; Kim et al., 1999), they do not often lead to testable IQ criteria. The most valuable contribution of these articles for the current project is the implication that one could compare how ubiquitously the sites in question could be used. For example, how many different types of media, data streams, APIs,[1] database downloads, etc. do the sites offer?

Third, a common set of authors speak of accessibility in terms of Web server speed, webpage existence, or access permissions (i.e., *availability*). For example, accessibility

[1] Application Programming Interface: a programmatic interface for accessing a site, service, or system's content or features.

is "The degree to which information is available, easily obtainable or quickly retrievable when needed" (Al-Hakim, 2007, p. 165). Others who share this view include: Michnik and Lo (2009), Stvilia et al. (2009), Su et al. (2008), Zhu (2008), and Stvilia (2006). Although availability is one of the criteria deemed too trivial in the present research context for in-depth consideration, the definition of availability given by Moran and Oliver (2007), namely searching webpages for "Deadlinks," could be fruitful in this context.

Finally, several researchers define accessibility in terms of international *standards and compatibility*. Llinas et al. (2008) compares British, Spanish, and US hospital websites using TAW, a tool that analyzes a website code's compliance with the Web Content Accessibility Guidelines (WCAG) 1.0, published by the World Wide Web Consortium (W3C). Likewise, the framework constructed by Fritz and Schiefer (2003) for evaluating the IQ of agribusiness websites mentions having website accessibility checked "by a robot" (p. 5). Coulter et al. (2006), a lengthy management-oriented report from several health institutes at Oxford that offers website development and dissemination advise in general, mentions (p. 10-11) complying with recommendations by the Royal National Institute for the Blind (RNIB), a UK charity that publishes its own "See it Right" Web accessibility checklist. The RNIB website also provides links to the WCAG, the US government's Section 508 checklist, and IBM's checklist (RNIB, 2008). One paper, von Danwitz et al. (1999), also mentions manually comparing sites' compatibility across different Web browsers, by which is meant the ability of different browsers' rendering engines to properly display the code of the page. This literature suggests that it would be beneficial to check the sites under study against the aforementioned guidelines and checklists.

### 2.3.2 Readability

This criterion has two meanings in the literature. In the first sense, which draws upon old work and is almost exclusively used by medical literature, *formulae* characterize the "ease" with which a text can be read, in terms of various descriptive document statistics. The goal is often to approximate the US grade school level of a text. This is intended for cases when conducting a statistically rigorous readability survey of actual readers is infeasible, as is the case for large collections on the Web. The most popular such formulae are the Flesh-Kincaid Grade Levels and the Flesh Reading Ease tests (Flesch, 1948, 1962). Both of these tests are linear equations that use counts of the total words, sentences, and syllables of a text, though with different constant (i.e., heuristic) intercepts and weights on the variables. Medical papers that have applied these equations to websites include: Llinas et al. (2008), Langille et al. (2006), and Abbott (2000). In decreasing order of popularity in the IQ literature, other indices include the Simple Measure of Gobbledygook (SMOG; McLaughlin, 1969), which incorporates polysyllables; the Fry Readability Graph (FRG; Fry, 1977), which plots the average number of sentences by syllables per 100 words; the Lexile Framework for Reading (Stenner, 1996), apparently a proprietary, industrial creation; the Gunning-Fog Index (Gunning, 1969), which incorporates complex words; the Automated Readability Index (ARI; Smith and Senter, 1967), which uses characters; and the Coleman-Liau Index (Coleman and Liau, 1975), which resembles ARI. These non-Flesch indices are applied to websites by several medical works, including: Berland et al. (2001), D'Alessandro et al. (2001), and Estrada et al. (2000). Is there any reason why these studies did not use a Flesch index? The first four used multiple formulae simply as a

comparison to the Flesch measures. Fitzmaurice and Adams (2000) offered no explanation as to why Gunning-Fog was used, other than to cite Vahabi and Ferris (1995), which uses Flesch, Fry, and Gunning Fog. Murphy et al. (2001) explain that they used Gunning Fog because it is "a more conservative assessment that yields a readability score two grade levels higher than the Flesch-Kincaid" (p. 100). The reader should also note that Wikipedia's IQ criterion on this matter – namely, "**well-written:** [original emphasis] its prose is engaging, even brilliant, and of a professional standard" – leaves the judgment of reading ease to either the reader or the site's editorial population (Wikipedia, 2009d).

Blumenstock (2008) and Stvilia (2006) were the only LIS studies encountered that applied readability metrics to websites. The first study used most of the metrics just mentioned, taking the same "[throw everything at the problem, including the] kitchen sink" (p. 1096) approach as did the medical authors. Seeking to demonstrate that page length is a fairly good indicator of quality, that study applied 30 metrics from natural language processing and machine learning, to Wikipedia pages and found that page length alone correctly predicted 93.31% of the time whether a page would have been honored with "featured article" status (Wikipedia, 2009d) by the Wikipedia population. Exact results are not reported for the readability tests, however the authors note that no other metric achieved greater than 97.99% predictive accuracy. By comparison, Stvilia (2006) advocated "text readability indices" (p. 65) – naming Flesch, Kincaid, and Fog without explanation – for measuring texts' complexity. For the current project, it would be a simple matter to run a sample of pages from the sites under study through all of these readability metrics. There appears to be no principled reason why most authors prefer one metric over another; this makes sense, due to the heuristic nature of the formulae themselves.

The second meaning of readability in the literature is *legibility*. The IQ literature reviewed by Cline and Haynes (2001) speaks of legibility in the same sentence as "Color coordination, lack of clutter, [and] unobtrusive backgrounds" (Design features section, list item 4), making the point that aesthetic and formatting characteristics "contribute to comfort and use" (same list item). Kihlstrom (2001) studied pharmacy benefit management sites, where "the print size was too small to read easily" (p. 66). Martinez-Lopez and Ruiz-Crespo (1998), studying webpages on rotator cuff rupture, noted that some had insufficient color contrast between the text and background, and complained of "abuse of different fonts and character sizes" (quoted in Eysenbach et al., 2002, p. 11). For the current research context, since all of the pages under study are in HTML, their CSS (or deprecated style tag) font properties could easily be compared. It should also be possible to identify font sizes and font faces/families appearing in images on the sites.

### 2.3.3 Portability, standards compliance, & interoperability

The first and largest group of authors writing on this criterion is again from medicine, and is most concerned with whether pages conform to government guidelines and mandates. For example, Davison (1997), surveying 167 sites about nutrition, note degrees of non-compliance with 11 Canadian nutritional guidelines. Since no known governmental or other guidelines exist for fansites, this literature is of little relevance to the current project.

Similar to the digital divide researchers, but with a more technical angle, a second, and much smaller, group of authors in this area are concerned with making websites that are adaptable, changeable, replaceable, reusable, traceable, testable, and/or installable.

Two papers, Leung (2001) and Zeist and Hendriks (1996), argued that portability can be achieved through compliance with open standards, measuring the quality of intranet applications against the Extended ISO[2] 9126 model popular in computer science. Validating the results against a survey of user satisfaction, Leung (2001) notes that this model shows promise, though the user sample size surveyed was inadequate for statistically rigorous conclusions. One paper, López-Ornelas et al. (2005), in creating an instrument for evaluating online journals, emphasizes achieving portability through distributing content in formats with broad "external recognition" (p. 135). After validating the instrument against a survey of journal editors, the authors conclude that their criteria "were mainly clear and pertinent, but were not enough and there were still more important questions to include" (p. 139), though those questions were not discussed.

For the current project, this literature again suggests the value of comparing the international standards, and other ways, in which fansites make their content portable and ubiquitously accessible. In terms of website software portability, only the wiki-based sites use open-source software code (i.e., MediaWiki), which can be freely replicated and modified on other sites. The edited sites probably do use content management software systems, judging by the webpage templates and large amount of content on those sites. However, the details of those systems are kept hidden from the user.

### 2.3.4 Trivial and inapplicable

The following variables are either too trivial in, or inapplicable to, this research context to deserve a detailed literature review.

[2]ISO: International Organization for Standardization

**Maintainability** refers to how difficult it is for users or site proprietors to maintain a website and its contents (Leung, 2001). Wikipedia and Wikia are both wikis, using the MediaWiki software, which has very public, popular, and well-studied features. For example, there is little point to rehashing how anyone with Web access could post content to Wikipedia (Wikipedia, 2009c), or how normative editorial rules have emerged from the virtual population of Wikipedia users (Butler et al., 2008). On the other hand, the inner workings of IMDb and GateWorld were proprietary/mysterious and editorially controlled.

**User feedback** , a common requirement for evaluating information system quality (e.g., Bernstam et al., 2008; Yadav, 2008; Coulter et al., 2006), was also possible on all of the sites. Wikis, as a feature of the software, received all of their content from users, and editorially controlled sites had their own writers and editors create original, as well as modify user, content.

**Searchability** is a criterion often encountered in literature about the IQ of electronic health information (e.g., Bernstam et al., 2008; Häyrinen et al., 2008; Llinas et al., 2008). All of the fansites provided a search feature, all of which were integrated into the sites' infrastructures, with the exception of GateWorld's, which was provided by Google.

**Language:** The portions of the fansites studied for this project were all primarily written in English. Both wiki sites provided the capability for users to create Stargate-related material in other languages, and IMDb had sister sites in Germany, Italy, Spain, France, and Portugal. Pages in non-English languages were not studied, because the researcher would be unable to consistently assess whether pages were about Stargate-related topics or

not, especially on the wiki sites, which can include cultural reference pages. Recent cross-language website IQ literature includes: Llinas et al. (2008); Rahnavardi et al. (2008), and Weissenberger et al. (2004).

**Cost** is a common criteria for automated IQ assessment frameworks in commercial environments (e.g., Bizer and Cyganiak, 2009; Häyrinen et al., 2008; Yadav, 2008; Burgess et al., 2007). Although IMDb charged for access to social networking features, which were targetted towards members of the entertainment industry (e.g., resumes, message boards, contact information IMDb, 2010c) such features were not descriptive of a specific media phenomenon, and hence were not under study here. Other than exposing users to advertisements, all of the phenomenon-descriptive content on IMDb and the other sites was available at no cost to the public.

**Availability:** These sites were all fairly large and, therefore, highly available, in terms of global download speeds and server uptimes (Netcraft, 2010a, 2010b, 2010c, 2010d). The smallest site, GateWorld, was hosted at ThePlanet.com, which had six data centers in Dallas and Houston, Texas, USA (ThePlanet, 2010). GateWorld also had a backup site, called the GateWorld Alpha Site (GateWorld, 2010a) – a reference to the off-world base in the Stargate canon to which members of the human race could escape in the event of Earth's impending destruction, to which it could switch, if the primary site failed. During the study, that site only contained server maintenance and status update information. For literature defining accessibility in terms of availability, see §2.3.1.

**Personalizability:** Both of the wikis and IMDb allowed basic interface customizations, including modifying themes/skins and page content sorting options, user bookmarking or tracking of favorite content, and updating of personal profile information shown to other users on the site. MediaWikis also allowed the user to customize their date and time format as well as timezone, in addition to content editor interface preferences. GateWorld only enabled personal profiles on its message forum, and was otherwise not personalizable. Recent personalization IQ literature includes: Halaris et al. (2007), Barnes and Vidgen (2003), and Leung (2001).

**Distributability:** All of the media were digital text and multimedia, and were, therefore, highly distributable. Relevant literature includes: Coulter et al. (2006), Meyer (2006) and Rosenzweig (2006).

### 2.3.5 Artifactual fitness research sub-questions

The literature reviewed in this section suggested the following research questions:

- **AF1:** *Accessibility as simplicity*: how do the sites' navigation, search interfaces, sitemaps, help documentation, and similar sections compare?
- **AF2:** *Accessibility as digital divide*: by how many, and which, channels can fansites' content be accessed?
- **AF3:** *Accessibility as availability*: do any of the webpages contain broken links, and how many?

- **AF4:** *Accessibility as standards-compliance*: to what extent do the pages' HTML and CSS code comply with popular accessibility guidelines?

- **AF5:** *Readability as formulaic*: what is the formulaic reading level of these pages, using different metrics?

- **AF6:** *Readability as stylistic*: how do pages' visual style features (e.g., colors and fonts) compare with each other?

## 2.4 Representational IQ: Accuracy

The term "accuracy" is used generally in this literature to describe the degree of representativity of an informational artifact or representation of that which it represents. For example, one may speak of a painting accurately depicting a person.

From most-to-least frequent, the IQ literature includes the following 14 accuracy related IQ variables: currency, citing sources, identifying authors, peer review, objectivity, empiricality, consistency, association-affiliated buttons, advertisements, inlinks, PageRank, traceability, rate of change, and number of edits.

**Recommendations & original research** will be discussed alongside advertisements and peer review, respectively, in this section, rather than in the following section on completeness. This is because the ideas are closely related, and because there is little IQ literature on these variables, though they are relevant to this research context.

### 2.4.1 Currency

Also called currentness, and bearing no relation to financial currency, information currency is most often discussed in the sense of *timeliness*. Timely information must be dated, ideally with both creation/posting and last-modification dates, and the last-modification date should be "recent," where recency varies by the urgency or importance of the information. A number of authors center around this general definition. For Kunst et al. (2002), a study of 121 webpages on five common health topics, currency is displaying "the date of the original document or content posting on the internet [sic], and that of any updates" (p. 581). 49% of the sites in that study complied with that definition. Griffiths and Christensen (2000) deemed depression websites to be recent, if modified in the last month. Pérez-López and Roncero (2006) studied 75 websites about postmenopausal osteoporosis, defined currency as "date of publication or update clearly stated on all pages" (p. 670), and found non-profit and governmental sites to be of higher overall quality (results for individual variables were not reported) to be higher than for industrial sites. Bharosa et al. (2008), a theoretical survey of IQ criteria relevant to the emergency response domain, noted that "Emergency situations changes [sic] time by time [sic] so it is very important to know the order of events" (p. 557). Other authors with similar definitions include: Stvilia et al. (2009), Marriott et al. (2008), Burgess et al. (2007), Harland and Bath (2007), Lewiecki et al. (2006), and Frické and Fallis (2004). Also, the following medical authors are among those who have estimated recency to be around six months for medical literature on the Web: Berland et al. (2001), Stausberg et al. (2001), and Griffiths and Christensen (2000).

To test this literature's definition of currency on fansites, since the sites have pages

with serial index numbers (i.e., TV series and episode), it should be straightforward to randomly sample content entries and describe the distributions of creation/posting and possibly last-modification dates. MediaWiki-based sites, as part of their default functionality, have public and complete revision history records for every page. The edited sites, on the other hand, have only very sparsely dated information; hence, the next definition of currency may be more appropriate for edited sites.

Currency has also been defined in terms of the up-to-dateness, or state-of-the-artness, of information. Seidman et al. (2003), a paper re-forming the many criteria from Eysenbach et al. (2002) into a conceptual framework designed specifically to evaluate the IQ of diabetes consumer-information sites, notes that "Perhaps the most commonly-cited definition of quality of care [for medicine in general] is the one developed by the Institute of Medicine, which states that quality in health care is 'the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge"' (Results, para. 3). Latthe et al. (2000b), a study of 32 websites on emergency contraception, similarly defined currency as "keeping up to date with [sic] present state of medical knowledge" (Materials and methods, para. 4), and found that only 42% of the sampled sites met that definition. Al-Hakim (2007), a business study that sought IQ criteria affecting the innovation management process in an IT firm, found up-to-dateness to be a decisive IQ dimension in that context, and defined currency as follows: "how up-to-date information is with respect to customer's [sic] needs or the task at hand. It reflects also [sic] how fast the information system is updated" (p. 165). Finally, Feng and Liu (2008) sent questionnaires asking 100 instructors and postgraduate students of a distance learning course taught between four universities in China and the

Netherlands to rate the importance for the course of 10 IQ criteria on a Likert scale. Using their own custom clustering computations, they determined that those surveyed perceived up-to-dateness to be similarly important as being able to provide user feedback, and that the course information be "appropriate" and easily understandable. Other authors adopting this definition of currency include: Su et al. (2008) and Yadav (2008).

For these fansites, much like comparing the up-to-dateness of news sites in general (cf. Abdulla et al., 2002), one could survey the cross-site coverage of several recent Stargate-related news events, comparing the facts asserted by the different sources both against each other and against what the show's producers say for themselves (e.g., on the producers' personal blogs, in interviews they have given, etc.). Unfortunately, this was not feasible for the current study. Comparing the articles in each site's news stream, finding several recent news events that they all cover, and investigating factual discrepancies between news stories by consulting several independent sources, to determine which site(s) is/are most often correct, would have been too time-consuming.

Finally, several writers defined the "freshness" of a site in terms of whether it discloses an update and maintenance frequency policy. Ekman et al. (2005), a study of cancer risk websites, adopted the popular Health on the Net Code of Conduct guidelines (HONcode, HealthOnTheNet, 2008), which, in the original form drafted by the EU and Swedish National Board of Health and Welfare, included a definition of currency as "Date last updated and the frequency of these updates" (Ekman et al., 2005, p. 767). The current version requires only "The date when a clinical page was last modified" (HealthOnTheNet, 2008, Principle 4, view complete version). The IQ literature review by Cline and Haynes (2001) included in the definition of currency "policies and methods regarding updating" (Evalu-

ating health information on the Internet section, para. 7, list item 1). The review by Kim et al. (1999) repeatedly defined currency as "frequency of update [sic], freshness, [and] maintenance of [sic] site" (Discussion, para. 1). Also, von Danwitz et al. (1999) recommended that sites disclose their "date[s] of technical maintenance." Finally, one author, Dragulanescu (2002), recommended judging site freshness by whether a recent copyright date is attached to every page.

Looking for recent copyright dates on the fansites is a trivial matter, for they all have them and they are all copyrighted up the present year, and update frequency policies are rare on fansites. As a feature of the software, the content of wikis is updated by users from moment to moment, rather than on a schedule. Additionally, both browsing and searching of the two edited sites under study indicate that neither have posted an update frequency policy.

### 2.4.2 Citing sources & identifying authors

Every IQ article that mentioned citing sources, of which there were 65 (e.g., Charnock and Shepperd, 2004; Gagliardi and Jadad, 2002; Meric et al., 2002), described "*attribution*" or "*referencing*" by website authors of the sources of their information, as is expected of academic work. Even Wikipedia's criteria for "featured articles" requires that claims made in such articles be "verifiable against reliable sources" (Wikipedia, 2009d, Principle 1, list item (c)), where "reliable sources" are defined as "credible published materials with a reliable publication process; their authors are generally regarded as trustworthy or authoritative *in relation to the subject at hand* [original emphasis]. How reliable a source

is depends on context" (Wikipedia, 2009g, Introduction, para. 2). However, very little of the IQ literature goes so far as to say that cited sources should be reliable. Perhaps this is because websites tend to cite their sources much less frequently than do academic works (so any citation behavior on the Web is to be encouraged), or because artifact-oriented IQ researchers recognize that evaluating the credibility, relevance, etc. of sources involves perceptual judgment by readers in context, and credibility perception has its own separate branch of the IQ literature. In any case, if sources are cited by an author, and are available for a reader to consult, presumably the reader can form an interpretation about the nature and quality of those sources.

On the other hand, the artifactual IQ literature does recommend that more authority-related details be provided by authors or organizations *about themselves*. One definition, by Silberg et al. (1997), is referenced by many: "Authors and contributors, their affiliations, and relevant credentials should be provided" (p. 1245). A body of literature of similar size (50 papers), and with similar members (e.g., Buhi et al., 2010; McKemmish et al., 2009; Kunst et al., 2002), as the one in the previous paragraph converges on these same criteria. This focus on author identities is similar to the literature's focus on author agendas/goals, target audiences, objectivity, and to the absence of advertisements (discussed throughout this and the next section). In this way, there appears to be a consensus that, if only one can identify the author of a text and their motives, one can know how trusting to be of the accuracy and completeness of content produced by that author. For example, a number of studies note that corporate websites, which are primarily motivated by making money, do not usually identify a single person who is responsible for the content (Robbins and Stylianou, 2003), are more likely to contain content that is incomplete or misleading in the

company's favor (Green et al., 2004), and to mix advertisements with page content (Sellitto and Burgess, 2005).

For the current project, this literature suggests the value of surveying source-citation and author-identification practices on the sites under study, as well as possibly cited-source types.

### 2.4.3 Peer review & original research

Although the IQ literature prizes websites that "specify editorial review processes and identify reviewers" (Cline and Haynes, 2001, Evaluating health information on the Internet section, para. 5, list item 3), and which display proof of "evaluation by professionals" (Coulter et al., 2006, p. 11), no evidence of these things appears on the fansites under study. Editor-controlled sites put editors in a dictatorial position, where all content is filtered through them. As is common among corporations, IMDb and GateWorld provide little insight into their editorial review processes, other than offering "submission guidelines," which describe the formats and syntaxes in which users should submit content, but not how the editors or reviewers will fact-check or otherwise judge content submissions (IMDb, 2009i; GateWorld, 2009f). Since editors can quickly change and instantly republish (without notifying the public) webpages that receive criticism from readers, the need for peer reviewers to ensure the quality of an archival copy likely is not felt by these sites (Cline and Haynes, 2001; Lacroix et al., 1994, p. 417). On Wikipedia, the label "peer review" is given to the process by which featured articles are reviewed (Wikipedia, 2009i). Though this has prompted at least one scholar to call the process a sort of "populist peer-review" (Rosen-

zweig, 2006), the "free-for-all, communitarian approach of Wikipedia" (Cronin, 2005) makes no claim to authority. In Wikipedia's words: "It is not academic peer review by a group of experts in a particular subject, and articles that undergo this process should not be assumed to have greater authority than any other" (Wikipedia, 2009f, Introduction, para. 1).

Nevertheless, regardless of its authority, content on fansites often includes original research – using the term "research" in the journalistic sense of informal personal investigation, such as if a fan speaks with a producer themselves and reports what the producer told them – resembling a sort of "unsanctioned scholarship" (Bradley, 2005). As Jenkins (1992) says: "the intimate knowledge and cultural competency of the popular reader also promotes critical evaluation and interpretation, the exercise of a popular 'expertise' that mirrors in interesting ways the knowledge-production of the academy. ... Within the realm of popular culture, fans are the true experts; they constitute a competing educational elite, albeit it one without official recognition or social power" (p. 86). Gelder (2004) similarly speaks of a need for work that can "pitch itself against neglect [i.e., of fan-made literature] elsewhere, especially in academia, replacing it with a much fuller, para-academic knowledge of the field: giving itself the amateur's time and space to fill in the gaps" (p. 88). On the sites under study, though forbidden from Wikipedia (2009e) and generally from Wikia, categories for pages containing original research do exist in some Wikia communities (e.g., Wikia, 2009c). Therefore, while it would not be productive to study peer review processes on these sites, the question of what forms and contents the original research of fans takes/includes is of interest.

### 2.4.4 Objectivity

In the IQ literature, the objectivity of an artifact is usually inferred from the *consensus of experts*. Bernstam et al. (2005), a survey of which medical website IQ instruments may be simple and objective enough to actually be usable by normal patients, speaks of objectivity as being tantamount to inter-rater reliability, when two or more experts – assuming (unrealistically) that these experts have never been influenced either by each other, by the same social norms, or by the biological substrates common across our species – independently come to the same conclusion. Burkell (2004) – a study showing that, whereas health information seals of approval often only indicate author/source disclosure readability, undergraduate students believe them to signify much broader information quality – notes that "for most if not all health information, there exists no gold standard that can be used to determine whether the information is in some absolute sense correct" (p. 12). And, Naumann and Rolker (2000), a German survey of IQ criteria, assesses objectivity merely in terms of "expert input." In these and 22 other sources, though inter-subjective expert consensus is not actually objective, there is a sense that it is the closest approximation of objectivity available (e.g., Jakobsson and Giversen, 2009; Feng and Liu, 2008; Al-Hakim, 2007). This manifests in Wikipedia's IQ criteria as "neutral point of view" (Wikipedia, 2009d), which requires that "all *significant* [original emphasis] views that have been published by reliable sources" (Wikipedia, 2010a) be represented equally and without preference. 'Published by reliable sources' refers to Wikipedia's empirical verifiability criterion (cf. §2.4.5). For fansites, one could look for whether accepted descriptions and interpretations exist in fan reviews and analyses on the same topic across sites.

Additionally, quantitative researchers speak of the objective measurements made by automated/computational processes, typically in a passing/undiscussed way (e.g., Michnik and Lo, 2009; Yadav, 2008; Zhu, 2008). As in content analysis, there is a notion that, whereas inter-rater reliability is the best objectivity that humans can achieve, if a computer can adequately count/measure a phenomenon, the results will be as objective as possible. Many of the variables throughout this dissertation were measured by automated parsing programs, as described in the Methods chapter (3).

### 2.4.5 Empiricality

Not to be confused with objectivity, empiricality refers most commonly and generally to the citing of *evidence of any kind for claims made by/in information* (e.g., Marriott et al., 2008; Ilic et al., 2004; Rolland et al., 2000), in order to verify (Jakobsson and Giversen, 2009; Naumann and Rolker, 2000; Veronin and Ramirez, 2000) the reality or factuality (Yadav, 2008; Stvilia, 2006; Bogenschutz, 2000) of those claims. Wikipedia's IQ criteria also agree with this general definition (Wikipedia, 2009d). Besides that general definition, the medical literature in particular often speaks of "hierarchies/levels of evidence," where randomized and controlled trials (RCTs) are at the top of the hierarchy, and are preferred, and descriptive/observational case studies are at the bottom, and are less trusted (e.g., Selman et al., 2006; Kunst and Khan, 2002; Latthe et al., 2000b). Seidman et al. (2003) also places systematic meta-reviews of multiple studies above RCTs on the hierarchy, because such work can span multiple RCTs. Finally, several authors define the empiricality of websites in terms of their methodological validity and rigor. Kunst et al. (2002) connects method-

ological quality directly with the notion of hierarchies of evidence, saying "We looked at the hierarchy of evidence posted on each website, examining whether the levels assigned to various pieces of information were related to their validity or methodological quality" (p. 581). Gillois et al. (1999), a study of eight cardiovascular risk prediction sites, criticized sites with "no valid use of information" (cited in Eysenbach et al., 2002, Online Table B). Also, the following studies criticized websites for either citing or using inappropriate analysis procedures and drawing incorrect implications: Gordon et al. (2001), Türp et al. (2001), and Martinez-Lopez and Ruiz-Crespo (1998).

Implications for the current project include examining whether the websites under study cite any evidence for their conclusions.

### 2.4.6 Consistency

This literature most often speaks of the consistency of *conceptual structures* and *graphic design*. About conceptual structure, Fisher and Kingma (2001) explains that "the representation of the data values is the same in all cases. It [i.e., consistency] implies that there is no redundancy in the database and that referential-integrity is enforced." Kahn et al. (2002) says that consistency "ensures a minimum level of interpretability and [that maximum] understandability is achieved" (p. 189), and Shanks and Corbitt (1999) defines it as "well-defined (perhaps formal) syntax" (p. 790). Others that provided more general definitions in this direction include: Su et al. (2009), Miettinen and Korhonen (2008), and Stead et al. (2008).

About consistency of graphic design, many authors speak of the sameness of fonts,

styles, and presentation in general (e.g., Magnus, 2006; McInerney and Bird, 2005; Katerattanakul and Siau, 1999), specifically that there be no deviations from a common style (Eppler and Muenzenmayer, 2002), and that the same images be presented again and again to the user (Halaris et al., 2007).

For the current project, one could examine whether a consistent set of data records/fields, as well as stylistic constants, emerge across the sites.

### 2.4.7 Advertisements & recommendations

Closely related to biased violation of objectivity in many medical IQ authors' minds (e.g., Marriott et al., 2008; Sellitto and Burgess, 2005; Frické and Fallis, 2004), ads are thought to indicate information that may be *misleading* or not 100% accurate (e.g., Yadav, 2008; Beredjiklian et al., 2000; Galimberti and Jain, 2000), especially if accompanied by testimonials, guarantees, or claims of persecution (Morahan-Martin and Anderson, 2000). Doupi and Van Der Lei (1999), Kihlstrom (2001), and Tu and Zimmerman (2001) are especially wary of ads *not being distinct from content*, and Howitt et al. (2002) and Fritch and Cromwell (2001) insist that ads be accompanied by *disclosure* of the author's agenda, along with any conflicts of interest. Finally, Haddow (2003) notes that much advertising on the Web is not governed by national regulations, making it potentially all-the-more shady.

On non-studio-run fansites, which are entertainment-oriented and which receive most of their content from either devoted fans or corporate sponsors, advertisements are more likely to be about collectible products (DVDs, memorabilia, etc.) than about drug treatments. There also are not likely to be as serious of concerns about users being misled into buying

a faulty product, because user expenditure and risk are minimal, and the vendors being linked-to are often large corporations (Amazon, iTunes, etc.). However, interesting results were found from inferring the vendor affiliations and target audiences had by each of the sites, from the ads that they post.

*Recommendations* about products and treatments are a rare criterion in the IQ literature, probably due to the suspicion of advertisements and author agendas had by many medical IQ authors. Citing McClung et al. (1998), Cline and Haynes (2001) noted that "only 12 of 60 articles from traditional medical sources adhered to treatment recommendations of the American Academy of Pediatrics, even when websites were from major academic medical centers" (Roadblocks section, para. 21). Also, in the DISCERN IQ instrument developed by Charnock and Shepperd (2004), users are encouraged to evaluate whether "recommendations and suggestions concerning treatment choices are realistic or appropriate" (DISCERNonline, 2009, question three, second list item).

On the fansites, since they attempt to catalog and critique a phenomenon, though citations are often made to similar content, authors generally refrain from making outright recommendations. None of the sites are financially affiliated with the studios, so little attempt is made by users or site proprietors to sell the shows, beyond representing them in a flattering way. The extent of salesmanship is that Amazon.com places product recommendations at the end of IMDb pages; GateWorld provides links to Amazon.com and iTunes from each episode page; Wikipedia and Wikia provide "External links" to corporate studio, network, and TV listing websites; and Wikia and IMDb host banner advertisements for a variety of products, often unrelated to Stargate.

However, fans and editors do make recommendations about individual episodes in the

form of *ratings*. All of the sites, except Wikipedia, allow users to rate episodes. IMDb allows this only for registered users, and, though it publicly tabulates all ratings given to individual titles in terms of user genders and ages, it prohibits collection of this data. GateWorld allows one rating per episode per IP address, and does not correlate ratings with demographics. GateWorld's editors also provide their own subjective quality rating for each episode, as well as the first-run and syndication ratings given to each episode by The Nielsen Company. Wikia offers a public (probably IP address-specific) star-rating system at the bottom of each page. Several interesting questions emerge from this. How do fan ratings compare with editor and Nielsen ratings on GateWorld, as well as across the sites on similar topics? Do high or low fan ratings correlate with any other quantitative characteristics of the fan literature available for each episode?

### 2.4.8 Inlinks & PageRank

Though many researchers have studied the properties of hyperlink graphs between authors and webpages, only a few have used it as an approach to evaluating webpages' IQ. These authors give the most attention to Google's PageRank algorithm, which is a corporate-proprietary variation on the Eigenvector centrality measure long known to network analysts (Pinski and Narin, 1976). Intuitively, Eigenvector centrality and PageRank estimate the value of a node in a network, based on the number of connections that node has to (or, in the case of hyperlinks, *coming from*) nodes with high "prestige." The prestige of a webpage "is proportional to the sum of the prestige scores of pages linking to it," (Chakrabarti, 2003, p. 209) recursively throughout the network. By comparison, inlinks refer to merely

counting the number of other pages that link to a page in question, regardless of their network centrality.

This idea has prompted website IQ researchers to surmise that, because inlinking and PageRank are conceptually similar to article citation counts, they may indicate endorsements of the reputational authority (Stvilia, 2006; Yadav, 2008), popularity (Ilic et al., 2004; Zhu and Gauch, 2000), or methodological quality (Frické and Fallis, 2004) of webpages. To this end, Griffiths and Christensen (2005), a study of 24 depression websites, found that PageRank values do indeed correlate highly with IQ evaluations made by medical experts following a popular non-network-oriented IQ instrument (i.e., DISCERN). That study concluded that PageRank would be good for the average consumer, who would not have the time or expertise to use DISCERN, though DISCERN did give more nuanced results. Similarly, Hu et al. (2007), one of several engineering-oriented authors working on predicting featured article status in Wikipedia, showed that users who contribute more featured content tend to have stronger network positions.

For fansite research, it would be interesting to know if those pages that have the most inlinks or the highest PageRanks correlate with any other quantifiable and IQ-related aspects of the pages, such as their lengths, amount of analysis or trivia, episode ratings left by fans or editors, or Nielsen TV network ratings. However, since only half of the sites are wikis, and edited sites do not always provide author information, it would be difficult to model author authority across these sites.

Other than PageRank, link-oriented researchers tend to be somewhat qualitative and interpretive, resembling what Schneider and Foot (2005) might call "Web Sphere Analysis" or what Herring (2007) might call thematic, link-based content analysis. At least one IQ

researcher (Sellitto and Burgess, 2005) speaks of qualitatively coding the destinations of links, in order to inductively typologize linking behavior on different types of sites. Several others offer more perfunctory/structural content analyses of links, including whether links are broken (Zhu and Gauch, 2000; Martinez-Lopez and Ruiz-Crespo, 1998) and how many links of any kind exist on webpages (e.g., Diering and Palmer, 2001; Hoffman-Goetz and Clarke, 2000; Breul et al., 1999). This research suggests the value of characterizing link frequencies and types on fansites.

### 2.4.9 Trivial and inapplicable

Only three of these variables are not of significant importance in this research context to warrant a literature review.

**Association-affiliated buttons:** such as the quality and security certification icons that appear on many medical and consumer websites (e.g., Buhi et al., 2010; Hanif et al., 2007; Kasal et al., 2005), are absent from these fansites, which seem to prefer a self-made public image. IMDb only shows the logo of its parent company, Amazon.com. Wikipedia shows "A WIKIMEDIA project" and "Powered by MediaWiki" buttons. Wikia shows its own logos, with only a textual link to the MediaWiki project. GateWorld shows its own logos, along with small textual and icon links to GateWorld-managed accounts on Facebook, Twitter, Youtube, and iTunes.

**Traceability** refers to the existence of provenance records for an artifact (Madnick et al., 2009; Eppler et al., 2003; Eppler and Muenzenmayer, 2002). Records of past page versions

and editors are built into MediaWikis, as is the capability for administrators to return a webpage to an earlier version. Pages on the editorially-controlled sites contain no apparent provenance information with which to trace the page's editing or distribution history.

**Volatility & number of edits:** a small group of quantitative research has characterized website quality in terms of content change. Anthony et al. (2005), a study of 7,058 contributors to the French and Dutch Wikipedias, defined quality as the "survivability" of a content contribution over time, assuming that high quality content survives the longest without being changed. That study found that the longest-living content comes either from committed zealots who produce a large volume of valued work, or from passers-by who fix small mistakes that no one edits again. Stvilia (2006) similarly defined "volatility" as the length of time that information remains valid, and found, on Wikipedia, that article volatility is usually caused by edit wars between authors or vandalism, and is remedied either by avoiding such wars or frequently monitoring articles for vandalism, and reverting any vandalized pages to a last-good version. Adler and de Alfaro (2007) used a similar definition of quality, building a reputation system where users gain reputability if their changes survive for a long period without being edited. Finally, Lih (2003) called the total number of edits on a wiki page an indicator of quality, naming it "rigor." As with modeling author authority, these IQ measures are only feasible in a wiki context, where every page has a history log of past edits, and only half of this project's sites are wikis. Wikipedia's IQ criteria also give preference to content that "does not change significantly from day to day" (Wikipedia, 2009d). This topic may be explored in future work.

### 2.4.10 Accuracy related research sub-questions

The literature reviewed in this section suggested the following research questions:

- **RA1:** *Currency as timeliness*: what are the distributions of content creation/posting and last-modification on fansite pages?
- **RA2:** *Citing sources and identifying authors*: what types of sources and authors are cited on fansites?
- **RA3:** *Original research*: what forms and contents does original research take on these fansites?
- **RA4:** *Objectivity as impartiality*: to what extent do accepted descriptions and interpretations exist in fan reviews and analyses on the same topic across the sites?
- **RA5:** *Empiricality as verifiability*: what types of evidence are cited by articles on fansites?
- **RA6:** *Consistency in concepts and styles*: to what extent does a consistent set of data fields and stylistic motifs emerge across the sites?
- **RA7:** *Advertisements as agenda-loaded*: what can be inferred about the vendor affiliations and target audiences of the fansites from their advertisements and user profiles?
- **RA8:** *Recommendations as ratings*: how do fan ratings compare with editor and Nielsen ratings on GateWorld, as well as across the sites on similar topics?

- **RA9:** *Recommendations as ratings*: to what extent do high or low fan ratings correlate with any other quantitative IQ characteristics of the fan literature available for each episode?

- **RA10:** *Inlinks and PageRank*: to what extent do pages with high PageRank values or counts of inlinks correlate with other IQ characteristics?

- **RA11:** *Link analysis*: which pages on fansites contain the most links, and where do links on pages usually go?

## 2.5 Representational IQ: Completeness

The term "completeness" is used generally in this literature to refer to that which possesses all that it could or should possess in some context, lacking nothing. For example, a news story might attempt to completely recount all that occurred during some event. Wikipedia's IQ criteria keep to this general definition (Wikipedia, 2009d).

From most-to-least frequent, the IQ literature includes the following 11 completeness-related IQ variables: author agenda, citing similar sources, statements of benefits and risks, length, collaborative filtering, recommendations, critical analyses, original research, number of authors, descriptive synopses, and copyright statements and disclaimers.

### 2.5.1 Author agenda

As in §2.4.4, this criterion is a large part of assessing the objectivity of a webpage. However, whereas objectivity is a form of accuracy, one can also describe how completely the

author's agenda has been disclosed. The most common such theme in the IQ literature regards *money*. Influenced by Silberg et al. (1997), Griffiths and Christensen (2000) calls for disclosure of "ownership of the site and sponsorship" (Methods, para. 10), and found that all but one of the 21 depression sites under study disclosed an owner, and only three mentioned sponsors. Langille et al. (2006) evaluated 50 sites on bowel diseases "for integrity, that is information about funding and ownership" (para. 1), and found that only 23 of them disclosed their funding sources. Other studies emphasizing funding sources include: Ekman et al. (2005), Ilic et al. (2004), and Marriott et al. (2008). Gillois et al. (1999) also criticized cardiovascular risk prediction sites for not identifying institutional or business partnerships, and a number of authors called for the disclosure of organizational or author affiliations (e.g., Buhi et al., 2010; McKemmish et al., 2009; Yadav, 2008). Evaluating these issues on fansites is relatively straightforward, as each of the sites offer History, About Us, Mission Statement, and similar pages detailing their ownership, sources of funding, and affiliations. These things can also be inferred from an analysis of advertisements on the sites.

The other area in which IQ authors desire disclosure of agendas regards the goals, audience, and general nature/type/location of the organization. Many of the medical authors expect some kind of statement of purpose from websites (e.g., McKemmish et al., 2009; Diering and Palmer, 2001; Kihlstrom, 2001). Charnock and Shepperd (2004), Meric et al. (2002), and Griffiths and Christensen (2000) also expect sites to identify their topic of primary interest, specialization, or disciplinary scope, respectively. Target audience is often closely related to purpose and goals in many IQ articles, as sites tend to have serving a particular population as their primary purpose (e.g., Ekman et al., 2005; Charnock and

Shepperd, 2004; Ilic et al., 2004). Several also invoke vague notions of organization "type" and location, without further definition, as a way of situating an organization in some kind of cultural or geographical context (e.g., Bizer and Cyganiak, 2009; Pérez-López and Roncero, 2006; Meric et al., 2002). For fansites, as in the previous paragraph, statements of purpose, primary interest, and organizational type and location are common, however target audience is rarely explicitly disclosed. Nonetheless, demographics are available from user profiles on the wiki sites, and a certain amount can be inferred from a site's advertising and reading level (e.g., a high school reading level and ads featuring pictures of scantily clad women points to a teenage, heterosexual, male audience).

### 2.5.2 Citing similar sources

In addition to whether either sources or evidence are cited to validate claims made by a webpage, the IQ literature also distinguished whether sites provide citations or links to resources that are similar to themselves (e.g., *competitors*). Only two authors generalize beyond links: Charnock's DISCERNonline (2009) says that sites should "provide details of additional sources of support and information" (Principle 7), and Moult et al. (2004) that a quality site "Contains details of other sources of information" (Results, Table 1). Every other author speaks specifically of links: sites should have "referral links to other resources" (Ilic et al., 2004, p. 115), a "selection of external hyperlinks" (Rolland et al., 2000, p. 864), "links to other Internet medical sites" (Howitt et al., 2002, Results, para. 5), "provision of links" (Seidman et al., 2003, Methods, Table 1), links available about the disease and supporting conditions (Coleman, 2003, p. 165), "links to further information sources"

(Coulter et al., 2006, p. 33), "links to other listings of resources" (Fritch and Cromwell, 2001, Specific evaluation criteria section, list item 3), and links to "supplementary services" (Llinas et al., 2008, p. 126). Fansites offer considerable cross-linking between each other. Hence, there would be value in examining why, or for what, the sites link to other fansites.

### 2.5.3 Length, collaborative filtering, & number of authors

These criteria are most often used by computer scientists and philosophers to infer completeness. On the quantitative computer science side, Hu et al. (2007) – a study offering three heuristic approaches to predicting featured article status on Wikipedia from article length (i.e., word count) combined with a notion of network centrality that incorporates author and reviewer edits and links – found that length does improve prediction when only author link authority is considered, but not with reviewer authority considered. Intuitively, this means that the highest quality content on Wikipedia tends to be either lengthy and by influential authors, or short and by anonymous authors. This finding is in line with a similar study by Anthony et al. (2005), discussed in §2.4.9. Stvilia (2006) also recommended using sentence and word lengths to measure cognitive complexity, alongside readability metrics (Flesch, Kincaid, and Fog) that also incorporate those measures. In addition to counting words, Blumenstock (2008), the study that found that featured articles on Wikipedia tend to be longer than average, also counted complex and mono-syllabic words, characters, tokens, sentences, and total syllables. Lih (2003) also used total number of unique authors as an indicator of the "level of good standing" of Wikipedia pages.

For fansites, especially since all of these studies were of Wikipedia, these studies sug-

gest the value in counting text features, authors, and edits, where available. However, since this project cannot use featured article status as a benchmark of quality across all of the sites under study, the value of these variables comes primarily from being able to predict or correlate other variables from/with these variables.

More qualitative and philosophical researchers speak of these same variables as being detriments to conciseness and eloquence. In a lengthy article for the history field, Rosenzweig (2006) argued, following Wang and Strong (1996), that collaborative writing is usually less concise and well-written, with less structured arguments, is less well-organized and formatted, and is best at providing long lists of facts and trivia. The positive side of this is that the content is usually accurate, and resistant to vandalism. That paper also noted that articles in Encarta, a digital multimedia encyclopedia published by Microsoft, are, on average, one quarter the length of articles in Wikipedia, and that Wikipedia's articles vary in length much more widely than do traditional encyclopedias. This latter observation echoes many of the medical papers cited throughout this review, which often complain that coverage of their medical subject of interest is of erratic completeness. For Wikipedia's part, their IQ criteria do encourage users to create articles having a concise lead/summary section, appropriate page sectioning and indexes, a consistent citation style, appropriate-yet-succinct image captions, and to stay "focused on the main topic without going into unnecessary detail" (Wikipedia, 2009d).

Rosenzweig (2006) explains this by quoting the editor in chief of Encyclopedia Britannica, Dale Hoiberg: "Wikipedia authors write of things they're interested in, and so many subjects don't get covered; and news events get covered in great detail. The entry on Hurricane Frances is five times the length of that on Chinese art, and the entry on the British

television show Coronation Street is twice as long as the article on Tony Blair" (para. 32). Fallis (2008) made a similar argument, that collaborative writing by open-source communities takes longer to have its errors fixed than does computer code produced by open-source communities, because such writing does not "bump up against reality" (i.e., break the system, if written incorrectly). So, unlike the uniform treatment given to content by editors, its quality is checked only as often as someone in the user population sees fit. On the other hand, more content is often produced overall by wiki populations than by editors, and is often more timely.

Outside Wikipedia, Al-Hakim (2007), the business study of IQ criteria affecting innovation management, found that conciseness was not particularly important to workers in that context. However, in the emergency response domain, Bharosa et al. (2008) argued that asking workers to sort through too long/much information can be a hindrance to achieving prompt responses. A similar view – "Is the information to the point, [de]void of unnecessary elements?" (p. 920) – was expressed by Su et al. (2008), in the field of disaster management.

For fansites, this literature asks whether wiki-produced pages are more list-oriented / tedious, verbose, and poorly argued and organized than are edited pages, and whether errors persist longer on wikis than on edited sites. Also, the business view, that conciseness is of little importance, is probably more appropriate to the fansite context than the emergency response view, because the fansite context involves retail marketing, and because fans are probably not making momentarily crucial decisions based on the information.

### 2.5.4 Copyright statements & disclaimers

Discussed only briefly in the IQ literature, several medical authors have noticed when sites under study did not display copyright notices (Kihlstrom, 2001; Hoffman-Goetz and Clarke, 2000; Doupi and Van Der Lei, 1999; Shon and Musen, 1999). Studies by Fallis and Frické (2002) and Frické and Fallis (2004) have also concluded that copyright notices, on webpages pertaining to treating fever in children, statistically significantly correlate with the accuracy of information on those sites, as judged by subject-specific criteria developed by medical experts. However, the authors note that such indicators are not fool-proof, as inaccurate sites merely need to learn that accurate sites tend to have copyrights, and imitate them. Similarly, Buhi et al. (2010), a study that asked 34 young adults to evaluate the IQ of 177 websites about online sexual health in terms of 15 criteria, noted that government websites are not allowed to be copyrighted. Finally, Wikipedia's IQ criteria mandate that content with free/open-source licenses be used when available and adequate, and have a number of criteria regarding the inclusion of non-free content (Wikipedia, 2009d).

Disclaimers, as website IQ criteria, occur only in the medical literature, and include such statements as "not designed or intended to replace the relationship between the visitor and a medical care provider" (Kihlstrom, 2001, p. 66) and "this product has not been evaluated by the FDA [i.e., the US Food and Drug Administration]" (Veronin and Ramirez, 2000). As for confidential and privacy statements and policies, Galimberti and Jain (2000) noted on hysterectomy sites that patients' photographs and full names were often included. Similarly, Kihlstrom (2001) and Ogushi and Tatsumi (2000) both noticed when the medical sites they studied did not display clear confidentiality or privacy policies.

Copyright, disclaimer, and confidentiality statements exist, in some form, on all of the fansites under study. The interesting question is to what end they are used in this non-medical context.

### 2.5.5 Critical analyses & descriptive synopses

The presence and contextual completeness and appropriateness of descriptive synopses, as well as of critical analyses, are among the possibly relevant context-dependent IQ factors one might encounter on fansites. In the medical IQ literature, this takes the form of experts (usually the authors) evaluating whether all conventional and mandated information on the topic in question has been included on a website (e.g., Pandolfini and Bonati, 2002; Lissman and Boehnlein, 2001; Chen et al., 2000; Latthe et al., 2000a), and that detailed descriptions, possible causes of medical conditions, possible outcomes, and possible repercussions are given (Charnock and Shepperd, 2004). For example, when studying the quality of sites on menorrhagia (abnormal menstruation), in addition to judging general IQ features of the sites (e.g., cited sources, currency, etc.), Latthe et al. (2000c) also studied whether anti-inflammatory drugs, anti-fibrinolytics, contraceptive pills, progesterone, the intrauterine system, and surgery were mentioned. Similarly, in the fan context, although Bradley (2005) mentions that – in addition to the "unsanctioned scholarship" (i.e., critical analyses by fans) mentioned in §2.4.3 – descriptive reviews are especially "prevalent in fan publications about girls' series books," one must know more about both girls' series books and the milieu of fan publications about them, in order assess whether fans' reviews and analyses of such books are contextually complete.

As discussed in §1.3, in order to adequately assess the presence of contextual IQ criteria, one of several non-LIS literatures would need to be added to this review. In order to limit the project to a manageable size, and to keep its scope within LIS, only the IQ literature has been formally consulted for this project.

### 2.5.6 Trivial and inapplicable

**Statements of benefits and risks** are not often found in science fiction fansites, unlike on medical websites. Although viewing a TV show can affect one's state of mind and beliefs, can have various psychosomatic effects on the body, and can physically harm the body after prolonged or extreme (e.g., loud volume) exposure, exposing oneself to entertainment-oriented mass media is not usually considered a "treatment" by western *popular* culture, and so is not discussed as such by the authors of fansites. While there may be academic value in studying fans' emotional reactions to specific episodes, such reactions were only textually recorded in discussion forums on all of the sites in this study, which were separate from descriptive and analytical fan accounts of episode content, which were the subject of this project.

### 2.5.7 Completeness-related research sub-questions

The literature in this section suggested that the following research questions:

- **RC1:** *Author's financial agenda*: to what extent does each site detail its ownership, sources of funding, and affiliations?
- **RC2:** *Author's general purpose*: to what extent does each site detail its purpose,

primary interest, organizational type, and location?

- **RC3:** *Reasons for citing similars*: why do the fansites link to other fansites?
- **RC4:** *Length and mass collaboration as thorough*: to what extent does page length or number of authors correlate with other markers of quality?
- **RC5:** *Length and mass collaboration as inconcise, ineloquent*: to what extent do wiki-produced articles contain more lists of facts and trivia, longer texts and sentences, and less organization than editorially produced articles?
- **RC6:** *Presence of copyrights and disclaimers*: how do copyright statements and disclaimers differ across wiki and edited fansites?

## 2.6 Conclusion

This chapter critically surveyed the literature pertaining to artifact-based measures of the quality of amateur, collectively-produced information on the Web. Drawing on work from a variety of fields, including LIS, the most prominent variables cited by these works were identified and categorized according to those that deal with the potential artifactual fitness, representational accuracy, and representational completeness of the information. No known literature has done, in this context, exactly what this project proposes, namely to evaluate and compare the quality of descriptive and analytical information compiled by several large virtual populations of fans of a particular mainstream science fiction franchise. Because the works reviewed in this chapter survey similar contexts to the current project, only in different fields/topics, it is believed that the variables identified may be relevant to the current

project. Also, by investigating IQ variables from other contexts in this context, further meta-analyses comparing amateur information collections such as these across multiple domains could be facilitated.

In addition to the variables identified in this chapter, this general research program would benefit from consulting additional literatures outside LIS, which could allow the quality of more of the local-contextual information to be evaluated. For example, the information could be evaluated for the clarity with which it represents the realities of the Vancouver film industry, or the cultural references and narrative techniques used in the science fiction franchise under study, as well as for the degree to which the sites facilitate online fan practices of engaging with the mass media, as have been observed in other contexts.

# Chapter 3

# Methods

## 3.1 Introduction

This chapter discusses issues related to sampling, to operationalizing the research questions raised in the previous chapter, and to the spectrum of analytical techniques appropriate for answering the research questions as well as for otherwise exploring the data.

Section 3.2 discusses the rationale for the chosen group of websites, discusses how statistically principled sampling of public websites can be done in a responsible and legal way, describes the downloading and sampling techniques required for these specific sites, and describes the types and quantities of content obtained from each site under study. Section 3.3 describes how the sampling and analysis processes were organized so as to answer multiple research questions with the same data sets, describes how additional variables emerged from the sites' local contexts, and references a codebook of all variables used in every analysis, found in appendix A. Finally, §3.4 summarizes how artifactual and exploratory research methods differ from laboratory and confirmatory methods, summa-

rizes the methods most often employed by the IQ literature, and describes the spectrum of methods used in this project's analyses.

## 3.2 Population

### 3.2.1 Choice of websites

Sampling issues are important for assessing representativeness, and studies of multiple websites on the Internet usually share a common sampling dilemma, namely no authoritative index of the entire Web exists. Even if one were to have access to all of the search engine indices in existence, the unified set would be incomplete and would change too quickly for contemporary crawling software to keep track. Hence, studies of multiple websites invariably begin from some kind of *judgment sample*, often the returns of a search engine, or the pages compiled by a Web directory project, such as DMOZ.org. Once a website has been chosen for study, it may possess page or content indices that can be used for random sampling – and this is true for all of the sites under study here – though this is not true for every site on the Web.

The website sampling techniques and rationale encountered in the IQ literature can be summarized as follows. Many of the authors did not specify a sampling technique, usually because they were either studying only a single site (e.g., Wikipedia) or were surveying literature on IQ criteria in a non-empirical way. Of the empirical and multi-site studies, the vast majority (45 studies) identified search engine results as their sampling technique. Though the details of search engines' algorithms are proprietarily hidden, for the medi-

cal authors, the primary rationale for using them was because this is how most patients would find health information on the Web. For example, Selman et al. (2006) conducted a questionnaire survey "to ascertain the internet search strategy of a lay person who would potentially access internet information on cervical cancer treatment" (Methods, para. 1). The results of this survey prompted the researchers to use a variety of search engines, to use search phrases relevant to their subject matter, to filter out returns that seemed irrelevant to two independent coders, and to disregard pages not written in the language of the population under study (English, in this case). Other recent examples of this technique and rationale include: Rahnavardi et al. (2008), Pérez-López and Roncero (2006), and Lewiecki et al. (2006). Additionally, in medicine, institutions and associations often accumulate directories of sites that are known to offer reputable information on some topic. For example, Huang et al. (2005), a study of fertility clinic websites, chose 266 websites from a directory created by the Society for Assisted Reproductive Technology.

The current study used these techniques and rationales, and more, to choose its sample of websites. First, since no science fiction associations (e.g., Rensselaer Science Fiction Association or British Science Fiction Association) could be found that recommend fan-sites to the general public, most fans would probably find information about Stargate by searching the Web, using Web directories, or hearing about them at non-virtual fan conventions, a prominent one being ComicCon.

Hence, second, a search of the largest Web search engines (Google, Yahoo!, and Microsoft Live) was conducted for only the term "stargate". Discarding studio and network sites, as well as sites not about the Stargate media franchise (e.g., several IT companies use the Stargate name), the four sites under study had the following ranks on the search engines

just mentioned, respectively: Wikipedia ($3^{rd}$ on Google, $1^{st}$ on Yahoo!, $1^{st}$ on MSLive), IMDb (2, 2, 2), GateWorld (1, 3, 3), and Wikia (5, 5, 5). The fourth position was occupied by the Stargate Information Archive (2009), a small site containing only a news feed, TV schedule, and forum, which have not been updated since 17 March 2008.

Third, DMOZ.org was also searched, again using only the search term "stargate". Of 381 sites returned, GateWorld was number 1 and was the only entry given a preferential yellow star rating and special formatting by the site's editors. IMDb was $45^{th}$. The wiki sites did not place in the top 100, and the other sites were a mixture of studio, network, large corporate, and small fan sites.

Fourth, though many Stargate fansites exist, the number of sites must be limited, in order for the project to be feasible. Also, since this project aims to generalize to other large Web 2.0 sites, small fansites could be too idiosyncratic to include. Similarly, it was desired that the sites under study represent the two most common editorial models (edited vs. communal) often seen on fansites. A mixture of very large and general sites of both editorial types (Wikipedia and IMDb) with smaller and more specific sites of both types (Wikia and GateWorld) seemed to well-represent this spectrum. Finally, one may notice throughout the following sections and chapters that these four sites are well-aware of each other: they seem to have established content boundaries, and they are densely interlinked. All of these considerations suggest that these four sites represent the core large fansites about this media franchise.

### 3.2.2 Website sampling principles

In addition to how the four websites were chosen, one should also understand what constitutes the statistical "population" in this context, and how each website could be sampled. As with much research in information retrieval and corpus linguistics, the population is a corpus (i.e., a collection of documents/artifacts), rather than a group of people. In this study's case, the population of each of the four sites' artifacts include all of the webpages, images, and multimedia, contained in each site's database at a given moment, being literally a snapshot of the entire website. Section 3.2.3 describes how, and to what extent, each site's database was acquired. Unless doing an explicitly longitudinal study, the accepted research practice is to note on what date the data were collected, so that a future researcher could independently reconstruct the collection, using archives provided either by each site or by the Internet Archive (2009).

Regarding sampling bias, for sites that provide complete database dumps, the only sampling bias is a temporal one (i.e., are there daily, weekly, etc. regularities in the database that one would miss by only viewing a snapshot?). Without conducting a longitudinal study, the best one can do is to say on what date the data were downloaded, so that any researchers who are conducting a longitudinal study in the future could situate the current project's results within whatever temporal patterns they discover. Sites that do not provide their databases for download require crawling and "screen scraping"[1] of their webpages. The least biased crawling method is when a site provides some kind of unique index in their pages' URLs, such that one can reliably collect, or randomly sample, every page on

[1]Screen scraping is a computer science term that refers to downloading an entire webpage, and using a custom-written program to parse the page's content/data from out of the surrounding HTML markup.

the site. Without this, one must resort to more exotic samples, such as a "snowball sample" (Goodman, 1961), where, beginning from a set of random or somehow-important "seed" pages, all of the hyperlinks on those pages are followed recursively until the edge of the site is reached.

Finally, because each database is often too large to analyze in its entirety, one or more random samples of it can be taken and statistically characterized. Since the webpages are stored in relational database tables, a straightforward way to take random sub-samples of the tables' records is to assign a random number to each row from a uniform probability distribution, sort the table by those numbers, and remove any set of $n$ records from the table. If inferential statistical analyses are desired, the size of $n$ must be calculated using power analysis (Cohen, 1988). For exploratory analyses, the largest tractable sample, given computer hardware and human content-analytic coding constraints, is the most desirable, as it comes as close as is empirically possible to the population values.

Downloading a database dump prepared by a website is generally both technically and legally preferable, because it implies that the site's proprietors have packaged for widespread distribution those parts of the site that they do not mind being programmatically analyzed. Proprietors also often locate such dumps on servers that can handle large downloads without disrupting the primary website, and the code itself is free of the extraneous markup that would be downloaded and discarded for each page during screen scraping, thus saving bandwidth. If a dump is not available, and screen scraping is necessary, one must be certain that the website's administrators have not forbidden it either in the robots.txt file located in the public root directory of the website's server, or in the site's Terms of Service/Use statement. For the current project, the only site for which this was an

issue is IMDb, the details of which are discussed in §3.2.3.

### 3.2.3 Sampling these websites

#### Wikipedia

As an open-source community project, all revisions of all pages ever made on the English, Spanish, French, and German Wikipedias, including discussion and user profile pages, can be downloaded freely in SQL and XML formats, under the GNU Free Documentation License (Wikipedia, 2009j). The data are usually current to within the past week, and instructions are provided by the community as to how to obtain and query the database.

In early August 2009, the archive from 15 July of all English Wikipedia pages, not including discussion and user profile pages, which are not under study here, was downloaded in XML format. It contained approximately three million articles. The archive was a single XML file approximately 4 gigabytes (GB) in size, when compressed with the open-source algorithm possessing the highest compression ratio available (i.e., 7-zip, with "ultra" settings). A custom program was written by the researcher to read this file on a line-by-line basis, so as to avoid having to uncompress the entire file, keeping the records for only those pages that mentioned the term "stargate" (note: case insensitive). Five thousand seventy one (5,071) pages contained that term.

Wikipedia (and Wikia) pages are not sorted by their authors into a small set of categories comparable to those of IMDb and GateWorld (e.g., actors, titles, episodes, crew members, etc.). Also, it would be difficult for a computer to categorize pages in this way, or even to determine whether a page pertains to the Stargate media franchise – and not, for example,

to the Stargate Project run by the U.S. Federal Government from the 1970s to 1995 investigating the reality and potential applications of psychic phenomena (McMoneagle, 2006). Therefore, all 5,071 pages were individually and manually examined by the researcher, irrelevant pages were removed, and relevant pages were sorted into categories comparable to those found on IMDb and GateWorld. Only 1,564 pages were found to contain references to the Stargate media franchise. These data represent an exhaustive sample of all relevant Wikipedia pages, to the degree that one researcher could obtain such a sample.

Table 3.1 displays the categories and number of pages found for each category, sorted from most to fewest.

Table 3.1: Exhaustive sample sizes: Wikipedia

| Actors | Lists | Crew | Episodes | Characters | Authors | Peoples | General |
|---|---|---|---|---|---|---|---|
| 371 | 58 | 54 | 48 | 23 | 21 | 10 | 9 |
| Technology | Games | | | | | | |
| 6 | 4 | | | | | | |

The following are descriptions of what kinds of pages were sorted into each category by the researcher. Actor pages described actors who portray characters in one of the Stargate series. List pages contained lists of anything in the other pages (often technologies and peoples), but in a characteristic lengthy tabular format. Crew pages were about directors, writers, producers, and the like. Episode pages described individual episodes. Character pages described fictitious characters in one of the series. Author pages were about authors of Stargate-related books or comics. People pages were about groups of people, from small tribes and organizations to entire races. General pages were about entire series (e.g.,

a page entitled “Stargate SG-1”). Technology pages were about any type of technology, including: weapons, ships, computers, scientific instruments, and communication devices. Game pages were about relevant video, board, and arcade games.

One may notice that Wikipedia’s sample sizes are rather smaller than the other sites’, which are described in the following three sub-sections, and that the above page counts do not total 1,564. The unaccounted-for 960 pages each contained only one or two lines about the Stargate franchise, usually pointing out a cultural reference, such as that character Jack O’Neill makes frequent reference to The Simpsons TV show in Stargate SG-1, or that a real U.S. Air Force Chief of Staff (General Michael E. Ryan) once had a cameo on the show. These pages were not included in the analysis, because only one or two lines of these pages’ contents were actually about Stargate, and then only vaguely. Wikipedia pages generally only described those aspects of the franchise that have been noticed by mainstream society, such as the most popular actors, crew members, episodes, and characters.

Finally, in addition to data retrieved from the database dump, three research sub-questions required either querying Wikipedia’s MediaWiki application programming interface (API) or automatically downloading full HTML pages. The details of those queries and downloads are provided at the beginnings of the relevant Results sections, namely §4.3.1, §4.3.7, and §4.2.4.

**IMDb**

IMDb, as a subsidiary of Amazon, is the only site studied for this project that is owned by a large corporation. IMDb offers encyclopedic information that is donated to them from users as well as studios, and is compiled and modified by a team of editors. In total, IMDb

pages on individual films and TV programs can, but need not, contain as many as 95 unique data fields in 11 categories, any of which can be either edited or initialized by users (IMDb, 2009j).

As such, IMDb's content is considerably more legally restricted and difficult to access than the other sites. When users contribute content to IMDb, they agree to sign over their copyright to the company, which then owns the content, and licenses the content to other commercial entities for a minimum annual fee of $15,000 USD (IMDb, 2009k). However, they do provide a limited dataset for "personal and non-commercial" use, which is primarily intended for individuals who wish to have a searchable snapshot of the database for their own private use (IMDb, 2009c). This dataset contains only the records provided on titles (i.e., movie and TV episodes) and names (i.e., actors and crew) – not the information about fictitious characters or user demographics – available on the main site. Crawling and screen scraping are forbidden. Also, the dataset is delivered in a non-standard relational format (not SQL), such that its tables must be joined, indexed, and searched using custom programs.

The current project uses the free dataset on the understanding that no actual IMDb content will be republished in this dissertation, that their data will only be used for private analysis purposes, and only the results of those analyses will be published in a non-commercial way. The free dataset was obtained from IMDb's FTP servers (IMDb, 2009a) on 28 July 2009, and was indexed and searched using the Alternative Movie Database (AMDb), an open-source implementation of IMDb's proprietary query programs (Siebert, 2002). Because only names and titles were searchable, it was not possible to search for all records containing the term 'stargate'; a seed list of actor, crew, and title names was needed. Since

GateWorld contains extremely complete and itemized lists of each of these entities, the lists from GateWorld were used as input for searching the IMDb database. Hence, though this sample may not be exhaustive or random, it is as complete a sample as the most respected and dedicated non-communal fansite on the Web (i.e., GateWorld) could obtain. From that dataset, 612 relevant actor records, 369 title records, and 27 crew member records were obtained.

Additionally, as will be discussed in §3.3, several of the research questions required content and link analysis of complete IMDb pages, so as to characterize things such as numbers of inlinks, PageRanks, validation errors, numbers of pictures and videos, lists and tables, vendor types, source of evidence types, original research types, and textual themes. Those analyses were conducted in the manually laborious way common to content analysis, namely by human coders viewing a random sample of the possible public webpages in a Web browser, counting features and themes present on each page, and saving no more than a screenshot for later personal reference. This type of analysis was also used to study relevant character pages, which are not in the free dataset, again based on a seed list of 869 character names retrieved from GateWorld. No screen scraping or other data mining techniques were employed in these content analyses – including for determining pages' inlinks and PageRanks, which only involved looking up IMDb URLs in Yahoo and Google's servers, placing no load on IMDb – and no IMDb content will be published in relation to those analyses.

**stargate.wikia.com**

The Stargate wiki hosted at Wikia, which was begun by the founders of Wikipedia, maintains nearly identical policies and systems to Wikipedia. Database dumps of all articles are available (Wikia, 2009f), as is the MediaWiki API, and all textual content is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license (Wikia, 2009d).

On 8 August 2009, the archive of all pages, including user discussion and profile pages, was downloaded from the Stargate wiki. The XML database was relatively small (20 MB uncompressed), and only contained Stargate-related records, making unnecessary the filtering process required for Wikipedia. This also ensured that the sample of this site used for the current project is exhaustive. The database contained 14,363 pages. As with Wikipedia, and unlike GateWorld or IMDb, these pages were not classified by their creators according to a small controlled vocabulary. Hence, a considerable semi-automated effort was required in order to introduce this structure to the database. Table 3.2 lists the resulting numbers of pages in each category, from most to least.

Table 3.2: Exhaustive sample sizes: Wikia

| Characters | Places | Technologies | Peoples | Actors | Episodes | Books |
|---|---|---|---|---|---|---|
| 830 | 802 | 496 | 468 | 373 | 371 | 325 |
| Miscellaneous | Ships | Crew | Videos | Games | | |
| 258 | 121 | 60 | 25 | 16 | | |

The category names have the same meanings as for Wikipedia, with the exception of the miscellaneous category, which included: battles, scientific/natural phenomena (e.g., black holes), foods, languages, and mythologies/religions. Only a handful of these were cultural

references, such as references to foods or languages not original to the Stargate universe. Cultural references were a much smaller phenomenon on Wikia than on Wikipedia, so were not isolated and removed. Also, the one category on Wikia that does not occur on Wikipedia is 'places', which are pages about planets, buildings, cities, landmarks, and any other location that is not also a ship (e.g., the city of Atlantis is also a spaceship, so is categorized as a ship).

The remaining 10,218 pages, which were not used in this analysis because they either only serve an administrative function or are user discussion and profile pages, fell into the following categories. Category pages listed categories into which regular pages had been categorized by users. Disambiguation pages listed several options for ambiguous queries, such as when an episode had multiple parts and a user searched for the episode without specifying a part. Files pages provided metadata about a file that had been uploaded to the wiki, such as an image, sound, or video file. Help pages provided documentation on using the wiki. Redirect pages resolved synonymous searches to a single canonical URL, such as a search for "The Beachhead" might redirect to "Beachhead, The". Finally, user profile and talk pages introduced editors to one another and allowed them to discuss articles before making changes. Profile and talk pages were not included because this study is neither about user demographics nor collaboration processes on wikis.

Finally, as with Wikipedia, three research sub-questions required either querying Wikia's MediaWiki application programming interface (API) or automatically downloading full HTML pages. The details of those queries and downloads are provided at the beginnings of the relevant Results sections, namely §4.3.1, §4.3.7, and §4.2.4. Unlike Wikipedia, popular Wikia pages also contained user voting/rating information; this information existed for

792 pages on the wiki. For more details, see §4.3.8.

### GateWorld

Perhaps the most prominent Stargate-only fansite, GateWorld is owned by one man (Darren Sumner), is run by volunteers and contractors, is financed by donations and advertisements, and is not affiliated with the studio or networks. GateWorld's staff of volunteers include: four editors, three writers, eight forum moderators, one system administrator, and one graphic designer (GateWorld, 2009e). Though founded in 1999, GateWorld's guides and reviews go back to the beginning of the Stargate franchise (i.e., the 1994 film; GateWorld, 2009d).

No templates or detailed forms were provided by the editors for content submissions to GateWorld, and any templates used by the editors were not publicized. The only means of submitting content to the editors was via the "Write To Us" form (GateWorld, 2009f), which contained fields for the user's name, comment, and a link to the source of their information. That page encouraged users to submit comments and suggestions, to report news and events, to submit Letters to the Editor for publication, and to report comment violations and broken site features. It referred users with questions about Stargate either to an FAQ page or to the GateWorld Forum.

Nevertheless, content on the site did follow several obvious patterns. There was information that originated from the studio/producers or other companies, from the site's users, and from the site's editors. *Corporate information* included: show credits (air dates, writers, directors, cast, crew, lengths, and technical details), Nielsen cable and network syndication ratings for many episodes, promotional photos and merchandise, and signs of

interaction between the site's users and editors with the show's the producers (e.g., interviews with GateWorld's editors, and blog and forum posts from producers). *User-made information* primarily included episode votes (one vote per IP address) and transcripts. Many episodes' dialog and action were transcribed either by a user or a staff member, with the author's name given at the beginning of the text. Additional non-descriptive user-generated information, which was not used in this study, included user forum posts, and a section of the site dedicated to fan fiction. *Editor-made information* made up most of the descriptive and analytical text on the site, including: plot synopses and analyses, encyclopedia entries describing and interpreting the roles of characters and technologies in the series, numerical episode ratings, and quotes compiled from outside sources (e.g., from cast and crew members) about individual episodes.

As might be expected from a small non-profit business, the website had no Terms of Use statement, made no mention of screen scraping or data mining, only limited Web crawlers to accessing one page every 60 seconds (via a robots.txt file), gave its webpages predictable names, and provided no database downloads. Therefore, automatically downloading and screen scraping pages on the site was neither difficult nor unethical, so long as excessive load was not placed on their servers.

On 16 June 2009, all indexed pages containing the aforementioned information were downloaded using POSIX utilities at a pace slow enough not to burden the site's servers. The site's pages all appeared to be neatly classified by the site's editors into a small controlled vocabulary, and either index pages listing all of the files in a section were present, or all of the sections' pages followed easily predictable naming conventions. Hence, if the site's index pages were accurate, the sample should be exhaustive, and the classifications

should all be as accurate as are the site's editors, who are obsessively meticulous, would have them be. Table 3.3 lists the numbers of pages in each category, from most to least.

Table 3.3: Exhaustive sample sizes: GateWorld

| Characters | Miscellany | Technologies | Episodes | Planets | Races | Ships | Books |
|---|---|---|---|---|---|---|---|
| 1037 | 540 | 479 | 361 | 283 | 149 | 75 | 67 |
| Comics | Videogames | | | | | | |
| 28 | 11 | | | | | | |

Several of the categories chosen by the GateWorld editors are slightly different from those that emerged from the wiki sites. Primarily because of choosing the 'planets' label over the more generic 'places', and by choosing 'races' over the more generic 'peoples', the editors had to have an expanded 'miscellany' (their term) category for locations and groups of people that are not quite either planets or races. Also in their miscellany category are things that were categorized by the researcher as miscellaneous on Wikia, namely: battles, scientific/natural phenomena, foods, languages, and mythologies/religions. As with Wikia, references to non-Stargate culture were infrequent.

GateWorld's content bears a remarkable similarity to the Stargate Wikia wiki. Indeed the GateWorld editors recently moved all of their encyclopedic content from static HTML to a wiki, though it is not open to the public. Comparing such similar sites, which use different editorial paradigms, should be interesting.

## 3.3 Instruments

The primary role of the IQ literature reviewed for this project is to suggest variables in terms of which the IQ of these websites may be evaluated. The project's research questions were derived from the subset of IQ variables that are either relevant or non-trivial to study in this context. As the questions require certain research tasks as well as the presence of certain data, and since there are many research questions in this project, the sampling and analysis process was organized according to those research questions that have common requirements.

In addition to the variables prescribed by the IQ literature, a number of contextual variables were also considered. While many of these were mandated by the research questions to be found through exploratory, hermeneutic content analysis (e.g., the question asking what types of advertisements occur on the sites), others were included in order to avoid turning a blind eye to variation existing in the webpages that is not described in the IQ literature, in order to represent these fan sites as well as possible, and hopefully to make a theoretical contribution to the literature on IQ criteria. Due to the scarcity of IQ literature on fansites, many of these contextual variables have only been documented by the current project.

On the editor-controlled sites (IMDb and GateWorld), a number of contextual variables originate from the controlled vocabularies imposed on the content of those sites by the sites' editors. For instance, many GateWorld pages contain a link to a review written by an editor, and many IMDb pages link to six standard types of references, namely: official sites (run by the artist or studio), photo sites, video sites, sound sites, miscellaneous sites, and

showtime listings. On the wiki sites (Wikipedia and Wikia), regular expression POSIX utilities were used to sort and count occurrences of all of the header tags – surrounded by '=' characters in MediaWiki's markup language – in the sampled articles. Those headers that occurred in more than one article were manually sorted by the researcher into categories of synonyms, because users often used similar or similarly spelled words without checking for consistency with other articles, and the occurrence of any similar lexical form of a category label was counted as indicating the presence of that category in a webpage. For example, if one page contained a header named '==External links==', and another page contained a header named '==external link list==', and if the contents of those two sections on both pages appeared to be highly similar to the trained human eye (i.e., if both contained a list of links to external resources, as often occurs near the end of Wikipedia pages), those sections were both treated as instances of the same variable (i.e., links.external.count; see table A.2 in the appendices). Finally, the `style` utility used for calculating textual readability scores was also capable of outputting a variety of lexical word usage metrics, including: verb phrases, conjunctions, pronouns, and nominalizations. Although most of the contextual variables found in this study are generalizable beyond these websites, with respect to the current data, most required measurement using custom parsers, written to exploit textual cues that would only be present in the context of these websites.

Appendix A contains content analytic codebook tables describing all of the variables, whether prescribed or contextual, used in all of the analyses. The tables include: the variables' names and types, descriptions of their meanings, examples of their occurrence, and which research question(s) each variable was used to answer.

The following sub-sections describe the three types of samples that were assembled, as

well as the research questions and variables that each sample was used to address. Note that the question of exactly which variables were available for study in each section of each website was the topic of the research sub-question about consistency, and is covered in §4.3.6.

Finally, this section ends with a presentation of how all of the variables in Stvilia (2006) were operationalized in the current project's context.

### All pages

Any research sub-questions that could be answered by automated means were studied on as exhaustive of samples of all Stargate-related pages of every website as possible.

Prescribed variables were measured in the following ways. Standard Unix POSIX utilities – such as `sed`, `awk`, `grep`, `sort`, `cut`, and `wc` – were used to count the occurrences of variables indicated by a single text string, such as media content channels and hyperlinks, which reliably had standard file name extensions or protocol prefix abbreviations. The IQ literature did not provide a list of recommended media content types for which to search, so every page on each website was tested for every content type that could be found in any of the pages on any of the sites. The Web spider function of the `wget` utility was used to check hyperlinks for brokenness. The `style` utility is capable of calculating all of the readability metrics mentioned in the IQ literature, except for the Fry metric, for which only heuristic manual reference Fry Graph diagrams and computer utilities exist (Fry, 1977). Rather than manually consult that graph for thousands of webpage texts, the researcher re-expressed and rotated the graph's curve into a straight line, and wrote a program to calculate Fry scores for each text, based on that line. Inlinks were computed by automatically looking

up the page's URL in Yahoo's free Site Explorer API (Yahoo!, 2009). PageRanks were found with a tool written by Walker (2007), based on the WWW-Google-PageRank Perl CPAN library by Karaban (2009), for querying Google's servers. Finally, validation and accessibility errors were found with Raggett's HTML Tidy utility (Raggett, 2009).

Contextual variables were each measured in a way appropriate to each website's code layout and linguistic variations, using many custom parsing programs. While some of these variables fell within the same general types of variables as did the prescribed variables, several new variable types also emerged. For example, while there were more types of links available on these sites than were prescribed in representational-accuracy research sub-question 11, no research questions itemized the types of short and long text fields, dates, ratings, or page types that should be studied. Hence, the exact operationalization of the research questions, as well as several entire categories of variables, had to be empirically defined in context.

**Random pages**

For the cases when variables could not be measured in an automated way, but rather required a degree of human interpretation, random samples were taken of the full datasets, because the exhaustive samples were too large to be coded by hand in their entireties. Much statistical literature exists on minimum acceptable sample sizes for latent variable models, such as principal components analysis (PCA) and factor analysis, which is the modeling family appropriate for the relevant research questions. General recommendations typically either endorse heuristics, such as certain round numbers of observations, usually ranging from 100-500 (MacCallum et al., 1999), or subject-to-variable ratios, usually ranging from

5:1 (five observations for every one variable) to 20:1 (Garson, 2008). Also, although hypothesis and significance tests were not called-for by the research questions, one may use statistical power calculations for multiple regression and correlation tests, as PCA constructs correlative models of linear dimensions of variables (Hsieh et al., 1998). Finally, the notion from survey research of preserving a sufficient amount of variation from a population of a certain size could be relevant, since most of the samples are exhaustive (Rea and Parker, 2005).

Forty nine (49) variables emerged from a pilot hermeneutic content analysis of random pages from any section of all four sites. Hence, the 5:1 heuristic from the previous paragraph would recommend that at least 245 observations be obtained for a principal components analysis. Similarly, the sample size calculations for multiple regression – assuming an $\alpha$ level of 0.05, 49 predictors, an anticipated effect size of 0.15, and a desired power level of 0.8 – suggest a minimum allowable sample size of 235. Treating the problem as being closer to survey research – assuming a 5% margin of error, a 95% confidence level, and a 50% response distribution – the suggested sample sizes for the four sites are as follows: GateWorld 346, IMDb 320, Wikia 354, and Wikipedia 234. Since all of these values are near or above 235, to be conservative, the larger sample sizes suggested by the survey calculation were used in determining random sample sizes.

Finally, the reliability of the content analytic data was ensured in two ways. Following Krippendorff (2004, p. 215), a "test, re-test" procedure was employed to illuminate unstable/idiosyncratic coding decisions on the part of the primary coder. This was accomplished by ensuring that an initial pass of coding was finished early in the analysis process (September 2009), and that enough time had passed until the end of the analysis process

(February 2010) that any subtle or peculiar decisions made in the initial pass has been largely forgotten. This process forced the analyst to re-negotiate an understanding of how the codes were determined. Additionally, a second coder, who was otherwise not involved in the project, re-coded a random sample of 10% of the data using only the descriptions provided in the codebooks in appendix A. The researcher discussed all discrepancies with the second coder until perfect reliability was achieved. The clarifications obtained through this twofold process are reflected in the codebooks.

**Certain matched pages**

One of the research questions, regarding objectivity as impartiality, required that a sample of matching pages be assembled across all of the sites, in order to characterize to what degree accounts on the same topics agreed across the four sites. Actor pages were excluded from this sample, because GateWorld does not have actor pages (they all refer to IMDb), because many of the actor pages were seen in the random samples, and because actor pages on IMDb tended to be highly listy, atextual, and formulaic. Also, the title and character samples were quite small, because Wikipedia only had 48 title/episode pages and 23 character pages. Nevertheless, 32 matched title pages and 21 matched character pages were manually found across the four sites, and their content analysis proceeded in the same way as for the larger random sample of all pages. In this case, sample size and power calculations were not considered, because the research question only wanted to know, given some set of common topics to evaluate, how much do the four website's accounts agree? Therefore, one of the most well-reputed agreement metrics in content analysis (Herring, 2007; Hong, 2005), Krippendorf's $\alpha$ (Krippendorff, 2004), was applied to each set of four

observations, and interpretations were developed about the cases in which the sites usually did and did not agree (§4.3.4).

**Stvilia's instruments**

The variables from Stvilia (2006, see p. 278 for a list) were studied in this dissertation in the following ways. Due to the large amount of unusual vocabulary on science fiction sites (e.g., fictitious alien races), checking for spelling errors, Stvilia's recommended objective metric, would be a manually laborious task. Checking the factual accuracy of fansites' texts in context would also be extremely manually time-consuming, because each bit of in-show trivia would need to be investigated either in the studio's literature or by watching the show's episodes. Completeness in context was studied in terms of the data fields present across each site (§4.3.6), and should be evaluated more qualitatively in future research. Relevance as aboutness was determined by having a subject expert (i.e., the author) manually verify that each page on the sites that included content about topics other than Stargate (i.e., IMDb and Wikipedia) were actually about the Stargate media franchise, consulting as many alternate sources as necessary to clarify ambiguous cases, such as resources about minor characters or locations having unusual names (§§3.2.3 and 3.2.3). For cognitive complexity, all of the readability metrics and length metrics used by Stvilia were employed, and more, as well as a number of word usage variables (§4.2.5). Regarding currency, several aspects of 'age' were studied in §4.3.1. For volatility, though the information on the sites under study did not expire, but rather described an archival mass-media phenomenon, last-modification dates on the wiki sites could give some sense of the speed at which the sites' information can change. Naturalness was studied as the

amount of unique page sectioning/templating (§4.4.4). Precision/granularity was studied throughout this project in terms of levels of trivial vs. substantive detail. Regarding accessibility, though speed was not an issue for these large sites, available media formats and WCAG as well as HTML validation 1.0 errors were studied (§§4.2.2 and 4.2.4). Security was not an issue for page content, because all of the information was public, though user information security and privacy were compared qualitatively (§4.4.1). Verifiability was studied via types of citations and evidence available on the sites (§§4.3.2, 4.3.5, and 4.3.11). Structural consistency was also studied in terms of via sectioning/templating. Consistency of this type was high on the editor-controlled sites and lower on the wikis (§§4.4.4 and 4.2.1). Great variation in the language present in the wikis' data fields made automatically comparing fields across pages difficult. Semantic consistency was studied via textual themes (§4.3.4). Most such information on these sites was either unstructured or loosely structured, requiring time-consuming human comprehension to study in greater depth. Finally, cohesiveness and informativeness, both of which Stvilia recommends studying via inverse document frequency (IDF), would be too time-consuming to study using the IDF of individual words. Also, which frequently occurring Stargate-specific words to disregard as stopwords is unclear. Redundancy was studied statistically throughout the project, by exploring different types of patterns in the information. On-topic-ness should be high on the sites, because pages were usually short and their topics encyclopedically narrow (§§4.4.3 and 4.4.4), though a qualitative study would probably be required.

## 3.4 Procedures

### 3.4.1 From the literature

The analytical methods applicable to this project span the range of approaches often associated with content analysis (CA), Web data mining, and frequentist statistics. Content analysts primarily count occurrences of some variable in a text or corpus, and then explicate those counts either directly or with basic descriptive statistics. By contrast, data mining researchers and frequentist statisticians seek mathematical patterns both within and between observations and variables, even if the observations are made by content analytic means. This distinction between descriptive, distribution-oriented statistics and modeling statistics can be found in most any statistics textbook (e.g., Tabachnick and Fidell, 2007, p. 7).

Within model-oriented statistics, one should also note the difference between exploratory and confirmatory approaches. Whereas "...'exploratory data analysis' is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as for those we believe might be there" (Tukey, 1979), confirmatory analyses take descriptive statistics, "find reliable differences and relationships, and estimate population values for the reliable findings" (e.g., Tabachnick and Fidell, 2007, pp. 7-8).

For the current project, an exploratory approach is often more appropriate than an inferential, hypothesis-testing approach. The primary reason is that, because the phenomenon under study is so new and unknown, it would be difficult to formulate hypotheses that would be both specific and important to people native to this social context. As is often the case with social scientific research – which, to some degree, requires that the researcher be immersed in the local psycho-socio-cultural context, in order to avoid ethnocentrism –

more contextually rich findings can be had by being guided by the research questions than by testing predictions that were formulated before actually engaging with the data in context. This preference, and these reasons, for exploratory analysis also predominate in the IQ literature. The goal of the research is to illuminate the variables that are most important in this context, and to find initial estimates of those variables, rather than to seek perfect values for well-known variables.

Finally, in these discussions, an important distinction between computer scientific and statistical methods should be noted. The fields of computer science, machine learning, and artificial intelligence have a long history of attempting to take human analysts out of the processes by which information is coded and analyzed. Such automation has been done for reasons of engineering a more efficient or convenient society, for understanding the human mind, etc. This automation has largely been accomplished by developing heuristical algorithms, basically series of tasks, which, through past experimentation, have been found by researchers to produce intuitive results on typical input/data. Hence, such heuristics act as filters by which to test to what degree a dataset conforms to general social expectations. By contrast, frequentist statistical methods are based on mathematical frequency distributions (e.g., normal/Gaussian, Poisson, Gamma, etc.), allowing an analyst to characterize the distribution of a *man-made* dataset, and to find patterns in, and draw inferences upon, that dataset with a high degree of mathematical reliability. In order for the results of statistical analyses to be interpretable, a human observer must have produced and understood the contextual meanings and distributions of the variables in the dataset. The human analyst is integral to the process, and the results are not necessarily socially generic. Both heuristical and statistical methods exist for describing and modeling datasets. This distinction will be

made throughout the following sub-sections.

### 3.4.2 Descriptive analysis

Heuristical, descriptive methods are most often used for summarizing the perfunctory or structural aspects of websites. For example, one can run a website's markup (HTML) code through programs that examine it for violations of national or international accessibility standards, as did Llinas et al. (2008) to the Web portals of 32 hospital systems. As in von Danwitz et al. (1999), a webpage's compatibility with various browsers can also be tested by viewing the site in different browsers, for which several automated tools also exist. It is likewise possible to automatically scan a webpage for "broken" hyperlinks (i.e., links that point to a destination that is not, or is no longer, available), and to summarize the character encodings, multimedia files, and Web 2.0 features that it uses (e.g., Moran and Oliver, 2007; López-Ornelas et al., 2005; Leung, 2001).

Statistical, descriptive analyses, which make up the majority of the IQ literature, require more human intervention. The study by Fulda and Kwasik (2004) is a particularly good example of this approach. First, they chose a sample of resource and access library, state library, and hospital websites from the South Central Chapter Region of the Medical Library Association. Though one may fault the authors for neither justifying their sample size nor accounting for the transient nature of the Web, their approach does highlight the degree to which statistical samples of social scientific data are often dependent upon human interpretation. The two authors then chose IQ variables relevant to their context (e.g., ease of accessibility, funding sources, presence of sitemaps and search engines, and presence

of disclaimers), and hermeneutically operationalized those variables by inductively exploring and discussing how the variables manifest in their research context. They then visited each website and constructed a spreadsheet of counts of which sites possessed which IQ features. In their results section, they summarize this spreadsheet in the form of count totals and percentages, interpreting what those descriptive figures say about the sites under study. For example, "Among the hospital Websites, thirteen of twenty-five (52%) used fee-based, commercially prepared resources. The most popular were HealthVision and Cerner Multum. ... Health sciences libraries and state libraries did not incorporate those types of resources in their sites" (Results and discussion, para. 3). Though the variables may be coded in either a more perfunctory or a more interpretive way, the result, an observation-by-variable data matrix, from which one can characterize the distributions of the variables, is constant across all such analyses.

### 3.4.3 Exploratory analysis

As a relatively young field, computer scientific "data mining" research (often called "exploratory data analysis" in the statistics field) approach data without pre-conceived hypotheses. To some extent, clustering and machine learning algorithms have been developed from this perspective, usually with an emphasis on improving either the runtime efficiency of the algorithm or increasing its capability to distinguish among certain scenarios in hypothetical, or otherwise contrived, data (e.g., the standard corpuses in information retrieval). Though desirable from an engineering perspective, this approach has resulted in a multitude of ways of modeling data, but in little understanding of how to determine when a

certain method is most appropriate or robust for real-world datasets.

Heuristic usage of such algorithms often employs a try-and-see approach, whereby data are run through several algorithms, hoping for interesting results. Blumenstock (2008, see §2.3.2) is a good example of this approach in the IQ literature. Similarly, researchers using both computer scientific and statistical methods sometimes apply familiar metrics or models (e.g., precision and recall, the vector-space representation, cosine similarity, k-means clustering, simple linear equations for readability formulae, and least-squares regression) without first exploring whether their assumptions have been met, or whether they are most appropriate, for a given dataset and research question. For example, in the IQ literature, see the use of precision and recall by Adler and de Alfaro (2007) and Gaudinat et al. (2007), and the use of $\chi^2$ tests by Frické and Fallis (2004), Meric et al. (2002), Huh and Cude (2004), and Buhi et al. (2010). From a statistical perspective, such approaches are problematic for a number of reasons, including: the method may not be resistant to "localized misbehavior" in the data (e.g., outliers; Hoaglin et al., 1983, p. 2), it may not account well for residuals left over after the model has been fit, it may not analyze the data in a scale or orientation that simplifies the data's structure, and heuristic approaches can rely more on faith in certain methods or assumptions about the data than on empirical diagnostics.

Alternatively, some data mining researchers create new/ad-hoc equations for describing phenomena, or solving a problem, in a certain context. For example, see the metric in Hu et al. (2007) for measuring IQ in terms of users' collaborative authoring behavior, or the one in Kasal et al. (2005) using inlinking practices. A statistical objection to this approach is that the properties of such idiosyncratic models may be poorly understood, and may make comparison with the results of other studies difficult, because they may disregard the

large body of literature that exists about more familiar models in the long-established field of statistics.

Statistical exploratory analyses often employ diagnostic and re-expression techniques, the analyst's knowledge of the mathematics of many established methods, iterative model fitting, and an open mind. The results of such analyses represent high quality model recommendations for, and initial estimates of, the phenomenon and questions under study, which other researchers may refine through either different avenues of exploration or separate confirmatory studies.

In this dissertation, descriptive and diagnostic techniques are always reported first, and are consulted when attempting to appropriately scale the data and choose models. The latest robust modeling techniques have been employed whenever possible, and are identified throughout. Regression models were also always built one variable at a time, following Emerson and Hoaglin (2006), with all available statistics being consulted at each step. In this way, all models were 'built' rather than 'tested' or 'confirmed'.

### 3.4.4 Methods employed

The following is a summary, from least to most technical, of which analytical methods were employed for which research questions, and why each was appropriate. These methods were not chosen out of mere preference or convenience, but because they either are standard diagnostic techniques for exploratively assessing the distributions of datasets, or are indicated as being the most appropriate model for the research circumstance by the distributional diagnostics and research sub-questions.

### Qualitative examination

Four research sub-questions required qualitative examination, namely: the question of how the sites' navigation, search interfaces, sitemaps, help documentation, and similar sections compare; the question of to what extent each site details its ownership, sources of funding, and affiliations; the question of to what extent each site details its purpose, primary interest, organizational type, and location; and how copyright statements and disclaimers differ across wiki vs. edited sites. These are qualitative problems, because these issues are either part of a site's overall template or part of its administrative sections, so are comprised of a small number of pages, often made by administrators rather than users. Because the few pages are either repeated or referenced many times across the site, macro-scale variation in them does not occur, making a quantitative analysis uninteresting. Rather, a close, critical, and comparative reading of the few relevant pages on each site should be more revealing. The qualitative perspective adopted here is hermeneutic-phenomenological, because it is particularly well-suited to the interpretation of texts. For more on this perspective, see Ricoeur (1976).

### Descriptive statistics

Most of the quantitative questions benefited from knowing at least basic descriptive statistics, such as the famous "five number summary" (i.e., minimum, first quartile, median, third quartile, and maximum; Hoaglin et al., 1983, p. 35), along with the mean, standard deviation (spread of the data), skew (whether the data have a bulge on the left or right), and kurtosis (the peakiness or flatness of the data). Such, or similarly meaningful, figures, as

well as diagnostic plots, when appropriate, are given for every quantitative variable studied in this project, and are used to determine and interpret their distributional properties. As discussed in §3.4.2, content analysis typically only goes so far as descriptive analysis. Six research questions were able to be sufficiently answered in this way, without employing statistical models, namely: what are the distributions of broken links on the websites, what are the most common HTML validation and accessibility errors for each site, what visual styles occurred on each site, to where do links on the sites usually go and from where do inlinks usually originate, and why do the sites link to other fansites? Also, recall from §3.3 that one research question required the application of several common inter-rater reliability metrics, which are either descriptive statistics or heuristic models. For more on the exploratory use of descriptive statistics, see Hoaglin et al. (1983).

**Latent variable models**

Eight research questions seek categories of the variation common to a set of variables, implying some kind of latent variable model (e.g., factor analysis, principal components analysis, structural equation modeling, etc.). For artifactual data, principal components analysis (PCA) is often most appropriate, because measurement error and the momentary transience of digital artifacts is less of a concern – because artifacts can be stored in their entirety, and artifacts' properties are typically recorded either by a computer program or by trained coders – than when measuring human beings. Whereas factor analysis maintains an error term for each observed variable, with the understanding that the measurements of the variables are likely to contain unexplained variation, PCA is essentially a dimensionality reduction technique, which assumes that the data were measured well enough, and merely

reduces variables' co-variation to orthogonal linear dimensions in a multivariate space. The questions for which this method was appropriate included finding correlated groups of the following: media content types, formulaic readability scores, original research types, sources of evidence, contextual/emergent data fields both from each site and across sites, types of advertisements, the four types of episode ratings on GateWorld, and variables measuring the conciseness and eloquence of the texts. Additionally, pairwise correlations were found for fan ratings of the same episodes on GateWorld and Wikia. For more on PCA, see Basilevsky (1994).

Canonical correlation, a type of latent variable model, was required for one research question, namely how the four episode rating variables on GateWorld relate to the other IQ variables. Canonical correlation was appropriate here, because, essentially, two latent variable models are being regressed against each other: one finding common linear dimensions/categories in the dependent fan-rating variables, and one in the independent IQ variables. The results of the analysis allow groups of similar ratings by different types of raters (i.e., fans, the GateWorld staff, and the Nielsen company) to be related to groups of IQ variables. Though it is unlikely that people give certain ratings based on certain IQ characteristics – especially if the rating was not actually given on the website in question, as is the case with Nielsen ratings – it can be interesting to know, for example, that the Nielsen network and syndication scores tend to rate episodes highly that these fansites fill with promotional pictures, transcripts, and poor-quality HTML code. On the other hand, fans rate such episodes lowly, preferring instead episodes with pages that contain reviews, high PageRanks and inlink counts, and contain texts that are more readable and have more complex word usage. GateWorld editor ratings sit in the middle of this spectrum, slightly

favoring the Nielsen side. For more on canonical correlation, see Tabachnick and Fidell (2007, pp. 567-606).

Note that, in interpreting this study's PCAs and the canonical correlation in §4.3.9, a common cutoff correlation/loading of ±0.32 was used, because loading matrices contain correlations, and squared correlations measure overlapping variance, so 0.32 would equate to 10% overlapping variance (Tabachnick and Fidell, 2007, p. 587). By the same reasoning, a 0.71 loading (50% overlapping variance) would mean that the variable is an excellent measure of a factor, 0.63 (40%) very good, 0.55 (30%) good, 0.45 (20%) fair, and 0.32 (10%) poor (Comfrey and Lee, 1992). Also, for both the PCAs and regressions in this dissertation, model coefficients and loadings are not typically reported, because, as an exploratory study that is attempting to provide a first picture of this research context, the precise magnitudes of those factors and coefficients are of less importance than their relative magnitudes and signs, both for grouping similar variables together and for answering the research questions.

**Multiple regression**

Multiple regression, though more common among laboratory and survey researchers than among artifactual researchers, was also required for two research questions. The question of to what extent pages' PageRank values or inlink counts correlate with other IQ characteristics implies predicting the levels of PageRanks or inlinks in terms of the other IQ variables. PageRank and inlink variables are continuous over a large enough range that regular multiple regression can be appropriate. The same can be said of the question about to what extent page length or number of authors correlate with other markers of quality.

Note that, for these and all statistical models run for this project, including the principal components analyses mentioned in the previous paragraph, all continuous variables were transformed/re-expressed via a power transformation of the form $y = x^p$ for some value of p necessary to make the distribution of the transformed variable roughly symmetric (cf. Hoaglin et al., 1983, Ch 3, specifically pp. 98-104), and were centered and scaled, in order to make the variables as linear and comparable as possible. Also, all variables were thoroughly examined for outliers, missing data, and violations of assumptions of normality, homoscedasticity, and multicollinearity. In most cases throughout this dissertation, the exact re-expressions done are not reported, because finding precise model coefficients is not the goal of this exploratory study. Finally, because even a careful examination will not identify all outliers or violations of assumptions for the large data sets considered in this study, robust regression renders results that are more likely to be insensitive to slight departures from assumptions, and hence is preferred to conventional least squares regression. (Computationally, robust regressions are only slightly more involved, as algorithms for them rely on iteratively re-weighted least squares algorithms.) For more on robust multiple regression, see Hoaglin et al. (1983, 2006).

**Logistic regression**

Finally, polytomous, ordinal logistic regression was most appropriate for answering one research question, namely predicting fan ratings on Wikia in terms of the other IQ variables. Fan ratings on Wikia occur as integers from 1-5, making them both discrete and ordinal, which violates several assumptions of continuous models. These models were run multiple times using several different model-building paradigms and software packages, in order to

verify the results. For more on logistic regression, see Paolillo (2002).

## 3.5 Conclusion

Using standard website data mining and statistical sampling techniques, cross-sectional snapshots of the databases of the four core large fansites associated with the Stargate media franchise were collected and coded, both for variables prescribed by the IQ literature and for variables that emerged from the local context. For all websites except IMDb, which restricts access to its database in various ways, exhaustive samples were obtained of the populations of all materials relating to the Stargate franchise on these sites. For many of the research questions, which required only automated parsing of these digital artifacts, it was possible to code and analyze the entire population of artifacts, allowing this project to report the most authoritative results possible for these data. Those research questions that required a degree of human interpretation were conducted either on exhaustive sets of the relevant pages or on statistically significant samples of the full datasets. The quantitative datasets were subjected to the most statistically principled and robust methods available for answering the research questions, including a number of principal component analyses, robust multiple regressions, canonical correlations, and polytomous ordinal logistic regressions. The results of these and the qualitative analyses shall be reported in the following chapter, and discussed in chapter 5.

# Chapter 4

# Results

## 4.1 Introduction

Following the same order as the literature review in chapter 2, this chapter presents the technical analysis details and results for each research question. The findings from this chapter, and their implications, are discussed more conceptually in the following chapter (5).

## 4.2 Artifactual fitness IQ

### 4.2.1 Accessibility as simplicity

**AF1:** How do the sites' navigation, search interfaces, sitemaps, help documentation, and similar sections compare?

The pages and sections of pages relevant for answering this question are either few in number or are repeated verbatim across every page on each site. Hence, this question will

be answered by qualitative examination of the relevant pages and sections of pages from each site, which will each be discussed in the following sub-sections.

### GateWorld

GateWorld's navigation was via drop-down menus located atop each page. They expanded down to two levels below the homepage, which was sufficient to deliver one to list pages of specific titles, characters, or the like. The site's content could also be browsed by clicking through the site's blog and RSS feeds, a media player's playlist, a TV listings calendar with links to episode guide pages, thumbnails of recent additions to the image gallery and GateWorld Store, or by subscribing to an email newsletter. The site did not include a sitemap. Its search engine was powered by Google's Custom Search product, which returned results pages that were integrated within a simplified version of the site's template. No advanced search options were available.

The site's help documentation took three forms. An FAQ page answered questions on the following topics: obtaining Stargate merchandise, the status of the GateWorld site and its policies, contacting the cast and crew, questions often asked by people new to the franchise, broadcast practices by TV networks, how the stargate and iris devices work, who are the antagonists from the first plot arc (i.e., the Goa'uld), differences between the film and the TV series, and how to participate in the Stargate USENET Newsgroup. Fans with questions not answered on this page may encounter either the Forum section of the site or the Write to Us page (GateWorld, 2009f). The forum section had over 30,000 registered members and seven million posts by those users at the time of data collection, and was highlighted in the global navigation bar atop each page. The Write to Us page, by contrast,

was at the bottom of each page in a small font, and sent messages to the small team of editors. The top of the Write to Us page encouraged only those users who had information or questions specifically for the editors to use the form on that page, and directed everyone else to the FAQ and Forum.

**IMDb**

Stargate information on IMDb may be navigated by the sub-menus of the Movies and TV drop-down menus, though this is quite time-consuming. The shortest path to any Stargate information available through the Movies navigation menu required clicking and browsing through five pages listing genres, ways of refining the search, and popular titles of that type. Therefore, someone looking specifically for Stargate would probably use the site's proprietary search engine, which returns many relevant photos, videos, titles, keywords, production companies, actors and crew, and partial matches to even the basic query "stargate". The photo gallery gave pages of 48 thumbnail images each, with 479 images in all, which could also be viewed as a slideshow. The videos page listed 30 videos per page, with 195 videos in all, many of them full episodes. The titles results returned links to pages about each series (e.g., SG-1), not each episode. Once on a series page, the user could browse to individual seasons, characters, actors, or any of the other information studied about IMDb in this project. Using the query "stargate," the "names" section of the search results did not find any of the people affiliated with the Stargate franchise. Production companies results found companies with the name Stargate in their title, which is not the case for all of the companies affiliated with this franchise. Keyword results matched against the folksonomy of terms that users can add to each title, and can deliver one to Stargate title

pages with an additional click.

The site did have a Site Index section, which listed high-level categories of the site's content (e.g., Advertising, Awards, Birthdays, etc.). As with the main navigation, many clicks would be required to find Stargate-specific information from this page. However, users might come to this page for help with using the IMDb site in general, for which there is also a section of Help pages. Help pages were sorted into sections, a list of the 10 most common questions, a link to lists of FAQ pages, and indices of data fields available to users on the site and how to correctly enter data into them. Both the Site Index and the Help pages could also be searched via the site's proprietary search engine. Advanced search forms existed for titles, names (cast and crew), and collaborations between titles and/or names. As on GateWorld, message forums existed, though IMDb attached one to each name and title page, and included a list of six posts from within the past day at the end of those pages. It is also possible to send email to the company via a contact form, though, like GateWorld, many requests are placed before the form, asking users to use the Help sections and a Help message board before resorting to the form. Unlike GateWorld, one may also only use the form if one is logged-in.

**Wikia**

As a site devoted to only the Stargate franchise, like GateWorld, stargate.wikia.com's navigation was directly relevant to finding information about the franchise. A drop-down navigation menu on each page had tabs for each series and movie as well as other forms of relevant media (e.g., books and DVDs) and races of characters. Most of these menus descended to the level of seasons or general categories, though the Universe and Infinity (a

non-canon animated spin-off, which lasted only one season) tabs went to the level of individual episodes, the games menu went to two popular titles, and the human characters menu went to three races of humans (one a race of our progenitors called the Alterans, and two current races of humans in different galaxies). The homepage also contained categories of links to series and movies, media, and "Around the Universe" (cast, characters, planets, etc.). There was also a brief blog-formatted list of current events, a featured article, a paragraph of background information about the franchise, a featured quote from one of the shows, a featured image, and instructions for creating new articles. As with most MediaWikis, featured articles/images/quotes were chosen through user nominations, and much of the navigation structure of pages was organic, in that each page was littered with links to other pages, both throughout the text and in semi-structured sections, such as Infobox and See Also sections. A built-in search engine was also available from each page, with advanced search options available either from the bottom of results pages or from a standalone Search page. Finally, from the Special:SpecialPages page, one could access a variety of sitemap-like reports listing both all pages, pages with certain properties, and users and their permissions. The nearest page to a sitemap would probably be the Category tree, which was a script for generating hierarchical displays of page categories created by users.

Help pages, again as with most MediaWikis, existed on the page level, in the form of Talk pages, and on the site level, in the form of Forum and Help pages. Talk pages were attached to each page on the wiki, allowing users to discuss the content of that page, and took the standard MediaWiki form of a lengthy threaded text file with users signing their posts with either their usernames or IP addresses. Such pages were regularly used by users,

though only actual pages and unique author counts were studied for this project (§4.4.3), as it is not a project about user collaboration on wikis. The Forum pages of this wiki had been used very little, with only 47 topic threads having been posted since 29 February 2008, when the forum was created. There also existed a Web-based Freenode IRC chatroom and #stargate channel, which one could enter only after creating an account on the wiki. At least a handful of users could usually be found in that channel, indicating that it might have been more popular for general questions about the wiki than the forum. Registered users also had their own Profile and Talk pages, as well as lists of each user's contributions, on/about which to interact with other users. The administrators of the entire wiki could be contacted via a simple contact form page, called Special:Contact, which was open to the public, and encouraged users to submit questions and feedback, rather than trying to dissuade and restrict them, like the editor-controlled sites.

**Wikipedia**

Like IMDb, Wikipedia is not devoted only to the Stargate franchise. Furthermore, large MediaWikis, such as Wikipedia, generally lack the kind of imposed navigation menu structures that one finds on editor-controlled sites, or even on smaller or more topically restricted wikis, such as the drop-down navigation menu on the Stargate Wikia wiki. Instead, one may either use the Category or Portal (i.e., super-category and generally explanatory) pages that have emerged through collective user cataloging activity, or use the search engine and then browse organically through inter-page links. The homepage of Wikipedia also contains numerous featured resources, though, given the size of Wikipedia, it is highly unlikely that even something similar to one's desired page would be elevated to the homepage. Enter-

ing “stargate” in the search box, which is located on every page, takes one directly to a page about the Stargate franchise, at the top of which are links to the film, a page about the stargate device, and a “disambiguation” page that lists all of the Stargate series, films, and major games and comics, along with pages on other topics containing the term. The same Special:SpecialPages pages also existed on Wikipedia, though its purview covered all of Wikipedia, not only Stargate, so the pages it generates could be much too off-topic. Although some Stargate-related pages have the word stargate in their titles, this is not always true, and the Category:Stargate category is not exhaustively inclusive of all relevant pages. However, delimiting by that category, as with the Category tree tool, does generate a sitemap of 112 prominent Stargate pages.

Regarding help pages, Talk pages existed for each page, as on Wikia, and followed the same format. As on IMDb, a large and well-documented general Help section, covering all of Wikipedia, existed, including sections on FAQs, getting started, browsing and editing Wikipedia, linking practices, using media on pages, tracking changes, policies and guidelines, communicating with other users, community consensus procedures, templating procedures, account maintenance, technical information about the MediaWiki software, and where to ask and address different types of questions and complaints. No forum existed. However, as with Wikia, user profiles and personal Talk pages were available, and, if none of these avenues were sufficient for a user, there were also Freenode IRC channels available for live chat with other users and administrators. Finally, though it was possible for users to email the Wikipedia administration directly, including founder Jimmy Wales, via the Wikipedia:Contact_us page, that page said in large bold red letters that Wikipedia has no editorial board, that “Content is not the result of an editorial decision by the Wiki-

media Foundation or its staff," and that "Although you can contact Jimmy Wales via one of these links, he is not responsible for individual articles or the daily operations" (Wikipedia, 2009b).

### Conclusion

Over all of the sites, navigation most obviously varied by site size and scope. The smaller sites were focused only on the Stargate franchise, so could have navigation menu structures and media-rich content (e.g., image and video galleries) specific to that topic on the home-page. The larger sites necessarily had to employ larger navigation structures, and could not display relevant media-rich content before asking the user to either browse or search first. On IMDb, this meant many editor-made menus and sub-pages, as well as a user-generated folksonomy. On Wikipedia, the categorization structures (e.g., categories and portals) were entirely emergent. Navigation structures also typically only delivered the user to franchise- or series-level title pages, from which one must browse through organic or semi-structured links either to more specific pages or to pages of other types (e.g., pages about cast, crew, or characters). Search engines were available on every site, and could deliver one to more specific pages than could browsing. GateWorld used Google's Custom Search, which had no advanced search options. IMDb had its own proprietary search engine that accurately delivered users to series-level title pages, and could do advanced searches of titles, names, and collaborations. Both wikis had MediaWiki's open-source search engine, which also provided advanced search options. GateWorld could have provided a franchise-specific sitemap, but did not. IMDb made no attempt to offer a franchise-specific sitemap beyond their search engine's results. The wikis offered customized sitemap-like report generation

scripts, which relied on user-generated categories, in order to filter out irrelevant pages.

All of the sites created documentation that was intended and organized to answer common questions. For the editor-controlled sites, help pages primarily had the tone of fielding questions received from users, so that users would know how the editors wish them to use and contribute to the site. This also manifested in large forums on these sites, presumably aimed at trying to get users to answer each others' questions, rather than turning to the editors. However, the wiki sites focused more on basic instructions about how to use the website's software, and otherwise tried to facilitate discussion on the level of content creation. Though, as a large site with well-established social norms, Wikipedia needed more documentation advising users on how to produce content that would be respected by the community. The wikis also had no functional forums, but instead relied on live IRC chat, apparently to facilitate even more rapid, or interactive, discussion among users. Also fitting with these different approaches, the editor-controlled sites included many disclaimer messages before their contact forms, and IMDb restricted the form's use only to registered members, probably in an attempt to decrease the amount of questions directed at them. The wiki sites, by contrast, had contact forms that were open to the public, and even encouraged submissions, though Wikipedia made it clear that administrators had less power to effect change in content than did rallying the community.

### 4.2.2 Accessibility as digital divide

**AF2:** By how many, and which, channels can fansites' content be accessed?

This question was answered both by examining the distributions of the individual media

type variables present on each site as well as finding common dimensions of variance in all of the variables occurring on each site, using principal components analysis.

## Descriptive statistics

### *GateWorld*

On GateWorld, RSS, CSS, JavaScript, JPEG, and GIF variables were present. Every page, except those in the Omnipedia (i.e., characters, races, planets, technologies, and misc), had one RSS feed, linked to two CSS stylesheets, had 12 JavaScripts, and had 34-100 JPEG and GIF files. Omnipedia pages were simpler: they had no RSS feeds, linked to only one CSS file, most contained three JavaScripts, and had one or two JPEG and GIF files. Descriptive statistics for JPEG and GIF files in the various sections of GateWorld are given in table 4.1.

This table shows that Omnipedia pages rarely have many JPEGs, which is due to those pages having been constructed using HTML frames, such that the page template was not sampled for every page. Generally, only one or two pictures of the character or object being described on the Omnipedia page existed. This has since changed; the Omnipedia has been redone as a wiki, and uses the regular site template without frames. Books pages had consistent counts, because those pages had minimal content. Popular comics pages had more content, and link-wrapped images of the comics often linked to similar titles. Episodes and video games had large spreads, with popular titles having large numbers of screen captures. Regarding GIFs, Omnipedia pages were much the same as with JPEGs. Books and comics both used GIFs primary for interface buttons. Episodes and video games again had large spreads, due to many clear.gif spacer images being used, and by GIFs being

Table 4.1: Descriptive statistics: GateWorld JPEG and GIF images

| | JPEG | | | | | GIF | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | books | comics | eps | omni | vg | books | comics | eps | omni | vg |
| present | 100% | 100% | 100% | 67% | 100% | 100% | 100% | 100% | 87% | 100% |
| stdev | 0.31 | 2.76 | 14.38 | 0.51 | 18.8 | 1.36 | 1.31 | 10.39 | 0.75 | 10.55 |
| min | 33 | 34 | 25 | 0 | 34 | 42 | 39 | 42 | 0 | 39 |
| 25% | 34 | 40 | 37 | 0 | 36 | 47 | 43 | 85 | 1 | 40.5 |
| median | 34 | 43 | 53 | 1 | 39 | 47 | 44 | 86 | 2 | 42 |
| mean | 33.89 | 42 | 51.59 | 0.69 | 47.3 | 47.2 | 43.78 | 83.69 | 1.56 | 46.2 |
| 75% | 34 | 44 | 61 | 1 | 50 | 48 | 45 | 89 | 2 | 43.5 |
| max | 34 | 45 | 127 | 2 | 91 | 49 | 45 | 97 | 3 | 66 |
| skew | -2.62 | -1.04 | 0.44 | -0.33 | 1.78 | -1.65 | -1.77 | -3.05 | -1.01 | 1.68 |
| kurt | 5.01 | 0.94 | 1.36 | -0.82 | 2.59 | 4 | 5.61 | 9.13 | -0.02 | 1.22 |
| n | 67 | 28 | 361 | 2563 | 11 | 67 | 28 | 361 | 2563 | 11 |

used in the user poll's buttons.

Correlating JPEGs and GIFs for each section shows that Omnipedia pages are highly positively correlated (0.92), and that book and episode pages are moderately correlated (0.41 and 0.39). Since JPEGs are usually used for more complex images and photographs (e.g., banner images), and GIFs for small interface elements (e.g., buttons), this suggests that the interface becomes more complex on Omnipedia, book, and episode pages, as more content-related images are added to those pages. Comics and video game pages follow the opposite trend, though not to an extreme degree, with correlations of $r = -0.02$ and $r =$

−0.07, respectively. Per-section JPEG-JPEG and GIF-GIF correlations were not possible, because different observations/pages existed in each section.

***IMDb***

For IMDb, recall from §§3.2.3 and 3.3 that only name and title page information were available in non-HTML from the free dataset, that the actual HTML pages had to be viewed manually in a Web browser in order to comply with their restrictions on screen scraping, and that the choice was made to only examine title and character pages, because IMDb actor pages were primarily listy/formulaic and used the same template as the rest of the IMDb site. The HTML-based IMDb findings reported in this chapter are based only on the certain matched pages sample (n = 32 for title pages and 21 for character pages).

RSS was not present on the IMDb pages examined, though every page linked to another page that provided site-wide (i.e., non-Stargate) RSS feeds. CSS, JavaScript, GIF, and PNG files were present on every IMDb page examined, and JPEGs were present on every title page. Every page referenced 7-8 CSS pages. Character pages reliably contained 10 JavaScripts and title pages 14-15. JPEGs were used on IMDb pages only near the top, when featuring thumbnails of a few photos from the image galleries. Most character pages contained none, though occasionally they had as many as 13 ($mean = 0.24, median = 0$). Title pages were the same, though most had 1-5 featured images ($mean = 3.63, median = 2$). Character pages' templates routinely contained 31 GIF images, without variance. Title pages were generally more content-rich and variable, containing the following spectrum of GIF variance: min 34, first quartile 40, median 41, mean 41.69, third quartile 43, max 60, standard deviation 2.76, skew 1.71, and kurtosis 10.73. Every page had the same 1-2 PNG

images: the IMDb logo, and the Amazon.com logo.

Regarding correlations, only title pages showed a marked JPEG-GIF correlation of $r = 0.33$, with character pages at $r = 0.01$. Like GateWorld, this probably suggests that title pages' interfaces become more complex as more photographic images are added. Character pages were considerably less content-rich and variable, but might suggest the same pattern.

### *Wikia*

Wikia, like Wikipedia, uses an HTML page template that is applied site-wide via PHP. RSS and Atom feeds are part of the template, and they show each page's recent changes. A CSS theme called Monaco, which is a modification of Wikipedia's Monobook theme, is used consistently across all of the pages. There are 39 JavaScripts in the template, including scripts for drop-down menus, Google Analytics, Xaw (the X Window System Athena widget set), and quantserve.com analytics. The template includes no JPEGs, though some graphical advertisements may be JPEGs. The template also contained 17 GIFs, primary for blank spacers and buttons. Finally, four PNG images were in the template, as background and logo images, probably because of PNG's high color space and more flexible transparency capabilities than GIFs.

In actual page content, JPEGs occurred in the following page types, with the following frequencies: episodes 87%, videos 60%, omnipediac (i.e., omnipedia-like) pages 55%, books 51%, games 40%, actors 20%, and crew 8%. JPEGs were primarily used for title banners and headshots for episodes and people, with no more than four occurring per page. Like GateWorld, popular episodes had as many as 16 screen captures. Also, though most omnipediac pages contained only the usual handful of JPEGs (mean: 1.73, median: 1),

a small number contained lengthy lists of stargate addresses, technologies, members of certain groups/races, etc. One such page had 196 JPEGs.

Only one page, titled "Military rank" contained GIF images, specifically two images showing tables of enlisted and officer rank insignias of the US Air Force. One or two PNG images occurred in a small number of actor, book, episode, and omnipediac pages. These were usually photos of someone, such as actors at a convention, a picture of a book or CD-ROM cover, or screen captures. One book page linked to a Flash animation gallery of Stargate props. 57% of episode pages linked to 272 PDFs, usually of screenplays and transcripts. Finally, the Stargate film's page linked to a QuickTime MOV of the trailer.

Regarding pairwise correlations between these variables, the only positive correlation was $r = 0.22$ between PNGs and PDFs on episode pages. Apparently, episode pages where people have gone to the trouble of posting a special photo (e.g., a lossless PNG) are also somewhat more likely to contain a PDF of supporting material. JPEGs were also slightly negatively correlated with GIFs on book pages ($r = -0.05$), and PNGs ($r = -0.09$) and PDFs ($r = -0.03$) on episode pages, possibly indicating that JPEGs were less common on pages with either more complex interfaces or more high-quality media files.

### *Wikipedia*

Wikipedia's template also contained RSS and Atom feeds showing recent page changes, and used the familiar Monobook theme. Fifteen (15) JavaScripts related to dynamic interface elements, their Usability Initiative, and server administration. No JPEGs occurred in the template, though possibly in the funding banners atop each page.

In actual page content, JPEGs occurred in the following proportions: 100% of general,

peoples, and technology pages; episodes 98%; characters 95%; games 67%; actors 38%; crew 23%; authors 20%; and lists 16%. As with Wikia, most pages had fewer than four JPEGs. However, technology pages ranged from 4-10, and general/franchise-level pages from 2-9, indicating that such pages always show pictures of their subjects. List and character pages also had higher maximums (12 and 10, respectively) than most.

Only three GIF images occurred in the sample, namely: a stargate iris/shield animation on a technology page, and two military medals on pages describing military personnel related to the franchise. PNGs occurred more frequently than on Wikia, in the following proportions: peoples pages 100%, technology 80%, games 33%, lists 16%, chars 14%, general 13%, episodes 4%, and actors 2%. Also, no more than six PNGs occurred on any page. This indicates that PNGs are primary used for reference images of peoples and technologies. PDFs occurred only in small numbers, and were used in much the same way as on Wikia. One author page linked to a PDF of a magazine issue; one list had a PDF of award nominations; nine actors pages linked to PDF supporting material (e.g., a vitae); 21 episodes linked to PDF screenplays; and three general pages linked to PDF posters from the TV network's advertising. SVG images occurred on 20% of actor pages, on 15% of crew pages, and on one author's page, in the form of generic "replace this image" filler images, for when no public domain photo existed for that person. Five percent of characters had SVG images representing their race/group (e.g., the brands/tattoos seared onto their subjects by Go'auld system lords), and 20% of technology pages contained SVGs (e.g., of the stargate chevron glyphs). Only one page, that of "Bill Nye" ("the science guy"), contained an mp3, and it was of a podcast. Also, the pages for "Stargate Atlantis" and its score's composer, "Joel Goldsmith," linked to an Ogg Vorbis file of the Atlantis theme.

Positive correlations were high between image pairs for a number of page types. JPEGs and PNGs were correlated on character ($r = 0.46$), list ($r = 0.75$), and technology ($r = 0.75$) pages, indicating that photographic reference images for these subjects are common in either format. JPEGs and PDFs were correlated on author ($r = 0.75$) and general ($r = 0.42$) pages, as were PNGs and PDFs on general ($r = 0.8$) pages, suggesting a similar point as on Wikia, that pages that have been given more attention probably receive both many photographic images and many supporting PDF documents. Finally, JPEGs and SVGs were correlated on character ($r = 0.89$) pages, as were PNGs and SVGs on character ($r = 0.55$) pages, suggesting that characters that warrant a photograph also often warrant an SVG image indicating their race/group. One notable negative correlation also occurred – JPEGs and SVGs on crew ($r = -0.22$) pages – probably because crew members are likely to have photographic portrait images, but unlikely to receive images typically only given to characters.

**Modeling results: Principal components analysis**

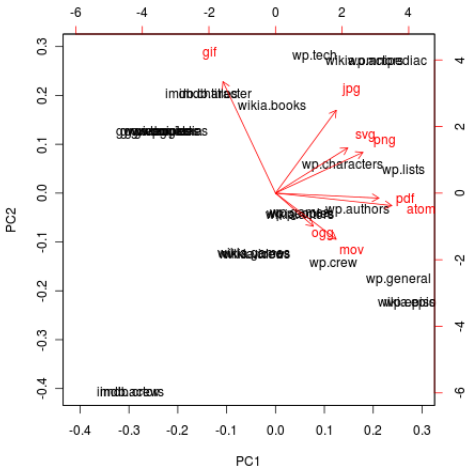
In addition to descriptive statistics of each variable, in order to identify any multivariate correlative structures in the data, as well as to confirm the previous findings, a principal components analysis (PCA) was conducted of the binary incidence matrix of which media variables (e.g., jpg, png, etc.) occurred on which site sub-sections (e.g., gateworld.books, wikia.actors, etc.). In that analysis, the first three eigenvalues were greater than 1, with isotropic variance clearly beginning at principal component (PC) 4; hence, the first four dimensions are worth interpreting. Figure 4.1 shows biplots of (a) PCs 1-2, (b) PCs 2-3, and (c) PCs 3-4. The PCAs and plots in this section were computed with the `prcomp` and

`biplot` functions in `R`.

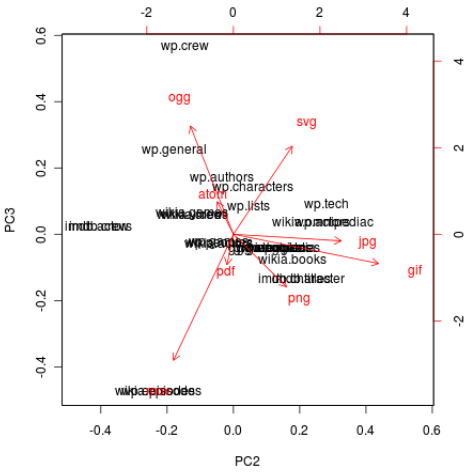
In PC1, Wikia episode pages as well as Wikipedia episode, list, and general pages are contrasted against IMDb actor and crew pages as well as most GateWorld pages. The former are associated with almost all of the media formats, whereas the latter only GIFs. This suggests that the starkest distinction was along editorial lines, with the editor-controlled sites prolifically using GIF images in their interfaces, which is a somewhat antiquated (i.e., pre-CSS and PNG) design technique.

PC2 distinguishes which site sections used which types of media. In the positive direction, Wikipedia technology and actor pages as well as Wikia omnipediac pages were associated with JPGs, whereas IMDb title and character pages had many GIFs. This is the same stark editorial contrast as seen in both the descriptive findings and PC1. Moving increasingly in the negative direction, Wikipedia character and list pages had SVG and PNG images, consistent with the correlation results. Wikipedia list, actor, episode, and author pages were known for PDFs and Atom feeds, consistent with the descriptive results. And Wikipedia general and crew pages as well as Wikia episode pages OGG and MOV files, which were the only pages containing those file types. These first two PCs account for the majority of the variance in the data.

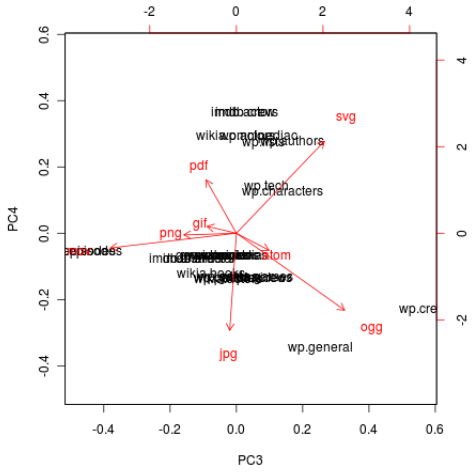
In PC3, an association between Wikipedia crew, general, author, and character pages with OGG and SVG files is evident. This is set in contrast to Wikia episode pages being associated with MOVs. As in PC2, Wikipedia crew and general pages were the only ones containing OGG files. Also, confirming the descriptive results, SVGs were common on Wikipedia character and people-related pages. Finally, MOVs on Wikia were indeed a unique occurrence.




(a) PCs 1-2




(b) PCs 2-3




(c) PCs 3-4

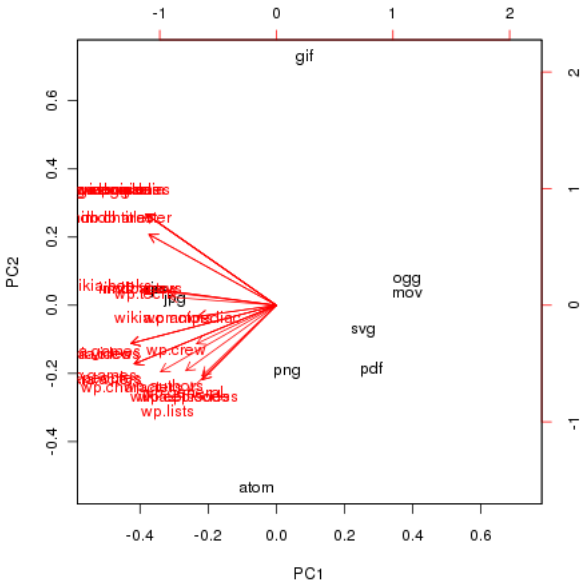
Figure 4.1: PCA biplots: Media types and website sections

Finally, in PC4, IMDb actor and crew pages as well as Wikia omnipediac and Wikipedia actor, author, and list pages are associated with SVGs and PDFs. As no IMDb page included either SVG or PDF images, this is rather strange; perhaps this is due to GIFs also loading in this direction. The Wikipedia page types are more expected. This is in contrast to Wikipedia general and crew pages being associated with OGG and JPG files, which also confirm the descriptive findings. These last two PCs identify relationships that are less common in the dataset. The overlap in results with some of the more prominent relationships could be due to the small and binary nature of the matrix being analyzed.

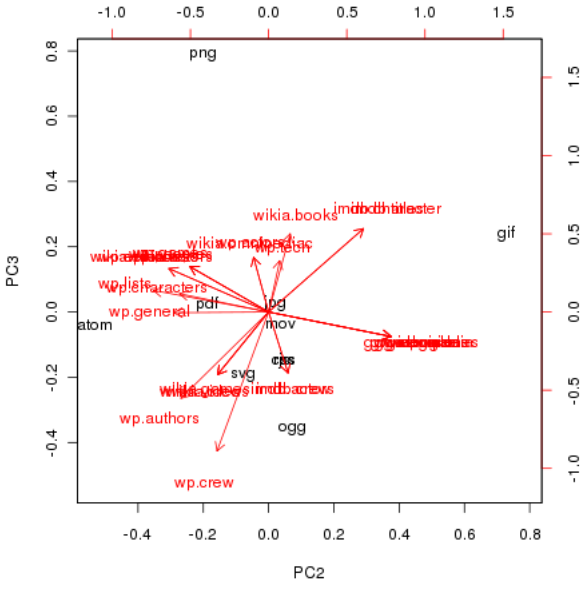
To confirm these results in another way, the transpose of the incidence matrix from the previous analysis was taken, and another PCA was conducted on that matrix. In this case, the first two eigenvalues were greater than 1, and the third showed a clear beginning to an area of isotropic variance. Figure 4.2 gives biplots for this secondary analysis.

In this analysis, as is common in PCA, the first component shows things that are more vs. less common. JavaScripts, CSS files, RSS feeds, and JPGs are more common overall than MOV, OGG, PDF, and SVG files, which is consistent with the descriptive results. In the second PC, the same pattern as observed in the first PC of the previous analysis is again visible, namely: IMDb and GateWorld pages use GIF and RSS files, and the wikis use Atom, PNG, and PDF files. Finally, the third, isotropic component showed infrequent associations, such as Wikipedia crew and authors having SVGs, Wikipedia lists having PDFs, and PNGs on Wikia actor, book, and episode pages.

These analyses indicate a general divide along editorial lines: GateWorld and IMDb use media content similarly, as do the wikis.




(a) PCs 1-2




(b) PCs 2-3

Figure 4.2: PCA secondary biplots: Media types and website sections

### 4.2.3 Accessibility as availability

**AF3:** Do any of the webpages contain broken links, and how many?

This question can be answered with basic counts and investigation. GateWorld's pages did not contain any broken links. A number of their links to YouTube and Amazon appeared broken, though this was due to those sites' Web servers prohibiting Web spidering programs from checking the integrity of links. When visited manually in a Web browser, the URLs were not broken. IMDb's pages similarly contained no broken links. This suggests that these sites routinely scan their pages for broken links, and that the editors always either investigate or remove them.

The wiki sites, however, did contain many broken links, perhaps because they cannot rely on their users/editors to perfectly correct broken links. Wikia actors pages had two broken links to a fansite on a related franchise. One Wikia book page linked to a closed Geocities account. Episode pages included many broken links, namely: 43 to the SyFy network, 27 to GateWorld reviews, 34 to ad index or no-longer-existing fansites, and one to a page on Joseph Mallozzi's blog. Games pages contained: four to a game-developing company's site, two to a defunct fansite (stargate-tcg.co.uk), two to a page on a gaming magazine's site, and two to an MMORPG company's site. Finally, all of the many Omnipediac pages contained only two broken links to a page located on an offline SG-1 role-playing game producer's site.

Despite having many fewer Stargate-related pages than Wikia, Wikipedia's pages contained many more, and greater variety of, broken links, possibly suggesting a less devoted and/or more chaotic user base. Unlike Wikia, broken links on Wikipedia rarely repeated.

Wikipedia's actor pages, for example, contained 40 broken links, all of which were unique. Their destination types followed the site's overall distribution of link destinations, which is discussed in §4.3.11. Author pages contained 10 broken links: three to interviews on a book company's site, two to a Dr. Who convention, two to BBC news articles, one to a novella archive, one to a sci-fi convention in Newfoundland, and one to a user's personal site. Character pages had eight broken links, all to defunct cast profile pages on the SyFy channel's website. Crew pages had six: four to articles in the mass news media, one to an online audition site, and one to a cast member's personal site. Finally, episode pages contained four broken links, all of which went to episode guides on the SyFy channel's website.

This again suggests that the sites' differences are primary due to different editorial approaches.

### 4.2.4 Accessibility as standards-compliance

**AF4:** To what extent do the pages HTML and CSS code comply with popular accessibility guidelines?

This question can be answered in two ways: through counting accessibility errors on the sites' sections overall, and through counting and describing which errors occur most on which sites' sections.

Since scanning for, and counting, both HTML validation errors and accessibility errors on webpages is trivial with Raggett's HTML Tidy utility (Raggett, 2009), both types of errors will be described in the following results. Also note that only the W3C's Web

Content Accessibility Guidelines (WCAG) 1.0 were evaluated. This was because version 2.0, which was released in December 2008 and somewhat conflicts with 1.0, has not yet received widespread use, nor has permeated across most accessibility testing tools, including Tidy. Also, the guidelines from the US Congress's Section 508 Amendment to the Rehabilitation Act of 1973 were not evaluated, because that amendment applies only to federal government websites. Other countries' accessibility guidelines were not evaluated, because all of the websites under study are based in the United States. Finally, be aware that, whereas the HTML validator identifies actual errors in HTML markup code, some accessibility issues can only be judged to be correct by a human expert (e.g., whether the site uses CSS as the W3C intended, or whether JavaScripts avoid flashing in ways that could cause epileptic seizures). Hence, WCAG validators tend to produce a considerable amount of "make sure that" messages, to remind the user to check that certain aspects of their code are accessible, even though those aspects of the code may be accessible. This will inflate the numbers of WCAG error counts in the following findings.

Also, unlike the editor-controlled sites, both wiki sites imposed a single HTML template on all of their pages, rather than developing different templates for different page types. This means that the variation in errors due to each section's interface had been made constant, leaving only the variation due to user content. Additionally, users of the wikis entered content using the MediaWiki markup language, which the system translated into XHTML code. This means that the system's capability to generate valid and accessible HTML markup from different types of MediaWiki markup is being observed, rather than the ability of end users to enter valid and accessible code. The Wikia and Wikipedia database dumps included the MediaWiki code, not processed XHTML code, and the wiki

sites' robots.txt files limited crawling activity to one page request per second. Therefore, for just this and the research sub-question discussed in §4.2.6, all of the HTML versions of the Stargate-related pages on Wikia and Wikipedia were downloaded at a rate no faster than one page per second. The results in this section were found using those datasets.

### Descriptive statistics

Descriptive statistics for HTML and WCAG errors for the sections of **GateWorld** can be found in table 4.2.

Table 4.2: Descriptive statistics: GateWorld HTML and WCAG errors

| | **HTML** | | | | | **WCAG** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | books | comics | eps | omni | vg | books | comics | eps | omni | vg |
| stdev | 10.35 | 6.06 | 60.03 | 3.35 | 6.59 | 10.57 | 10.67 | 78.69 | 15.4 | 74.7 |
| min | 32 | 30 | 39 | 3 | 31 | 307 | 288 | 297 | 8 | 286 |
| 25% | 33 | 32 | 92.75 | 4 | 35 | 336 | 332 | 555.5 | 34 | 301.5 |
| median | 35 | 33 | 113 | 6 | 38 | 336 | 334 | 581 | 40 | 335 |
| mean | 39.61 | 36.93 | 120.51 | 6.32 | 39.5 | 338.68 | 333.26 | 566.59 | 37.66 | 362.3 |
| 75% | 43 | 41 | 133.25 | 8 | 42.5 | 346 | 336.5 | 617 | 45 | 430.5 |
| max | 81 | 50 | 748 | 76 | 52 | 358 | 347 | 695 | 234 | 487 |
| skew | 2.27 | 0.84 | 4.41 | 9.22 | 0.83 | -0.67 | -3.02 | -1.83 | 1.76 | 0.48 |
| kurt | 5.12 | -0.59 | 38.25 | 144.83 | -0.04 | 1.12 | 12.75 | 3.59 | 24.73 | -1.49 |
| n | 67 | 28 | 361 | 2563 | 11 | 67 | 28 | 361 | 2563 | 11 |

For HTML validation errors, these figures say that books, comics, episode, and video games pages always had at least 30 errors, but usually fewer than 80, the latter of which was also true for omnipedia pages. Some omnipedia pages had very few errors, probably

because some had very little content. Content-rich episode pages, on the other hand, can have many hundreds of errors. Episode and omnipedia pages both had very definite (leptokurtotic) means, though, for episodes, the mean was about 120 and, for the omnipedia, about 6. Books had a moderately definite mean about 40 errors, and comics and video pages had fairly ambiguous (platykurtotic) means in the upper 30s. The distributions of all except comics and video pages were skewed to the left, indicating that pages with relatively high numbers of errors were not the norm.

For WCAG accessibility errors, these figures say that all sections except omnipedia pages had at least about 300 errors, with some episode pages going as high as 700. Some omnipedia pages had as few as eight errors. Book, comic, and episode pages were mildly right-skewed, indicating that most of these pages tended to have relatively higher error counts. Video games and omnipedia pages followed the opposite pattern, usually having fewer errors. These differences could again be due to pages with more complex contents presenting more opportunities for editors to make errors.

Descriptive statistics for **IMDb**'s HTML and WCAG errors can be found in table 4.3.

Like GateWorld, in IMDb's figures, the majority of pages have many errors of both types, the more complex pages (i.e., title pages) have higher error counts of both types than the pages with usually less content (i.e., character pages), there are more WCAG errors than HTML validation errors, and all of the sections are slightly left-skewed. This suggests that both IMDb sections follow the most common trend also seen on GateWorld, namely that more complex pages tend to have more errors, and only a few exceptionally lengthy/complex pages accumulate many errors.

Statistics describing HTML validation and WCAG errors for **Wikia**'s sections can be

Table 4.3: Descriptive statistics: IMDb HTML and WCAG errors

| | **HTML** | | **WCAG** | |
|---|---|---|---|---|
| | characters | titles | characters | titles |
| stdev | 2.97 | 8.63 | 5.22 | 19.54 |
| min | 19 | 35 | 172 | 235 |
| 25% | 20 | 52 | 177 | 289 |
| median | 20 | 57 | 177 | 301 |
| mean | 20.72 | 57.23 | 178.4 | 299.64 |
| 75% | 20 | 60 | 178 | 311 |
| max | 44 | 123 | 223 | 386 |
| skew | 4.76 | 1.88 | 5.01 | 0.16 |
| kurt | 22.19 | 9.96 | 27.12 | 2 |
| n | 42 | 12 | 42 | 12 |

found in tables 4.4 and 4.5.

These Wikia figures show many of the same trends seen on GateWorld and IMDb. Page length is generally proportional to number of errors, which actually makes sense more for the wiki sites than the editor-controlled sites, because all user-created content is being passed through the same XHTML-generating filter. Since the filter is programmed to consistently produce certain output from certain input, which may contain validation and accessibility errors, the more output it produces, the more errors there will be. On Wikia, book, episode, and omnipediac pages were often the longest, because topics such as

Table 4.4: Descriptive statistics: Wikia HTML errors

| | **actors** | **books** | **comics** | **crew** | **eps** | **games** | **omni** | **videos** |
|---|---|---|---|---|---|---|---|---|
| stdev | 4.25 | 14.88 | 3.63 | 4.56 | 11.13 | 6.47 | 7.71 | 5.69 |
| min | 13 | 11 | 19 | 19 | 19 | 16 | 17 | 19 |
| 25% | 19 | 20 | 24 | 19 | 33 | 19 | 19 | 20 |
| median | 24 | 24 | 28 | 19 | 33 | 19 | 24 | 28 |
| mean | 23.03 | 26.49 | 25.12 | 21.43 | 35.29 | 22.25 | 24.67 | 26.72 |
| 75% | 24 | 28 | 28 | 21.5 | 35 | 24 | 26 | 32 |
| max | 55 | 222 | 28 | 32 | 169 | 39 | 307 | 32 |
| skew | 2.31 | 8.65 | -0.84 | 1.66 | 7.3 | 1.8 | 19.59 | -0.33 |
| kurt | 12.26 | 101.5 | -0.85 | 1.17 | 70.93 | 2.5 | 674.79 | -1.81 |
| n | 373 | 325 | 24 | 60 | 371 | 16 | 2975 | 25 |

books and episodes probably contain more content for fans to document than, for example, a DVD box set. While this was also true for GateWorld's pages, it appears that Wikia users may have been more motivated (or have felt less encumbered by copyrights) to create lengthy documentation for books than were GateWorld's editors. The lengthy types of pages were also the most positively skewed, suggesting that, for Wikia too, only a few pages generate large numbers of errors. However, like GateWorld and not IMDb, WCAG errors for all pages except episodes and omnipediacs were skewed slightly to the right, with actors, crew, and games being most extremely so. This may indicate that smaller fansites can have difficulty keeping accessibility errors low, especially on more obscure pages.

Table 4.5: Descriptive statistics: Wikia WCAG errors

| | **actors** | **books** | **comics** | **crew** | **eps** | **games** | **omni** | **videos** |
|---|---|---|---|---|---|---|---|---|
| stdev | 24.9 | 40.75 | 23.81 | 49.41 | 42.84 | 81.79 | 73.69 | 78.05 |
| min | 126 | 26 | 344 | 107 | 235 | 75 | 282 | 306 |
| 25% | 352 | 358 | 381 | 346 | 485 | 349 | 363 | 353 |
| median | 377 | 388 | 404 | 346 | 494 | 353 | 381 | 472 |
| mean | 373.5 | 385.7 | 389.4 | 360.4 | 499.1 | 352.1 | 390.7 | 417.7 |
| 75% | 382 | 403.5 | 406 | 361.2 | 510 | 368 | 395.8 | 476 |
| max | 477 | 625 | 427 | 466 | 744 | 479 | 3359 | 563 |
| skew | -1.87 | -0.69 | -0.84 | -1.7 | 0.28 | -2.62 | 25.04 | -0.24 |
| kurt | 26.92 | 24.96 | -0.34 | 12.56 | 8 | 10.03 | 977.12 | -1.32 |
| n | 373 | 325 | 24 | 60 | 371 | 16 | 2975 | 25 |

Descriptive statistics for **Wikipedia**'s HTML and WCAG errors can be found in tables 4.6 and 4.7.

First, note that the abbreviation "N.E.O." is used in the Wikipedia tables for cases when Not Enough Observations were available to calculate a statistic. Though the samples were exhaustive, one should always be wary of findings with particularly small sample sizes.

The largest/longest pages on Wikipedia were those for actors, general pages (e.g., about entire series), and list pages, because Wikipedia's focus is on documenting topics at an abstract level. Consistent with the other sites, these longer page types tended to have more HTML validation errors, and most of the sections were left-skewed, as is the norm on the

Table 4.6: Descriptive statistics: Wikipedia HTML errors

| | **actors** | **authors** | **crew** | **eps** | **games** | **general** | **lists** | **omni** |
|---|---|---|---|---|---|---|---|---|
| stdev | 22.87 | 9.93 | 16.54 | 12.94 | 14.85 | 89.96 | 63.85 | 24.51 |
| min | 2 | 2 | 2 | 4 | 12 | 9 | 2 | 5 |
| 25% | 4 | 2 | 4 | 9 | 17.25 | 29.5 | 36 | 8.75 |
| median | 6 | 3 | 6 | 16 | 22.5 | 68 | 68 | 21.5 |
| mean | 13.57 | 7.65 | 12.3 | 19.81 | 22.5 | 83.88 | 80.88 | 28.92 |
| 75% | 14 | 8 | 12 | 27 | 27.75 | 85 | 128 | 40.5 |
| max | 259 | 39 | 78 | 58 | 33 | 294 | 246 | 87 |
| skew | 5.65 | 2.14 | 2.7 | 1.1 | N.E.O. | 2.22 | 0.68 | 1.07 |
| kurt | 44.82 | 4.43 | 7.37 | 0.56 | N.E.O. | 5.57 | -0.28 | 0.1 |
| n | 371 | 21 | 54 | 48 | 4 | 9 | 58 | 39 |

other sites as well. The unusually high minimum values for games is due to there only being two pages in that category: one a general page about all Stargate-related games, and the other a lengthy page about the popular Stargate Worlds game. Also, with the exception of the episodes WCAG errors, which are very slightly right-skewed, all of those errors follow the large site, left-skewed pattern, as does IMDb.

Finally, at every level of the five number summaries, the wikis had a little more than half the number of HTML validation errors than the editor-controlled sites, though the division with respect to accessibility errors was according to site size, with the larger sites having around 25% fewer WCAG errors than the smaller sites. IMDb did well at capping

Table 4.7: Descriptive statistics: Wikipedia WCAG errors

| | **actors** | **authors** | **crew** | **eps** | **games** | **general** | **lists** | **omni** |
|---|---|---|---|---|---|---|---|---|
| stdev | 65.95 | 34.44 | 54.78 | 59.07 | 38.18 | 158.81 | 367.74 | 106.13 |
| min | 81 | 81 | 91 | 140 | 355 | 344 | 86 | 295 |
| 25% | 108 | 99.75 | 99 | 351 | 368.5 | 437 | 439.2 | 355.2 |
| median | 122 | 114 | 117 | 368 | 382 | 491 | 519.5 | 369 |
| mean | 140.4 | 120.6 | 132.3 | 380.9 | 382 | 523.5 | 630.6 | 401.1 |
| 75% | 145.2 | 137.5 | 142 | 418.5 | 395.5 | 578.8 | 849 | 416 |
| max | 527 | 240 | 374 | 518 | 409 | 838 | 1475 | 947 |
| skew | 3.57 | 2.29 | 2.76 | -0.84 | N.E.O. | 1.07 | 1.12 | 4.17 |
| kurt | 14.79 | 7.25 | 8.79 | 5.25 | N.E.O. | 1.35 | 0.76 | 20.76 |
| n | 371 | 21 | 54 | 48 | 4 | 9 | 58 | 39 |

the upper limit of their HTML errors, though this could be due to their limiting the sizes of their pages. GateWorld consistently had the most HTML validation errors, and also made no attempt to conform to any W3C markup standard. Regarding accessibility, the wikis again had the best minimum error values, though IMDb and Wikipedia emerged as consistently having the fewest WCAG errors. Overall, Wikia's pages had the most WCAG errors, with GateWorld consistently having around 20% fewer.

### Ten most frequent errors

So that one may also see which errors were most common on each site, the error files generated by Tidy were automatically combined, sorted, and each error type counted, using POSIX utilities, both for each site sub-section and for each site overall. The following four tables, 4.8 - 4.11, present the 10 most frequent errors – whether validation- or accessibility related – with each table being followed by discussion of the meanings of the error texts as well as any notable within-site variations.

Table 4.8: Ten most common markup errors: GateWorld

| count | error text |
|---|---|
| 44,602 | Access: [2.1.1.1]: ensure information not conveyed through color alone (image). |
| 40,304 | Access: [6.1.1.3]: style sheets require testing (style attribute). |
| 34,567 | Access: [7.1.1.5]: remove flicker (animated gif). |
| 31,044 | Access: [1.1.1.1]: <img> missing ‘alt’ text. |
| 22,436 | Access: [1.1.2.1]: <img> missing ‘longdesc’ and d-link. |
| 17,314 | Access: [5.5.2.1]: <table> missing <caption>. |
| 17,314 | Access: [5.5.1.1]: <table> missing summary. |
| 16,411 | Access: [11.2.1.5]: replace deprecated html <font>. |
| 16,014 | Access: [9.3.1.3]: <script> not keyboard accessible (onClick). |
| 14,073 | Access: [5.3.1.1]: verify layout tables linearize properly. |

These errors, all of which pertain to accessibility, say the following: the site should not use text coloring to convey information that is conveyed no other way, administrators are encouraged to check that the site’s CSS rules work correctly, the site uses GIF images that

may flash and cause epileptic seizures, its images and tables do not have either descriptive text or metadata records that a blind person's screen reading program could see, it uses the antiquated "font" tag instead of CSS, there are JavaScripts that only work if one is able to use a mouse, and it uses tables instead of CSS positioning for page layout. Essentially, the site was designed in a manner typical of about 10 years ago, before CSS became common, and it assumes that the user can use a mouse.

The only section having different most common errors was the omnipedia, which was designed using HTML frames (another antiquated technology for most websites), such that the usual page template was lost on those pages. On those pages, the errors about images missing alt and summary attributes, and about flashing GIFs, were replaced errors against using too short/cryptic of link texts, and with ensuring that JavaScript features can be accessed via other means and do not flicker.

IMDb shared a number of common errors with GateWorld, all of which were accessibility related. Only errors five, six, and height in the previous table are new, and they all refer to similar issues. On sites using JavaScript, one should provide textual equivalents to features otherwise only available via dynamic scripts, which either the user or their browser/screen-reader may not be capable of manipulating. The first new error (fifth error, table 4.9) instructs the site's developer to test that their JavaScript content can be accessed via other means. The second means that any textual equivalents to scripts should be updated whenever the script is updated. The third says that textual equivalents for every JavaScript should be placed inside <noscript> tags. The only within-site variation to these errors was on the titles page, when an error about including metadata records for every image was included.

Table 4.9: Ten most common markup errors: IMDb

| count | error text |
|---|---|
| 51,746 | Access: [9.3.1.3]: <script> not keyboard accessible (onClick). |
| 18,847 | Access: [2.1.1.1]: ensure information not conveyed through color alone (image). |
| 17,701 | Access: [8.1.1.1]: ensure programmatic objects are accessible (script). |
| 17,701 | Access: [7.1.1.1]: remove flicker (script). |
| 17,701 | Access: [6.3.1.1]: programmatic objects require testing (script). |
| 17,701 | Access: [6.2.2.2]: text equivalents require updating (script). |
| 17,701 | Access: [2.1.1.4]: ensure information not conveyed through color alone (script). |
| 17,701 | Access: [1.1.10.1]: <script> missing <noscript> section. |
| 17,439 | Access: [13.1.1.1]: link text not meaningful. |
| 16,014 | Access: [1.1.2.1]: <img> missing 'longdesc' and d-link. |

All of Wikia's top 10 errors have been described for previous sites. No notable variation occurs in this list across the site's sub-sections.

Wikipedia's top 10 errors also contain little new or different, other than the invalid character warning (third error), which means that a textual character made its way onto many pages that does not occur in the Unicode character set used on every page (i.e., UTF-8). The troublesome character is an emdash, which is probably auto-inserted by Wikipedia's XHTML generator, since it occurs multiple times in every file. The only variations to this pattern are that the peoples pages were warned not to use tables without summary texts for screen readers, the games pages were warned for using lengthy link texts (e.g., making an entire sentence into a link), and the technology pages were warned for using images

Table 4.10: Ten most common markup errors: Wikia

| count | error text |
|---|---|
| 147,291 | Access: [8.1.1.1]: ensure programmatic objects are accessible (script). |
| 147,291 | Access: [7.1.1.1]: remove flicker (script). |
| 147,291 | Access: [6.3.1.1]: programmatic objects require testing (script). |
| 147,291 | Access: [6.2.2.2]: text equivalents require updating (script). |
| 147,291 | Access: [2.1.1.4]: ensure information not conveyed through color alone (script). |
| 139,549 | Access: [1.1.10.1]: <script> missing <noscript> section. |
| 116,229 | Access: [6.1.1.3]: style sheets require testing (style attribute). |
| 99,417 | Access: [2.1.1.1]: ensure information not conveyed through color alone (image). |
| 81,121 | Access: [7.1.1.5]: remove flicker (animated gif). |
| 72,328 | Access: [1.1.2.1]: <img> missing 'longdesc' and d-link. |

without summary texts for screen readers.

Remarkable similarity occurs in these lists. Accessibility errors are clearly more common than validation errors, and the same few errors occur consistently. The most common accessibility challenge faced by these sites is to either make their JavaScripts accessible via multiple user interface devices, or to provide textual alternatives to those scripts. Secondary challenges are to ensure that all graphical and tabular content is accompanied by textual descriptions, which would be primarily of importance to blind users, and that images do not flicker, which would be of most concern to epileptic users.

Table 4.11: Ten most common markup errors: Wikipedia

| count | error text |
|---|---|
| 44,916 | Access: [6.1.1.3]: style sheets require testing (style attribute). |
| 20,951 | Access: [13.1.1.1]: link text not meaningful. |
| 5,945 | Warning: replacing invalid character code 128 (emdash) |
| 5,739 | Access: [8.1.1.1]: ensure programmatic objects are accessible (script). |
| 5,739 | Access: [7.1.1.1]: remove flicker (script). |
| 5,739 | Access: [6.3.1.1]: programmatic objects require testing (script). |
| 5,739 | Access: [6.2.2.2]: text equivalents require updating (script). |
| 5,739 | Access: [2.1.1.4]: ensure information not conveyed through color alone (script). |
| 5,739 | Access: [1.1.10.1]: <script> missing <noscript> section. |
| 5,310 | Access: [6.1.1.1]: style sheets require testing (link). |

### 4.2.5 Readability as formulaic

**AF5:** What is the formulaic reading level of these pages, using different metrics?

The phrase "these pages" in the question presents a categorization problem, namely: which pages? There are far too many pages on each site to give individual attention to each. One might assume that the categories used by either the sites' creators or the researcher (e.g., actors, episodes, books, etc.) adequately capture the types of writing that exist on those pages. However, what if there exist multiple writing types within a category, or what if, for example, the writing of some actor pages are similar to some episode pages? To solve this problem of typifying writing types across an entire website, since readability measures are essentially heuristic formulae combining basic word usage measures (e.g.,

word and syllable counts) into a metric, it seems logical to use the co-variance in the 24 word usage variables that the `style` program also calculates for each text (cf. §3.3 and appendix A, table A.9) as a way of finding writing types. Also, by including the readability variables in the same analysis, one could say which readability metrics tend to score which writing types highly and which lowly. As with many artifactual dimensionality reduction problems, a latent variable model, such as principal components analysis, would be most appropriate for this task (see §3.4.4 for discussion).

Every block of text on every page of every website was automatically scanned, and readability scores using nearly all of the metrics mentioned in the IQ literature were calculated. One metric, the Lexile Framework for Reading, which is a proprietary formula owned by MetaMetrics, Inc., was not available, because requests to the company for affordable access to the proprietary software that they require for computing Lexile scores went unanswered. Instead, the reputable Lix measure by Björnsson (1968, 1983), which the IQ literature does not mention, was used. As explained in §3.3, the Fry metric was calculated by re-expressing and rotating the Fry Readability Graph, so as to derive a formula, the details of which can be found in appendix B. Finally, for all four PCAs discussed below, varimax (orthogonal) rotation was chosen for rotating the principal component axes, because, after computing promax (oblique) rotations and the correlations between components, correlations between all pairs of components on each site were found to be on the order of $10^{-17}$ (i.e., very near zero). This strongly suggests that orthogonal rotations are satisfactory for these data. The PCAs in this section were computed with the `prcomp`, `varimax`, and `promax` functions in R. A power value (m) of 3, the middle of the recommended range, was used for promax.

The following sub-sections answer the research question first at an overall descriptive level, then at the level of modeling co-variance structures within each website.

### Descriptive statistics

Table 4.12 displays descriptive statistics for each readability metric, averaged (using the mean) across all pages collected from GateWorld. All collected pages contained blocks of text on GateWorld.

Table 4.12: Readability averages: GateWorld ($n = 3,030$)

| | **Kincaid** | **ARI** | **CL** | **Flesch** | **Fog** | **Lix** | **SMOG** | **Fry** |
|---|---|---|---|---|---|---|---|---|
| stdev | 3.02 | 3.53 | 3.53 | 18.16 | 3.88 | 11.5 | 3.23 | 3.26 |
| min | 0.98 | 1.44 | 1.6 | 9.98 | 0.76 | 7.34 | 0.6 | 0.2 |
| 25% | 6.11 | 6.78 | 10.31 | 58.19 | 8.59 | 34.9 | 8.52 | 4.9 |
| med | 7.62 | 8.66 | 11.67 | 66.81 | 10.23 | 39.19 | 9.6 | 7.6 |
| mean | 7.61 | 8.67 | 11.3 | 62.03 | 10.05 | 38.39 | 9.19 | 6.95 |
| 75% | 9.57 | 11.03 | 13.52 | 73.64 | 12.33 | 44.3 | 11.09 | 9.05 |
| max | 15.14 | 17.1 | 19.42 | 87.02 | 19.38 | 64.46 | 16.08 | 11.8 |
| skew | -0.29 | -0.08 | -1.32 | -1.66 | -0.79 | -0.85 | -1.46 | -0.9 |
| kurt | 0.75 | 0.44 | 3.13 | 3.31 | 1.48 | 3.31 | 3.14 | 0.76 |

From this table, one can say that the range of readability scores spans from the very low to very high ranges of each metric, that most of the pages' texts would be categorized between the seventh and nineth grade levels, and that all of the scores show a slight rightward

skew, indicating that somewhat more pages are at higher grade levels than lower ones.

Table 4.13 shows the corresponding figures for IMDb. Fortunately, there were a number of texts available in the actor, crew, and title records in the free IMDb dataset. Note that 61% (610/1008) of collected IMDb pages were empty of any textual records, essentially pages that had titles and possibly lists of links, but had yet to have blocks of textual content added to them. Such pages had readability scores of zero, and the following statistics do not include them.

Table 4.13: Readability averages: IMDb ($n$ = 398)

| | **Kincaid** | **ARI** | **CL** | **Flesch** | **Fog** | **Lix** | **SMOG** | **Fry** |
|---|---|---|---|---|---|---|---|---|
| stdev | 2.34 | 3.01 | 1.63 | 9.18 | 2.61 | 7.24 | 1.7 | 2.86 |
| min | 2.3 | 1.8 | 7.3 | 39.7 | 3.9 | 18.5 | 3 | 1 |
| 25% | 6.4 | 7.1 | 10.7 | 61.4 | 9 | 35.6 | 8.73 | 7 |
| med | 8.1 | 9.2 | 11.55 | 67.05 | 10.8 | 40.3 | 9.9 | 8 |
| mean | 8.17 | 9.27 | 11.66 | 67.62 | 10.79 | 40.46 | 9.9 | 7.21 |
| 75% | 9.6 | 10.98 | 12.6 | 74.3 | 12.4 | 45.38 | 11 | 9 |
| max | 15.6 | 18.7 | 17.4 | 94.1 | 19.3 | 61.3 | 15.6 | 12 |
| skew | 0.33 | 0.35 | 0.26 | 0.02 | 0.3 | 0.08 | -0.02 | -0.8 |
| kurt | 0.05 | 0.03 | 0.66 | -0.11 | 0.19 | 0.07 | 0.66 | -0.13 |

This table shows that IMDb pages also span the range of possible readability scores, though fewer pages have very low scores than on GateWorld. The median and mean values are also slightly higher, but rounding the most common scores to integers reveals the same

seventh to nineth grade levels. There is very little skew in any of the distributions.

Table 4.14 shows the statistics for Wikia. Of Wikia's pages, 24% (978/4144) were empty of blocks of text, so were not included.

Table 4.14: Readability averages: Wikia ($n = 3,166$)

| | **Kincaid** | **ARI** | **CL** | **Flesch** | **Fog** | **Lix** | **SMOG** | **Fry** |
|---|---|---|---|---|---|---|---|---|
| stdev | 2.49 | 3.23 | 2.55 | 11.58 | 2.86 | 8.18 | 2 | 2.69 |
| min | 1.4 | 0.1 | 5.2 | 12.5 | 3 | 14.8 | 3 | 1 |
| 25% | 6 | 6.6 | 10.9 | 59.6 | 8.4 | 34.2 | 8.5 | 7 |
| med | 7.5 | 8.5 | 12.4 | 67.7 | 10.2 | 39.2 | 9.7 | 9 |
| mean | 7.65 | 8.69 | 12.62 | 66.68 | 10.31 | 39.42 | 9.66 | 8.64 |
| 75% | 9.1 | 10.52 | 14 | 74.4 | 12 | 44.3 | 10.7 | 11 |
| max | 44.1 | 56 | 43.4 | 100 | 45.8 | 125 | 19.4 | 12 |
| skew | 1.61 | 1.64 | 1.5 | -0.55 | 1.01 | 0.69 | -0.29 | -1.03 |
| kurt | 15.76 | 16.47 | 10.34 | 1.06 | 8.3 | 4.49 | 2.05 | 0.91 |

Wikia's texts had readability scores with similar ranges as the previous sites, though their means and medians were slightly higher: between the eighth and nineth grade levels. Kincaid, ARI, Coleman-Liau, Fog, and Lix scores were all slightly left skewed, and had quite definite peak values around their averages. Flesch, SMOG, and Fry were slightly right skewed, with ambiguous peaks. Though it is difficult to speculate, based on only these variables and their heuristic equations, why certain metrics grouped together, the principal components analysis of these and other word usage variables in the next sub-section should

make such interpretations both more possible and more empirical.

Table 4.15 displays the descriptive statistics for Wikipedia. Four percent (22/522) of pages were empty of blocks of text, so were not included.

Table 4.15: Readability averages: Wikipedia ($n$ = 500)

| | **Kincaid** | **ARI** | **CL** | **Flesch** | **Fog** | **Lix** | **SMOG** | **Fry** |
|---|---|---|---|---|---|---|---|---|
| stdev | 2.19 | 2.79 | 1.91 | 9.81 | 2.52 | 7 | 1.62 | 2.63 |
| min | 2.2 | 1.5 | 7.6 | 26.5 | 4.8 | 20.1 | 6.2 | 1 |
| 25% | 7.58 | 8.5 | 12.2 | 55.58 | 9.98 | 38.5 | 9.6 | 7 |
| med | 8.6 | 9.9 | 13.2 | 62 | 11.4 | 42.45 | 10.5 | 8 |
| mean | 8.9 | 10.22 | 13.38 | 60.93 | 11.5 | 42.83 | 10.58 | 7.66 |
| 75% | 10 | 11.7 | 14.3 | 67.62 | 12.9 | 46.9 | 11.5 | 9 |
| max | 21.8 | 25.5 | 22.6 | 93.9 | 24 | 73.9 | 19.4 | 12 |
| skew | 1.08 | 0.94 | 0.9 | -0.64 | 0.83 | 0.31 | 0.55 | -0.81 |
| kurt | 3.78 | 3.37 | 2.36 | 0.92 | 2.72 | 1.59 | 2.14 | 0.66 |

Wikipedia's pages also had similar ranges as the previous sites, and, like Wikia, had slightly higher mean and median readability values than the editor-controlled sites, namely between the eighth and nineth grade levels. Values for all of the metrics, except Flesch and Fry, were slightly left skewed and had moderate peaks around their average values. Flesch and Fry scores are slightly right skewed and have less definite peaks.

### Modeling results: Principal components analysis

Focus shall now turn to principal component analyses of readability and word usage variables on the sites, so as to characterize the readability of as many unique text types on the sites as possible.

#### *GateWorld*

With four eigenvalues above 1.0, and the scree plot indicating the beginning of isotropic variance with the fifth eigenvalue, five components were chosen for interpretation. The following list shows the spectra of how the variables load onto the principal components, with positive and negative ends of the components indicated with '+' and '-' signs. Also, in [brackets] at the ends of each component are the page types that occur most frequently in association with that component, with the most frequent page types closest to the ends.

1. + [EPS, COMICS] words chars sent prep p.avlen.s conj s.short s.pass pron v.tobe v.aux nom s.b.conj s.b.sub s.b.prep s.b.art s.b.int s.quest s.longest.w CL w.avsyl flesch kincaid ARI fry lix fog w.avlen.ch s.avlen.w smog s.shortest.w [OMNI] -

2. + [COMICS, BOOKS] s.shortest.w s.longest.w nom prep chars words conj v.tobe s.b.art s.pass s.b.prep s.b.sub pron w.avlen.ch v.aux s.b.int s.short s.b.conj sent p.avlen.s s.quest lix CL smog w.avsyl fry s.avlen.w ARI fog flesch kincaid [VG, OMNI] -

3. + [OMNI, EPS, VG] w.avsyl CL s.shortest.w s.longest.w flesch smog lix pron conj prep v.aux kincaid words s.b.sub v.tobe fog chars s.pass s.b.conj s.b.int s.b.prep s.quest nom ARI s.b.art p.avlen.s s.short sent w.avlen.ch s.avlen.w fry [BOOKS, OMNI] -

4. + [OMNI, EPS, VG, COMICS] lix w.avlen.ch s.shortest.w kincaid smog s.b.art fog nom s.quest v.aux s.b.int chars p.avlen.s s.b.sub words prep s.b.conj conj s.pass v.tobe s.b.prep sent pron s.short flesch s.longest.w s.avlen.w ARI fry w.avsyl CL [COMICS, EPS, BOOKS, OMNI] -

5. + [OMNI, EPS] w.avlen.ch smog w.avsyl s.shortest.w CL s.quest s.b.int flesch s.b.conj fog v.aux p.avlen.s pron s.b.sub s.avlen.w conj fry words sent v.tobe s.pass prep chars s.b.art s.short s.b.prep nom ARI s.longest.w kincaid lix [BOOK, EPS, VG, OMNI] -

For the first principal component (PC1), most episode and comic pages are shown to have had large numbers of words, characters, and sentences (i.e., to be lengthy overall), which is in contrast to omnipedia pages, which usually had sentences with few words, many average-length sentences and words, and were given high scores by the SMOG and other readability metrics. This shows that episodes (and comics) followed a longer-is-better approach, that omnipedia pages a more minimalist or average writing approach, and that the omnipedia approach was generally rewarded with high scores by the readability metrics, whereas the episode approach was not.

In PC2, comics and books pages were contrasted with the omnipedia paradigm, to which video game pages were deemed close. Comics and book pages had especially short and long sentences, in terms of word counts, as well as many nominalizations. Anecdotally, such pages seem to have had sentences with greater ranges of narrative expression, sometimes using short sentences for exclamatory effect, and sometimes longer ones for greater narrative detail. Nominalizations are also especially common in science fiction writing, where concrete processes are reified into more abstract ones, upon which to base a fictional

story (e.g., mechanical into mechanization). Omnipedia pages are again shown to be the preference of the readability metrics, especially because those pages had many sentences and words of average length.

From PC3 onwards, because omnipedia pages were the most numerous in the sample and had much variety, the variability present in non-omnipedia pages can also be found in some omnipedia pages. PCA finds ever-smaller, orthogonal linear dimensions of variability in a dataset, so the variability from this point onwards is describing a smaller portion of the dataset's variability than did earlier PCs. The analysis' eigenvalues indicate that PC2 accounts for about $1/3$ of the variability accounted for by PC1, PC3 for about half of that by PC2, PC4 for about half of that by PC3, and PC5 for slightly less than that by PC4. For purposes of clarity, it will be assumed that all of the following patterns also occur on some omnipedia pages, but that omnipedia pages are not mentioned because they represent a kind of noise in the dataset.

PC3 shows that some episode and video game pages were similar, in that they had words of average length (in terms of syllables), and short and long sentences, which the readability metrics, especially Coleman-Liau and Flesch, regard highly. On the other hand, some book pages had many average length words (in terms of characters) and sentences, as well as many short sentences and many sentences overall, which the Fry and ARI metrics reward. This is effectively saying that the reverse of PC2 is occasionally true, namely that non-book pages can also have an array of sentence lengths, and that some book pages can have average sentences and more episode page-like lengths. Also, Flesch and ARI seem appropriately placed, as they measure readability in terms of syllables and characters, respectively. However, by this same reasoning, Coleman-Liau and Fry are in reversed

positions from what one would expect, indicating that their associations with this PC must be based on other than the most highly loaded word usage variables.

In PC4, an orthogonal pattern within the episode and book page dichotomy is contrasted, with episode pages having had more words with average numbers of characters as well as short sentences, which the Lix metric regards highly. This makes sense, as the Lix metric defines readability in terms of characters rather than syllables. Book pages could have more words with average numbers of syllables, sentences of average length, and long sentences, which the Coleman-Liau, Fry, and ARI metrics prefer. Similarly, in PC5, episode pages are most contrasted with video game pages, in that, whereas episode pages can have average word lengths and short sentences, game pages can have long sentences with many nominalizations and prepositional beginnings, somewhat like comic and book pages.

A comparative summary of GateWorld's results versus the other three sites will be given at the end of this sub-section.

***IMDb***

Three eigenvalues were above 1.0, and the isotropic area of the scree plot began with PC4. Therefore, four PCs will be interpreted. As with the previous GateWorld section, those PCs' spectra of loadings can be found in the following list.

1. + [ACTOR] words prep chars s.longest.w sent w.avlen.ch w.avsyl smog para conj lix CL fog pron s.short kincaid s.avlen.w [CREW] s.pass ARI v.tobe flesch s.long fry s.shortest.w nom p.avlen.s s.b.pron s.b.art v.aux s.b.prep s.b.sub s.b.int s.quest

s.b.conj [TITLE] -

2. + [ACTOR] s.shortest.w lix CL kincaid fog ARI smog flesch s.avlen.w w.avsyl w.avlen.ch para fry s.longest.w s.b.int prep s.b.pron s.b.conj conj words chars pron s.pass v.tobe sent s.b.sub s.short s.b.prep nom v.aux s.long s.quest s.b.art p.avlen.s [CREW, ACTOR] -

3. + [ACTOR, CREW] s.b.pron s.b.prep p.avlen.s nom s.shortest.w s.b.art s.b.sub chars words sent conj prep s.short pron s.avlen.w ARI s.pass kincaid fog lix w.avlen.ch smog para w.avsyl s.long v.tobe CL s.longest.w flesch fry v.aux s.b.conj s.b.int s.quest [TITLE] -

4. + [TITLE, ACTOR] s.b.int s.b.sub s.quest nom ARI v.aux s.avlen.w kincaid fog s.long lix pron s.longest.w s.shortest.w smog chars conj words prep s.b.prep v.tobe s.pass CL s.short w.avsyl w.avlen.ch para s.b.pron p.avlen.s sent flesch s.b.art fry s.b.conj [ACTOR, TITLE] -

The first PC accounts for the majority of variability in this dataset, 14 times more than PC2. In PC1, the difference between actor and title pages is shown to be paramount, with crew pages sitting all together at the center of the spectrum. Actor pages overall were long and verbose: having many words, characters, and sentences, many prepositions, many very long sentences, and many words with average lengths. Title pages, by contrast, were most characterized by how their sentences begin, as well as their use of modal auxiliary verbs (e.g., could). Neither of these patterns draws a strong association with any of the readability metrics, though SMOG, a metric that prioritizes large syllable counts, is closest to the actors

pattern, and Fry, which focuses on sentence counts and lengths within paragraphs, is closest to the titles pattern, probably due to the p.avlen.s variable.

PC2 contrasts the most common type of actor page with a type of actor page that is more similar to crew pages. Then, in PC3, which accounts for a similar amount of variability as PC2, this actor-crew type is contrasted against title pages. The main type of actor pages had shorter sentences, so are scored highly by those readability metrics that involve dividing by word or sentence counts, which many do. The actor-crew type, by comparison, had longer texts with more variability in length and word usage, which did not draw the attention of the readability metrics. In PC3, in comparison to the actor-crew type of pages, title pages had many questions, often began with conjunctions, had many auxiliary verbs, and drew the attention of readability measures that value texts with more syllables and characters.

PC4, with an eigenvalue of only 0.94, is rather ambiguous. It contrasts two types of pages, both of which are most like title pages, but also somewhat like actor pages. The former had many interrogative sentences, sentences beginning with subordinate conjunctions, and nominalizations, which the ARI measure, which prefers many characters overall and short sentences, rates highly. The latter pattern had mostly conjunctive and article-based sentence beginnings, more sentences, larger paragraphs, and more paragraphs, which draws the attention of the Fry and Flesch metrics for the same reasons as in PC3.

***Wikia***

Five components qualified for interpretation on Wikia, namely:

1. + [EPS] chars words prep s.pass v.tobe sent p.avlen.s conj s.short pron s.longest.w

nom s.avlen.w v.aux s.b.pron fog kincaid ARI smog s.b.art para s.b.prep s.b.sub lix w.avlen.ch w.avsyl CL s.b.int fry flesch s.quest s.b.conj s.shortest.w s.long [OMNI, VID] -

2. + [OMNI, BOOKS, GAMES, ACTORS] CL lix kincaid fog smog ARI s.shortest.w w.avsyl w.avlen.ch para s.avlen.w fry s.longest.w flesch s.long nom s.b.art v.tobe chars s.pass prep words conj sent p.avlen.s s.short s.b.int s.b.prep s.b.sub s.b.conj pron v.aux s.b.pron s.quest [EPS] -

3. + [EPS, OMNI, ACTORS] flesch fry para w.avlen.ch w.avsyl s.quest s.short p.avlen.s sent s.shortest.w s.longest.w s.b.conj s.b.int s.b.pron s.b.art s.pass v.tobe s.long words chars pron prep CL conj s.b.prep s.b.sub v.aux s.avlen.w fog nom kincaid smog lix ARI [BOOKS, GAMES, OMNI] -

4. + [OMNI, EPS, BOOKS] para w.avsyl w.avlen.ch s.shortest.w s.avlen.w s.longest.w v.tobe s.pass nom prep s.b.prep chars s.b.sub words s.b.art conj sent p.avlen.s v.aux s.short pron s.b.pron ARI kincaid smog fog lix flesch s.b.int CL s.b.conj s.quest fry s.long [BOOKS, EPS, OMNI] -

5. + [OMNI, EPS, BOOKS] fry s.b.conj CL flesch s.quest s.b.int fog lix kincaid smog ARI s.b.pron v.aux pron conj s.b.prep s.short s.b.art p.avlen.s sent words chars prep s.b.sub s.pass nom v.tobe s.avlen.w s.shortest.w s.longest.w para w.avlen.ch w.avsyl s.long [BOOKS, OMNI] -

As on GateWorld, the largest distinction (three times larger than PC2) is between episode pages, which tended to have many average length sentences using a variety of

language, and more variable-length omnipedia pages that had more questioning and conjunctive beginnings, most of which resembled video pages and drew the attention of the Flesch and Fry metrics. PC2 contrasts episode pages with those omnipedia pages that were more like book, game, and actor pages. Such pages rate highly on many readability measures, had many short sentences, average word lengths, and many paragraphs, possibly indicating lists of trivia or quotes.

PC3 breaks apart the variability in PCs one and two, saying that some actor and omnipedia pages were more like episodes and less like books and games. This new group scores highly on Flesch and Fry, has many paragraphs, many words of average length, and many interrogative and short sentences. Book and games pages, by contrast, score higher on most of the other readability metrics, had many nominalizations, and many sentences of average length.

PC4 similarly says that some book and omnipedia pages are more like episodes, and there are actually two types of such pages. The first had many paragraphs, average length words, and a variety of sentence lengths. The second type had many long and questioning sentences, sentences beginning with conjunctions, and scores highly on the Fry and Coleman-Liau metrics. In PC5, this latter group is contrasted with a group missing the episode page qualities, which had long sentences, average word lengths, many paragraphs, and many very long and very short sentences, but which the readability metrics generally ignore.

### *Wikipedia*

Six components qualified for interpretation on Wikipedia, namely:

1. + [OMNI, EP] char word prep sent p.avlen.s s.pass conj v.tobe pron s.longest.w s.short nom s.long s.avlen.w s.b.art s.b.prep v.aux ARI fog kincaid s.b.pron smog s.b.sub lix s.shortest.w s.b.int w.avlen.ch s.quest w.avsyl s.b.conj CL flesch fry [CREW, ACTOR, EP] -

2. + [ACTOR, AUTHOR, CREW] CL lix kincaid smog ARI fog w.avsyl w.avlen.ch s.avlen.w s.shortest.w s.longest.w fry char word s.long s.b.conj v.tobe nom p.avlen.s sent v.aux s.b.int s.quest prep s.b.sub s.pass s.short conj s.b.art pron s.b.prep s.b.pron flesch [ACTOR, EP] -

3. + [ACTOR, EP] s.avlen.w ARI v.aux smog nom kincaid lix s.b.int conj s.b.art fog s.b.sub s.b.prep s.b.conj s.quest prep s.long s.longest.w pron s.pass s.b.pron v.tobe CL word char s.short p.avlen.s sent flesch w.avsyl w.avlen.ch fry s.shortest.w [CREW, AUTHOR, ACTOR] -

4. + [ACTOR] flesch s.b.pron s.avlen.w s.long s.longest.w fog pron ARI kincaid s.short prep fry smog v.tobe lix word char s.pass s.b.prep p.avlen.s sent s.shortest.w conj CL s.b.sub v.aux nom s.b.art s.quest w.avlen.ch w.avsyl s.b.conj s.b.int [CREW, OMNI, EP] -

5. + [EP, ACTOR, CREW] fry CL flesch s.quest s.b.conj fog s.b.int lix s.b.art smog kincaid s.b.prep prep s.long ARI s.b.sub pron conj s.b.pron nom v.aux s.short sent p.avlen.s char word s.pass s.longest.w v.tobe w.avlen.ch w.avsyl s.avlen.w s.shortest.w [AUTHOR, CREW, ACTOR] -

6. + [ACTOR, EP, OMNI] s.quest s.b.conj flesch s.avlen.w s.longest.w s.b.int s.shortest.w

s.long fog word kincaid ARI char sent p.avlen.s s.short smog lix w.avsyl w.avlen.ch v.tobe prep s.b.art conj s.pass fry pron v.aux nom s.b.sub s.b.prep s.b.pron CL [OMNI, EP, ACTOR] -

As with the previous sites, PC1 most contrasts omnipedia and episode pages, which accounts for three times more variability than PC2. However, probably because only the most popular episodes are documented on Wikipedia, the split is not as clear as on previous sites, and actor and crew pages, which Wikipedia covers more, grouped together with episode pages. Also unlike previous sites, omnipedia (and some episode) pages were the longest and had the most variety of language, probably because that category included the lengthy list pages that are a staple of Wikipedia. Episode, actor, and crew pages, by contrast, are preferred by the Fry and Coleman-Liau metrics, probably from having had many average length words (in terms of characters and syllables, which are weighted heavily by those two metrics), and were also heavily interrogative and had conjunctive sentence beginnings.

PCs two and three are essentially mirror images of each other page-wise, indicating that two kinds of orthogonal spectra of variability occurred between people-oriented and episode-oriented (and some actor) pages. The first spectrum says that people-oriented pages rate highly on every readability metric except Flesch, and tended to have average word and sentence lengths. This is compared with more episode-oriented pages, which score highly on Flesch, and were characterized by many pronouns, as well as sentences beginning in prepositions and articles. However, some episode-oriented pages had average sentence lengths, auxiliary verbs and nominalizations, and scored highly on ARI, SMOG,

Kincaid, and Lix. On the other hand, some people-oriented pages had many sentences with very short words, many words of average length, and scored highly on the Fry and Flesch metrics, somewhat like the actor and omnipedia pages in PC3 of Wikia.

PC4 says in what way actor pages were usually different from the other page types, namely that they score highly on Flesch, began with pronouns, and had average-to-long sentences. PC5 compares the episode-author-crew pattern from PC1 with something close to the actor-author-crew pattern of PC2 two and three. The former is the preference of the readability metrics, and had interrogative sentences and conjunctive sentence beginnings. The latter had a variety of sentence lengths, average-length words, and to be verbs. The final PC says that a dimension of variability occurs within, and spans, the most common page types. On the one hand were interrogative sentences with conjunctive beginnings, high Flesch scores, and average-to-long sentences. On the other hand were sentences more characterized by their beginnings, nominalizations, and auxiliary verbs, which score highly on Fry.

### *Conclusion*

If one pairs the applicable of ends of each PC between the previous analyses – for example, placing the lengthy episode and comic page end of GateWorld's PC1 alongside the lengthy actor page end of IMDb's PC1, and so forth – a fairly consistent set of **six writing style types** emerges. This section briefly discusses the similarities and differences in those types across the sites, as well as their relation to the readability metrics.

The first and most prevalent type, as is often the case with PCA, are the most frequent/lengthy texts, and occur on whatever page type was most prevalent on each site. This

included episode and comic pages on GateWorld and Wikia, actor pages on IMDb, and omnipediac and episode pages on Wikipedia. Such pages had many words, characters, sentences, and paragraphs of average length, often many prepositions, and, on the wiki pages, many passive sentences. GateWorld and Wikipedia also shared the pattern of having many conjunctions. Such writing usually receives low readability scores from all metrics. Due to its resulting from the massive agglomeration of information on popular topics, without much regard for narrative subtlety, one might call this the *mass agglomerative style*.

The second most common type had very short sentences on GateWorld and Wikia, sentences with conjunctive and interrogative beginnings on all except GateWorld, and average sentence and word lengths on GateWorld and Wikipedia. All except IMDb score highly on the Fry metric for this writing type, the wikis score highly on the Flesch metric, and GateWorld scores highly on SMOG and Lix. Such writing occurs most on omnipediac pages, which included title pages on IMDb, and episode and actor/crew pages on Wikipedia, so one might call this the *general documentation style*.

The third type had many very short sentences, but otherwise average length words and sentences, and score highly on Coleman-Liau, Lix, Kincaid, Fog, and ARI everywhere except GateWorld, where the writing of this type also included very long sentences, nominalizations, prepositions, conjunctions, and general length, which does not attract any readability metrics. As this type of writing occurred most often on book, comic, and author pages, one might call this the *book/author style*.

Type four had many average length paragraphs and sentences; was often interrogative; had many conjunctions, pronouns, and prepositional sentence beginnings on wikis; and scored highly with the Flesch metric, most on GateWorld and Wikipedia, as well as some-

times the ARI, Kincaid, and CL metrics, more like the book/author style. IMDb and Wikia also shared a tendency towards auxiliary verbs in this type of writing. As a more discussant form of the omnipedia style, with aspects of the book/author style, this might be called the *style of reviewers, interpreters, or commentators*. This style can occur on most any type of page.

Types five and six are possibly sub-categories of the reviewer style, and occurred in the same pages as that style. Both styles most attract the attention of those readability metrics that weight word and character counts most heavily – the ARI and Lix metrics for the first style, and the Coleman-Liau metrics for the second style. The first style also sometimes rates highly on Flesch, and both styles often rate highly on the Fry metric, probably because both tended to have many sentences and syllables. In the first style, short sentences, average length paragraphs, and sentences beginning with pronouns and interrogatives were the norm. In the second style, sentences beginning with conjunctions, long sentences, and average length words were more common. These styles may merely refer to *two styles of reviewing*: the first pithy and questioning, and the second more about offering lengthy interpretations.

The writing style types and readability metric associations identified in this section are not only more empirically validated and precise than merely assuming that writing styles differ by types of website sections, but one could easily imagine them generalizing to other communal Web 2.0 sites.

### 4.2.6 Readability as stylistic

**AF6:** How do pages visual style features (e.g., colors and fonts) compare with each other?

To answer this question, a large number of regular expressions were applied via the POSIX `grep` utility to all of the HTML code downloaded from each site, extracting for each HTML page both all of the properties/rules of Cascading Style Sheet (CSS) versions 1.0 and 2.0 as well as all of the deprecated, proprietary, or otherwise pre-CSS style-related tags and attributes that were known to the researcher. The following lists give those tags and attributes:

**Tags:**

animate, applet, audioscope, b, basefont, bgsound, blackface, blink, blockquote, bq, center, code, comment, dir, embed, fn, font, i, ilayer, image, isindex, layer, limittext, marquee, menu, multicol, nobr, noembed, nolayer, nosmartquotes, s, samp, shadow, sidebar, sound, spacer, strike, u, wbr, and xml (which is different from the XML version declaration, <?xml version="1.0"?>)

**Attributes:**

alink, align, autoactivate, background, bgcolor, bgproperties, border, bordercolor, bordercolordark, bordercolorlight, cellspacing, cellpadding, color, compact, controls, dynsrc, face, frame, framespacing, halign, height, ibmlogo, internal-gopher-menu, language, left_arrow, left_margin, link, loop, marginheight, noshade, nowrap, red_bullet, rightmargin, size, start, target, text, type (except on script tags), usestyle, valign, value, version, vlink,

and width

Using the POSIX `sort`, `uniq`, and `cut` programs, those style files were semi-manually cropped, sorted, and combined both within and across the sections of each website. The results were manually examined by the researcher, who created summary lists of the styles used in each section and each site. On none of the sites were there found to be more than trivial modifications to the style rules being used between website sections (e.g., the same or very similar rules were used on the GateWorld episode and book pages). This is a biproduct of studying well-established sites, the designers and webmasters of which have either modularized or standardized their work to a great degree. Therefore, the following sections summarize only the styles used across each site, not per-section.

**GateWorld**

Though some inline CSS appeared, GateWorld's creators heavily employed deprecated font and similar tags. Figure 4.3 displays the color palettes that appeared on different areas of the site. The colors themselves are in the background of the boxes, and the foreground text gives the hexadecimal color value that either appeared on the site or can exactly generate the color that appeared on the site.

All of the colors are subdued blues, greys, and blacks – colors similar to those prominent in the set designs of the two most recent Stargate series. For a fan familiar with the series (e.g., the researcher), the colors remind one of the Atlantis characters' blue-grey uniforms and the Universe characters' black uniforms, the cold metallic facades of the futuristic vessels and military buildings where the characters often live, the sky blue back-

Paragraphs

#000000 #333333 #666666 #92B0CF #A8C4E1 #B0B0B0 #BBD1E8 #CCCCCC #EEEEEE #FFFFFF

Navigation links

#000000 #666666 #FFFFFF

Text colors

#000000 #000011 #18325D #213F67 #2A0532 #2C5F8B #333333 #3479A5 #374251 #444444 #589ECA #666666 #6699CA #92B0CF #999999 #A8C4E1 #BBD1E8 #C0C0C0 #CCCCCC #CFD6DE #FFFFFF

Figure 4.3: Color palettes: GateWorld

grounds often shown when the characters fly US Air Force fighting aircraft and the images of such aircraft that adorn offices and public spaces on the show, and the darkness of outer space as viewed from space ships. All of these themes also appeared frequently in the site's banner and background images.

Paragraphs of text always used the standard modern/industrialistic Windows Arial font. Most navigation links were white, so that they stood out. Links to add GateWorld's headlines to one's own site were black and subdued, as were Omnipedia page links, which were dark grey (#666666). External/off-site links to actor pages on IMDb received no special styling, and had a regular font size of '1', indicating they were not intended to stand out. External links to posts and interviews were black and de-emphasized, but had an increased line height of 15 pixels, and used some deprecated italics (<i>) tags. All of the site's fonts were sans-serif, giving a cold and modern look, the site's preferred fonts being Verdana, Arial, Tahoma, and Lucida Grande. Font weights were either bold or normal. Font sizes were usually '1' (user's default), with small text being 8 point and headers being 9-12, 14,

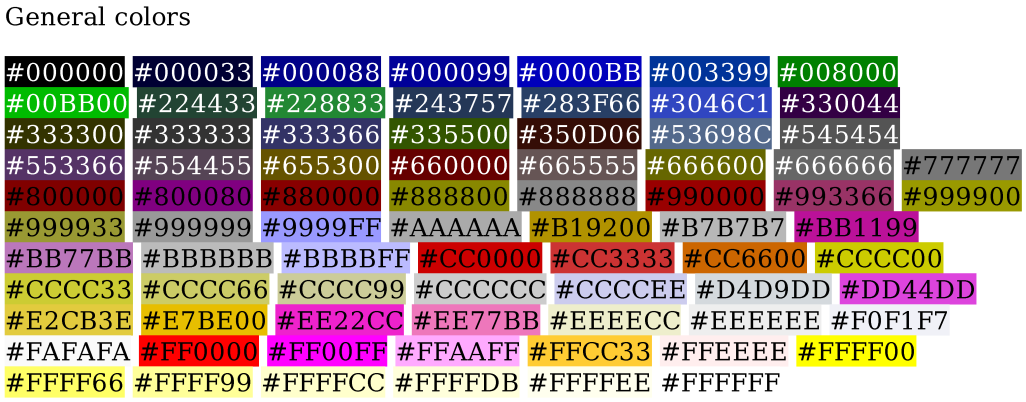



Figure 4.4: Color palettes: IMDb

and 18 pixels. These settings indicate that the site's own content was usually intended to visually stand out from that of other sites, though links to other sites did exist. It did not resemble social networking sites, where all links are designed to keep one on the same site.

### IMDb

IMDb's creators used all three forms of CSS (inline, top-of-page, and external file), as well as inline deprecated style tags and attributes. All pages drew from a single common color palette, shown in figure 4.4.

The palette is considerably brighter than GateWorld's, obviously intended to conform to a general style guide or brand identity of the company, rather than to a specific media franchise. So many primary colors might be an attempt to make the site socially generic. The yellow base of many of the colors suggests the gold of the IMDb logo, and of many movie studios' logos, as well as the usual color of subtitles in movies.

All fonts on IMDb were also sans-serif, the preferred family being again Arial, probably because most computer users run Windows, and Helvetica. Font weights were either

bold or normal. Font sizes had a wide range: 8 point for forms; 10 point for the disclaimer on each page; 12 point, 15 point, 20 point, 10 pixels, 13 pixels, 15 pixels, "xx-small," and "x-small" for regular text; 18 point or "medium" for first-level headers; 16 point for second level headers; "small" for third level headers. Most text was either aligned middle or justified. Backgrounds were either repeated or positioned. Relative positioning and deprecated width and height attributes were used for cast lists. Borders were either invisible or 1 pixel, were solid, often used the old <hr> tag. These settings also indicate a preference for modernism/industrialism, as well as a somewhat disorganized or unprincipled opportunism, common in business, of using whatever seems easiest for achieving the desired effect. Also, as is common on large corporate sites, all of the primary pages' links either returned one to the same site or to one of their advertisers. To find external content, one had to navigate to sub-pages containing lists of external links or user feedback.

**Wikia**

All styles on Wikia were done in CSS, within an XHTML template. The styles were a combination of the Monaco theme/skin, which resembled Wikipedia's Monobook theme and was used for the default/administrative Wikia pages, and a custom container theme for user-generated content, which presumably changed to suit each differently themed wiki on Wikia (e.g., the container of the Wookiepedia, the Star Wars wiki, contained images relevant to that franchise). Figure 4.5 shows the color palettes of both themes.

The Monaco theme was quite basic, with just a few Earth tones. The container theme, on the other hand, had many colors more resemblant of the Stargate franchise: black, greys, and blues. These colors were also most visually prominent in the interface. However, the

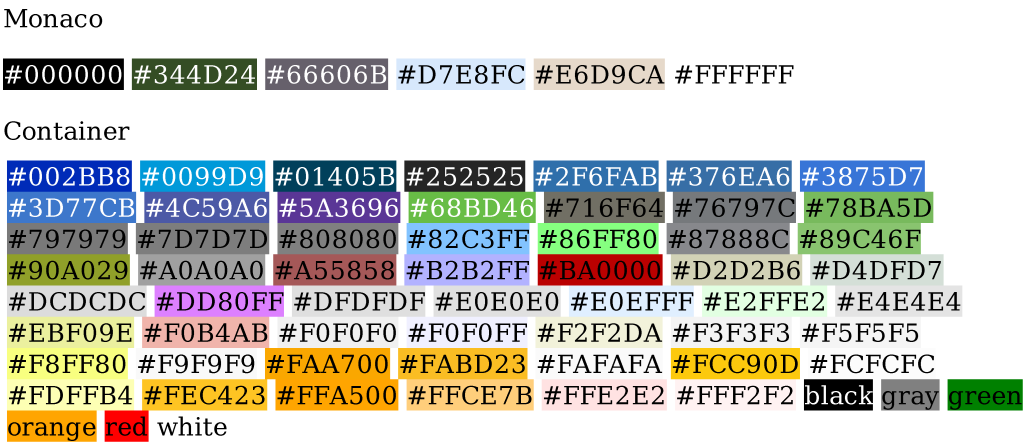



Figure 4.5: Color palettes: Wikia

theme also included yellow, green, and red colors. Green was used for administrative features, such as edit page buttons. Red appeared when highlighting advertisements. Yellow was not found prominently on any pages viewed for this project, other than in small icons, which were not colored by CSS, so may be unused.

Fonts were primarily sans-serif, again. However, Arial was not the preferred typeface; Lucida Grande, which appeared through Mac OS X at the time, was preferred, with the Windows Tahoma and Arial fonts coming second and third. Helvetica and Verdana were also recommended. Additionally, the serif fonts Times (again the Mac version comes first) and Times New Roman (the Windows version second) appeared in in-text quotes. As with IMDb, a mixture of font sizes and page division measurement types (percent, em, point, pixels, and built-in keywords) appeared, indicating a similarly unprincipled approach (i.e., neither a fluid nor static CSS layout), at least in this respect. Font weights were either 400, 700, bold, or normal. Text was aligned in every way possible, and occasionally had italics. CSS positioning was used consistently, textual overflow was managed, and borders

were 1-4 pixels wide, colored, and solid. Their CSS also consistently used the K&R and BSD/Allman indent styles, the former of which is common in open-source programming communities[1], and the latter more in business.[2] They also demonstrate greater knowledge of CSS selectors than either of the edited sites. Finally, most pages linked to external sites. These things indicate that the site designers probably preferred Macs, and that they have made a reasonable effort to comply with broader XHTML and CSS best practices. However, like the two editor-controlled companies, a degree of unprincipled opportunism was also apparent, which might result from Wikia's being a for-profit business.

**Wikipedia**

Wikipedia's styles were the most in communion with the open-source community. Its ubiquitous Monobook theme used all CSS and XHTML, and was modularized between several stylesheets. Although the frequent blues and greys of Monobook do accord well with Stargate, this is not intentional, as the template is not customized for individual pages. The color palettes stored in each stylesheet are depicted in figure 4.6.

As with IMDb, the palettes say more about Wikipedia as an institution than about Stargate. In this case, a non-profit organization wishes to offer a public service and keep itself in the background, as indicated by light pastel Earth tones and whites. As with advertisements on the other sites, reds did occasionally appear in fundraising campaigns on the site, and yellows highlighted the borders of interface elements and the site's disclaimers.

[1] K&R stands for Kernighan and Ritchie, a reference to those authors' famous book on the C language, in which most Linux programs are written.

[2] The BSD license has fewer restrictions than the GPL, which is most common in free/open-source projects.

main.css

#000000 #002BB8 #005896 #00BBFF #2F6FAB #5A3696 #716F64
#76797C #772233 #7D7D7D #A55858 #B2B2FF #BA0000 #DCDCDC
#DDDDDD #E2FFE2 #EEEEEE #F0F0F0 #F0F0FF #F9F9F9 #FAA700
#FABD23 #FBFBFB #FF0033 #FFA500 #FFAE2E #FFCC66 #FFCCCC
#FFCE7B #FFE2E2 #FFFFCC #FFFFFF black blue gray green orange white

shared.css

#2F6FAB #333333 #4C59A6 #AAAAAA #BBBBBB #BBBBFF #CCCCFF
#DCDCDC #DDDDDD #ECECEC #EEEEFF #F2F2F2 #F3F3F3 #F9F9F9
#FCFCFC #FFF2F2 #FFFFFF

commonPrint.css

#2F6FAB #AAAAAA #BA0000 #CCCCCC #CCFFCC #DDDDDD #EEEEEE
#F9F9F9 #FFFFAA #FFFFFF silver white transparent

combined.min

#0645AD #333333 #4C59A6 #666666 #777777 #999999 #A8D7F9
#AAAAAA #DDDDDD #EEEEEE #F3F3F3 #FAFAFA #FFFFFF

JQuery

#000000 #212121 #363636 #999999 #A7D7F9 #AAAAAA #C0C0C0
#CD0A0A #DADADA #E6E6E6 #FBF9EE #FCEFA1 #FEF1EC #FFFFFF

Figure 4.6: Color palettes: Wikipedia

Default font families (sans-serif, serif, and monospace) were used throughout Wikipedia, indicating that site's creators express no operating system preference. Fonts and page elements were also sized with relative forms of measurement (e.g., percent, em), so that the pages resized based on users' browser settings, indicating a conscientiousness towards users and awareness of low-vision accessibility issues. The only static font sizes were in the stylesheet used for printing, a medium in which page sizes can be somewhat taken for granted. Font weights were either bold or normal, as well as 400 or 700 in JQuery. Text was aligned on all margins, but not justified, and could be italic. CSS positioning was used consistently, textual overflow was managed, and borders were standard 1-2 pixels, dashed/dotted/solid, and colored. Their CSS stylesheets were consistently in the open-source K&R indent style, and they used some of CSS's most complex selectors. Finally, nearly every page links to some external content. These settings indicate an approach largely guided by accessibility and usability principals, and with no apparent corporate affiliations.

**Conclusion**

Stylistically, distinctions between these sites fell more along ideological or business model lines than editorial lines. The more that a site was either profit-oriented or industry-affiliated, the more that it tended towards opportunism in its own context and away from international standards and academic and open-source community best practices. The opposite appeared to be true, to the degree that a site was non-profit-oriented. Larger sites also featured fewer customizations specific to this media franchise, likely in order to simplify system administration. GateWorld was small and business-minded, made almost no at-

tempts to follow standards, and was highly customized. IMDb was large, business-minded, not customized, only peripherally engaging with standards. Wikia was a mixture of both of these trends, having enough of a profit agenda to draw it into opportunistic coding practices, but also a templating system for incorporating franchise-specific customizations into the administrative template, as well as making a concerted effort to engage with international standards and the open-source community. Wikipedia was large and offered no customizations, but was non-profit and made every effort to engage with international standards and academic and open-source communities. As with the previous section, these results could easily generalize to sites on any topic.

### 4.2.7 Conclusion

The following cross-site conclusions were drawn by the analyses in this section.

Section 4.2.1 showed that smaller sites can have more focused navigation structures, whereas larger sites must rely more upon taxonomies or folksonomies and search engines. Navigation structures typically only delivered the user to franchise- or series-level content, after which browsing was necessary. Whereas editor-controlled sites had more infrastructure around avoiding repeatedly answering questions about their policies and preferences, wiki sites relinquished control of content and invested in infrastructure to help users discuss.

Episode pages universally contained more photos and common media formats, such as JPGs and PDFs, than did other pages (§4.2.2). The editor-controlled sites used many GIF images in their interfaces, a somewhat outdated graphic design technique. On the other

hand, the wiki sites used more contemporary, and a greater variety of, formats, such as Atom feeds, PNG and SVG images, and OGG and MOV multimedia files.

Neither editor-controlled site contained broken links, suggesting that their editors automatically scan for and fix or remove them, though wiki sites contained many, perhaps because they cannot rely on their users/editors to systematically correct them (§4.2.3). Wikipedia's pages contained many more, and greater variety of, broken links than did Wikia's, possibly suggesting a less devoted and/or more chaotic user base.

Smaller sites had difficulty keeping accessibility errors low, especially on obscure pages (§4.2.4). Larger sites probably had more personnel for uniformly policing these sorts of problems. Wikis had about half the number of HTML errors as did the edited sites. Accessibility errors were generally more common on the sites than HTML validation errors. The most frequent accessibility problems were either JavaScripts restricting accessibility to only the mouse or obscuring page content from users who cannot use JavaScript, or media files and tables not having textual alternatives.

Average readability scores of texts on the sites were between the $7-9^{th}$ grades for the editor controlled sites, and were slightly higher ($8-9^{th}$) for the wiki sites (§4.2.5). All of the sites except GateWorld had many pages empty of content text, though this happened much less often on the wiki pages than on IMDb. Six writing style types were found in the pages of each site. The mass agglomerative style occurred on the most popular pages, where large amounts of information were dumped without much consideration for style. The general documentation style occurred on the more encyclopedic pages, and was rather descriptive and mundane. The book/author style occurred on pages about print publications, and often contained a greater variety of narrative devices and sentence structures than the general

documentation style. The style used by reviewers, interpreters, and commentators occurred on every page type, and comprised two sub-types: a pithy and interrogative style taken by reviewers, and a long-winded discussant style taken by commentators.

Finally, stylistically, the more that a site was either profit-oriented or industry-affiliated, the more that it tended towards opportunism and shortcut-taking in its own context and away from international standards and academic and open-source community best practices (§4.2.6).

## 4.3 Representational IQ: Accuracy

### 4.3.1 Currency as timeliness

**RA1:** What are the distributions of content creation/posting and last-modification on fan-site pages?

It was only possible to answer this question for the wiki sites, because the editor-controlled sites did not reliably provide dates for either their pages or their subsequent modifications of those pages. On 10 September 2009, both the Wikia and Wikipedia application programming interfaces (APIs) were queried for both the first and most recent revision records of all Stargate-related pages. The records were requested in XML format, only one record was requested by each query, and only the timestamp field was requested for each record, because that was all that was needed to answer the question. The download speed and frequency were set to rates well slower than required by the sites, and each query/download was set to retry until successfully completed. Standard POSIX utilities

(`wget`, `grep`, `sed`, etc.) were used for the downloading and record parsing processes.

Then, the Gregorian dates were converted to continuous Julian Day Numbers, which are commonplace in astronomy, based on equations provided in Seidelmann (1992, p. 604), using a custom program. This was done in order to avoid the complex discrete and cyclical numerical properties of the Gregorian calendar. Also, the time-of-day portions of the data were not used, because the times of day at which people create or modify content was not of great interest, and would have added complexity to the analysis.

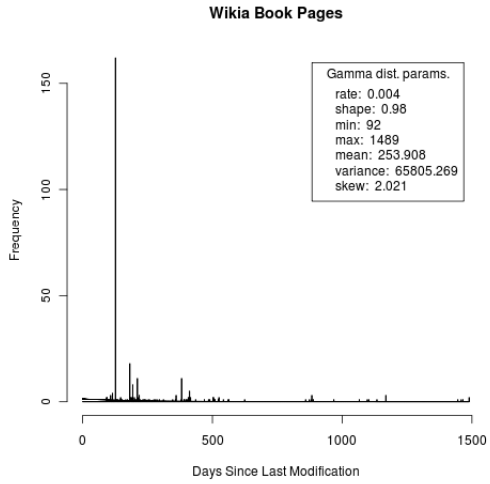
Finally, for phenomena that have either grown or grown-and-lost in popularity, as is the case with these websites, "time since..." measures are usually Gamma or exponentially (which is a type of Gamma) distributed, having a large amount of observations in the more recent past and a long tail trailing off into the more distant past. Therefore, the two sets of creation and last-modification measurements were converted from raw date variables into time-since-collected variables by subtracting each revision/creation date from the collection date. These variables will be studied in histogram form, by counting the number of times that content was either revised or created on a certain number of days since collection. The plots in this section were created with the `hist`, `lines`, `legend`, and `dgamma` functions in R, following the examples in Crawley (2002, pp. 487-490).

One unexplained error occurred during data collection, and had a small effect on the analysis. For some reason, no records were returned by the Wikia API for revisions made in the 85-126 (*mean* = 93.4) days prior to collection date, nor by the Wikipedia API for those made in the 57-59 days prior to collection. On 4 March 2010, the researcher confirmed that the queries had all actually been run on 10 September, via system timestamps; that it was possible to retrieve new/today's revision records from both the Wikia and Wikipedia

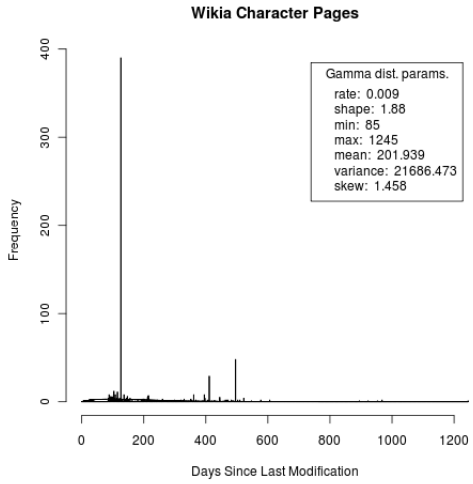
APIs, using the same code used in September; and that the date parsing, conversion, and subtraction programs work correctly. Since Wikia and Wikipedia are run by related groups, run the same Web server and wiki software (i.e., MediaWiki and Apache on Ubuntu or Debian Linux), and may share hosting infrastructure, perhaps they were both experiencing API delays, which they did not apparently advertise, on the collection date. This affects the following analyses by causing an unnaturally abrupt lower bound to appear on the left side of each variable's results.

As expected, most of the variables could be well-fit by a Gamma distribution. Figures 4.7 - 4.10 show all of the Wikia and Wikipedia variables that were obviously Gamma distributed, dividing the pages according to the websites' sections (e.g., books, characters, etc.). Note that the apparent discreteness in several of the graphs is from producing histograms with a relatively small number of observations. Figure 4.11 shows sections from both sites that were Gamma-like, having a large spike in the recent past, but also a small one in the distant past. Figure 4.12 shows two sections that had only a small spike in the recent past, and a larger spike in the distant past. Finally, figure 4.13 shows that the Wikia page creation distribution resembles a wave of several Gamma, or possibly normal, distributions. In the following figures, *shape* is the parameter representing how closely the peak of the distribution approaches the y-axis. *Rate* is an inverse scale parameter, such that larger rate values represent less spread in the data. The *mean* is $shape/rate$, the *variance* is $shape/rate^2$, and the *skewness* is $2/\sqrt{shape}$.

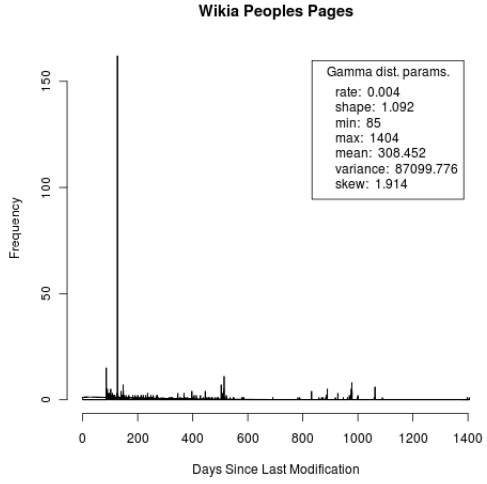
In all of these sub-figures, one sees a peak of values to the left of the graph, trailing off to the right. For the last-modification variables, this generally means that most pages had been updated relatively recently. Wikipedia's pages were the freshest overall, being



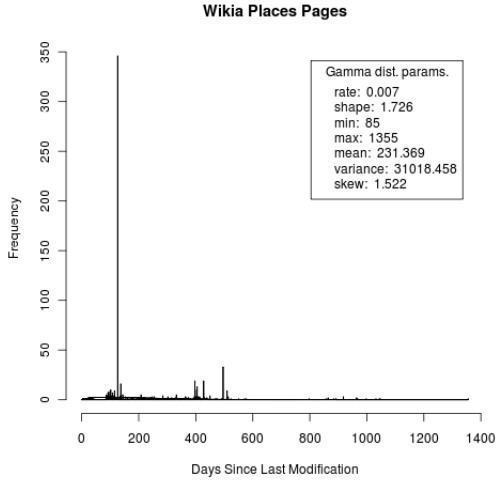

(a) Wikia Books




(b) Wikia Characters

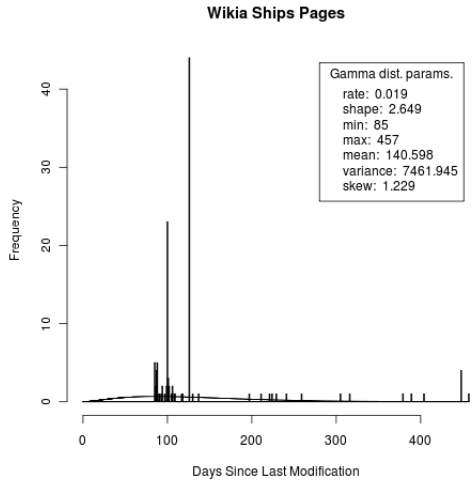



(c) Wikia Peoples

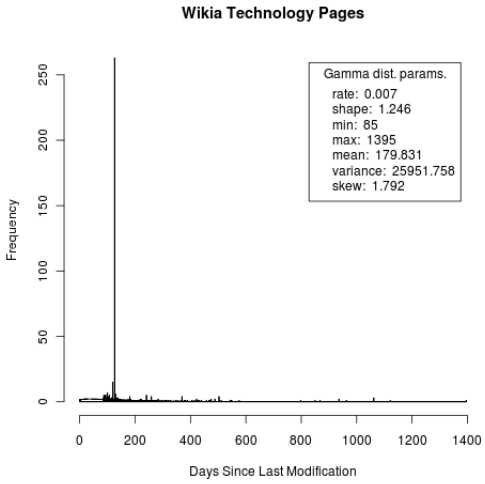



(d) Wikia Places

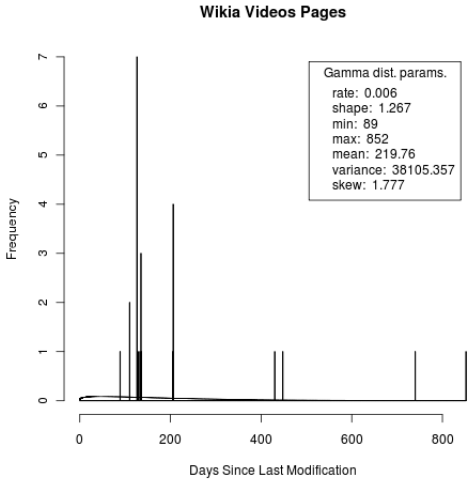
Figure 4.7: Gamma-distributed variables (1 of 2): Wikia




(a) Wikia Ships

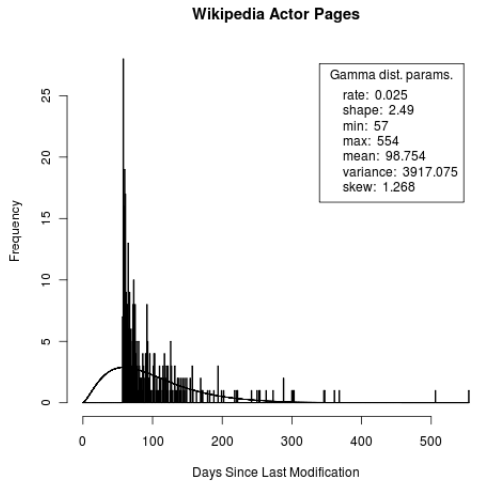



(b) Wikia Technologies

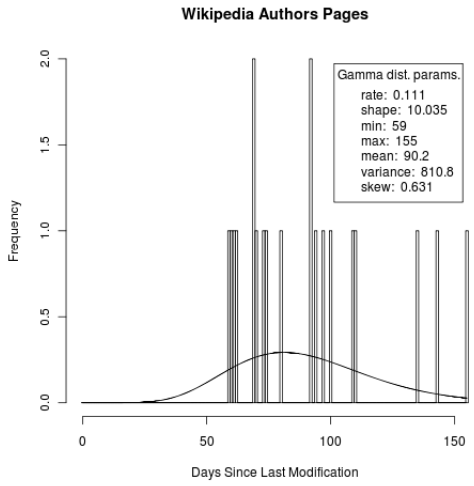



(c) Wikia Videos

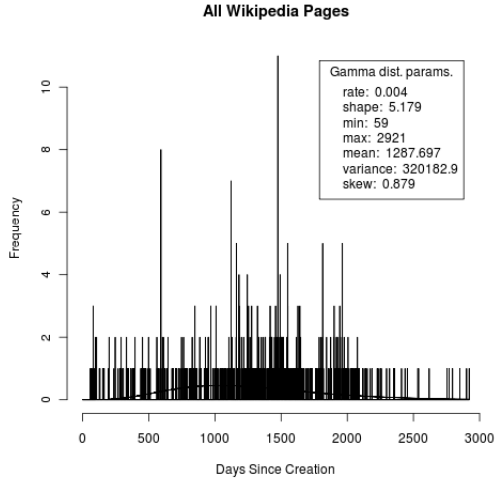
Figure 4.8: Gamma-distributed variables (2 of 2): Wikia



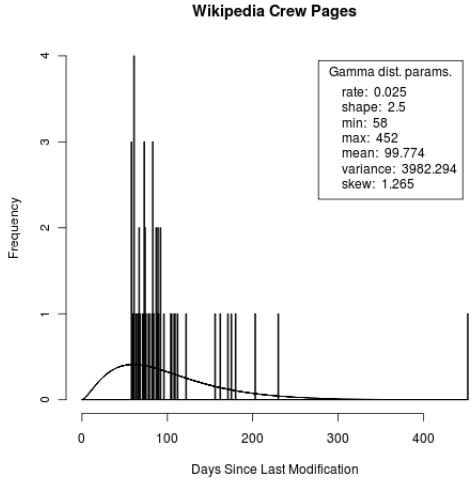

(a) Wikipedia Actors




(b) Wikipedia Authors

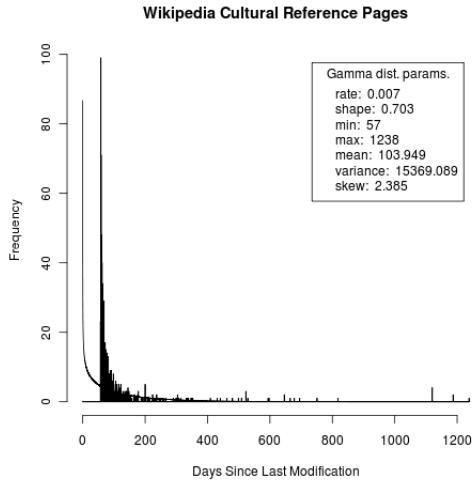



(c) Wikipedia Creation

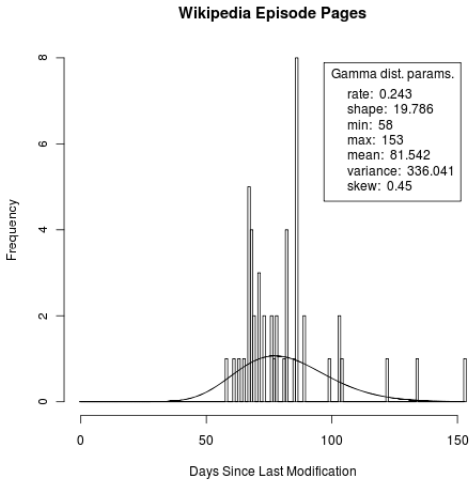



(d) Wikipedia Crew

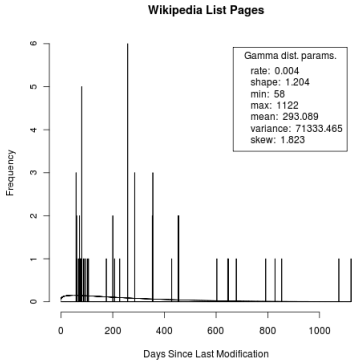
Figure 4.9: Gamma-distributed variables (1 of 2): Wikipedia



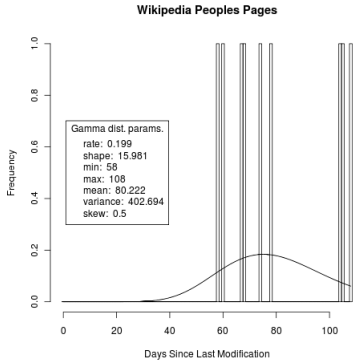

(a) Wikipedia Cultural References



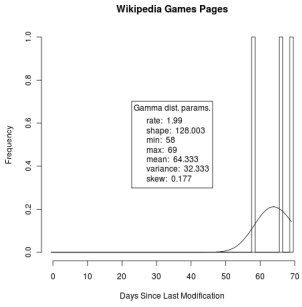

(b) Wikipedia Episodes




(c) Wikipedia Lists




(d) Wikipedia Peoples




(e) Wikipedia Games

Figure 4.10: Gamma-distributed variables (2 of 2): Wikipedia

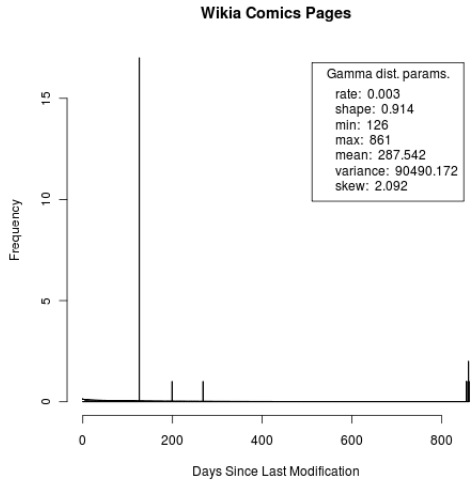
updated quarterly on average (*mean* = 114 days). Wikia's pages were updated more than every two thirds of a year (*mean* = 220 days). On Wikia, ship and technology pages were updated most frequently (*means* = 141 and 180 days), and peoples/races pages were updated least frequently (*mean* = 309 days). For Wikipedia, most page types – including game, peoples, episode, author, actor, and crew pages – were updated within the past 2-3 months, though list pages had a mean of 293 days.

In terms of rates and shapes, there were basically two types of distributions. Those variables with many observations were less affected by the API delay collection error discussed at the beginning of this section, and tended to group to the middle or left of the plots, having more or less variance. However, those variables with few observations (e.g., Wikia ships, and Wikipedia games and peoples) were more affected by the delay, appeared to begin towards the right of their graphs, and had higher rates (i.e., lower variance). On Wikia, books, peoples, and comics (see figure 4.11) pages had the most variance in their dates, but were also the closest to the y-axis. Ships pages had the least variance, and fell into the small-and-on-the-right type of distribution just discussed. High peaks on one or two days in the recent past for book, character, peoples, places, and technology pages might indicate a recent fan drive on that site to update pages (though the researcher could not find an advertisement for this). On Wikipedia, list pages had the highest variance and, along with cultural reference pages, were the most left skewed. Compared with Wikia, most Wikipedia page types had lower variances and were more shifted towards the middle, though did not have the high peaks on one or two days, indicating that Wikipedians updated pages in more of a continuous manner, whereas Wikians seem to have updated many pages at once. Furthermore, the discussion of figures 4.11 - 4.13 below will show that this behavior happens

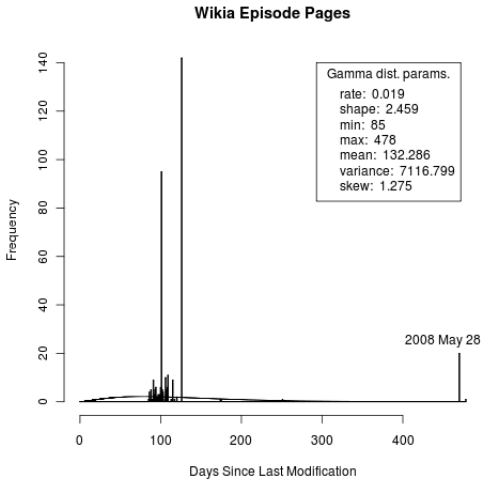
periodically on Wikia.

Wikipedia's page creation variable (see figure 4.9(c)) was also Gamma distributed. Wikipedia pages were an average of 3.5 years old, which means that most were created towards the end of Stargate SG-1 and the middle of Atlantis. SG-1 was canceled when it was still receiving high ratings, in order to draw general popular attention to Atlantis and to two direct-to-DVD SG-1 movies. Hence, it would make sense that that was when Wikipedia, which generally reflects only those things that have reached the surface of public awareness, devoted most attention to Stargate. The oldest Wikipedia pages were almost exactly eight years old, placing them around the time of the SG-1 season five mid-season finale, which was also the series' $100^{th}$ episode. That episode, like the $200^{th}$ episode five years later, was a joke show, filled with in-jokes and self-satire, and received high viewership. Hence, it would also make sense that that episode would have garnered the attention of Wikipedia.

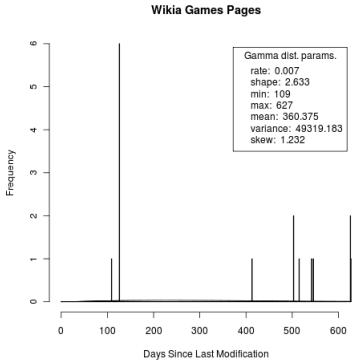
Before discussing periodicity in the above figures, it should be noted that the majority of observations of the variables displayed in figure 4.11 follow a Gamma distribution. On Wikia, the episode page type had a high peak around 132 days before collection (30 April 2009) and a small peak around 478 days (28 May 2008). The first peak is similar to the previous ship page results: episode pages were updated slightly more frequently than were pages about ships, the two distributions have the same rates, and ship pages' shape is only slightly larger. The first peak of the games pages had a similar rate and shape to the other omnipediac pages on Wikia, as the comics pages did to the book pages. All of these peaks are also focused enough in time that they could indicate a fan campaign to update pages en masse.



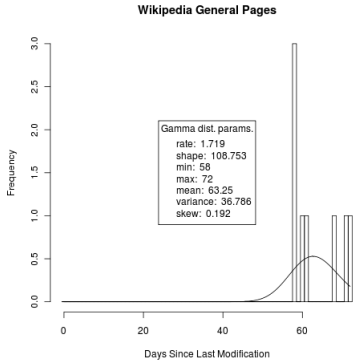

(a) Wikia Comics



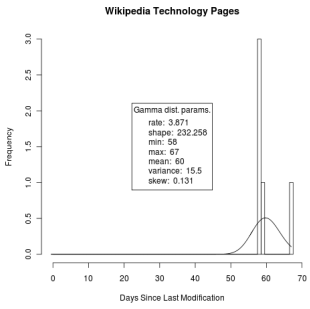

(b) Wikia Episodes




(c) Wikia Games

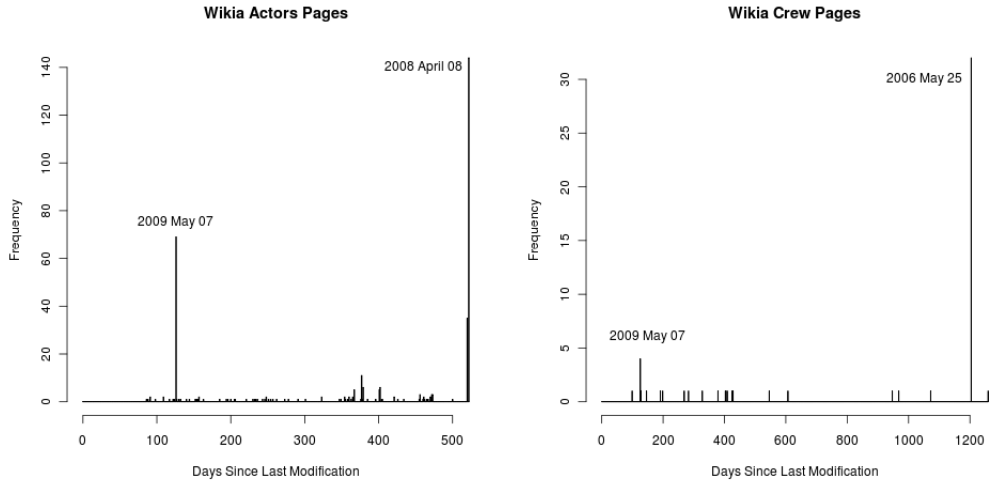



(d) Wikipedia General



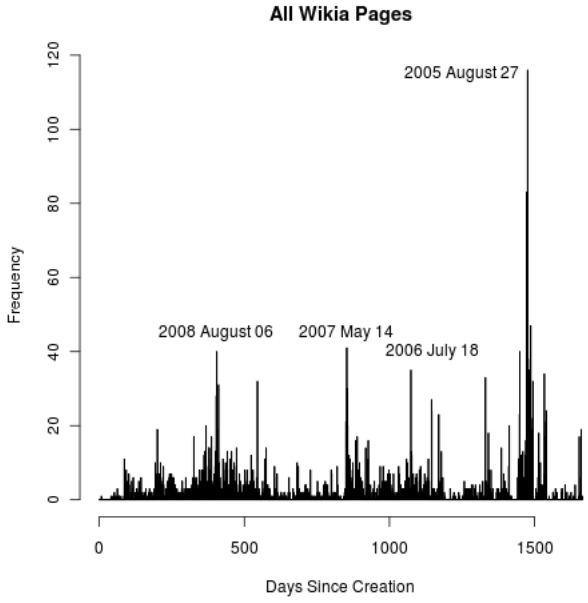

(e) Wikipedia Technologies

Figure 4.11: Gamma and periodic variables: Wikia and Wikipedia




(a) Wikia Actors (b) Wikia Crew

Figure 4.12: Largely periodic variables: Wikia and Wikipedia




(a) Wikia Creation

Figure 4.13: Periodic webpage creation: Wikia

The periodicity of last-modification behavior on these sites only emerges when the approximate dates of the peaks, which have been labeled in figures 4.11 - 4.13, are combined across the different variables. On Wikia, eight variables, spanning a range of types, contained peaks in early and late May, around the times that public and private universities end their spring semesters, going back to 2006. Two contained peaks around the times that public and private universities took Spring Break in 2008 (i.e., mid-March and early April). Finally, one peak, on the games variable, occurred over Christmas 2007. This suggests that the periodic updating frenzies on Wikia are probably not the result of fan campaigns, but of university students taking semester breaks, and the updates happen across all pages types. On Wikipedia, four omnipediac variables (i.e., games, general, peoples, and technology) had peaks in early July 2009 and one (peoples) in early June 2009. Though the periodicity is not as stark as on Wikia, this suggests that omnipediac pages also tend to be updated periodically on Wikipedia, namely in the early-to-mid summer.

Page creation dates on Wikia followed a different cycle. In all years except 2007, most pages were created in the mid-to-late summer, between the end of July and the end of August, before school resumes in the fall. In 2007, many pages were also created in early May, perhaps joining the earlier last-modification pattern at that time of year. This difference between page update and creation times of year might be due to the relative amount of work involved in the tasks. In the beginning of summer, and over brief breaks, students might be willing to make small changes to wiki pages, perhaps during the down-times in their vacations. However, by the end of the summer – as most other trips, projects, internships, and the like are finishing or becoming routine, and as the hottest weather of the year confines people indoors – students might be more likely to spend the time necessary to

compose a new page from scratch. In any case, it seems fairly clear that the primary user group of Wikia is university students, probably those living in the northern hemisphere.

Overall, page creation volumes on both sites seem to have peaked between 2005 and 2007, towards the end of SG-1 and the first half of Atlantis, and have been gradually decreasing. The occasional DVD movies and new Universe series appear to only be motivating users of these sites enough to either maintain or gradually decrease their participation.

### 4.3.2 Citing sources and identifying authors

**RA2:** What types of sources and authors are cited on fansites?

With the exception of the occasional photo taken by a fan of a cast member at a convention, as well as the first-hand descriptions that occur in plot summary texts, most of the pages on these sites were found to be historiographical in nature – compiling an account of the past doings of a person, media artifact, or event based only on documentary evidence from websites, popular print media, and occasionally academic literature. As such, there was no appreciable difference between citing sources and citing evidence. In §§4.3.5 and 4.3.11, types of evidence and link destinations are presented in depth. Therefore, this section will only give an account of to what degree, and in what way, each site identifies the authors of its pages.

**GateWorld** had quite a few sources of content besides the editors. Fans, who were cited either by name or username, were the usual source of lengthy reviews and transcripts for individual episodes. They also posted comments on the news blog, submitted videos to the videos podcast and archive, and were the source of fan fiction and forum posts archived

on the site. The "production" sections of episode guides also identified authors from the mass news media, other fansites, cast and crew sites, and the GateWorld editors' interviews with crew members as their sources. Book and comic pages identify the work's author, and typically had a summary provided by the publisher, which was identified. Game pages did the same for/with the game's developing company. Interviews conducted by the GateWorld editors identified both the interviewing editor and the interviewee by name. Photos in the gallery section gave credit to their owners, unless a GateWorld editor was the owner. For example, promotional photos had credits to the studio, and all episode screen capture pages contained a disclaimer saying that they were not provided for re-sale and were owned by MGM (the studio). All other content on the site was created by the editors, including: most news blog posts, all omnipedia articles, a video podcast (except for fan-submitted videos), and audio podcasts and their accompanying editorial notes.

**IMDb**, as one might expect from their insisting that users sign over copyright, was considerably more non-attributive about their content. Plot summaries, title/episode reviews, message board posts, biography paragraphs about cast and crew members, and news blog entries posted by fans were the only content to which a fan's username was attached. News blog posts that referenced another website or mass media outlet also identified that source, the provider of TV listings for the show was identified, and lists of interviews and articles with/out about a cast or crew member identified the professional interviewer or writer as well as the company that published the work. A number of fields were created by fans, whose work was not credited. Lists of genres, quotes, links to relevant external sites, advice to parents on a title's appropriateness for children, awards, user ratings, recommended similar titles, trivia, technical/production details, filmographies, cast and crew lists, and

connections between titles and actors were all compiled to some degree by registered users of the site, because they were all user-editable, but no credits or history of edits to those lists was available, even if one registered and logged-in to the site. Also, if any information came from the studio or network, it had likely been licensed for use without attribution, for there were no attributions to MGM or the SyFy Channel on individual title or cast/crew pages either. The only mention given to unnamed users and companies was on one of the Help pages (IMDb, 2009e), which said that they use "as many sources as we can get our hands on," followed by a long list of public and private sources that mentioned no parties in particular.

**Both wikis**, by contrast, used the same MediaWiki software and its default attribution practices, which attempt to make transparent the authors of all content appearing on a wiki site. All pages were created and edited only by users and administrators, whose usernames or IP addresses were attached to every edit, as viewable in each page's revision history. Usernames could link to the profile pages of individual users, where more information about the user was occasionally available, at the user's discretion. Many pages contained bibliographies, using various style guides, with references to print materials mentioned in the page's text. Also, lists of links believed by users to be relevant, to both internal and external sites, were available towards the end of many pages on both sites. Similarly, all multimedia objects appearing on the pages contained their own pages that provided copyright, ownership, revision history, and technical details about the object. In general, other than the webpage template created by the sites' administrators as well as the third party advertisements appearing on each Wikia page, the researcher was not able to find content on either wiki site without either a username or IP address as attribution.

In summary, the editor-controlled sites generally outsourced the most laborious aspects of content creation to users, and reserved either the more glamorous tasks or high-profile content for themselves, even if created by a user. For example, the GateWorld editors contributed most interviews with cast and crew members, convention coverage, audio-video podcasts, and commentarial episode guides. Lengthy episode reviews and transcripts were left to users. The only content on the site not fitting this pattern were the omnipedia pages, which evinced, on the parts of the editors, a kind of loving obsession for the minutiae of the franchise that is probably peculiar to small, volunteer fansites. IMDb's editors, other than giving others credit for lengthy writing, allowed users to edit any part of pages, but did not identify which information had been compiled by users and which by themselves. By contrast, both wikis gave up most editorial privilege in exchange for asking fans to edit and compile all content on their sites. Any fan could have their name attached to any part of the page's content, and the sites' editors took on only administrative roles. Of course, administrators could contribute as much content as they wished, though their credit for any contributions would be no greater than that received by normal users.

### 4.3.3 Original research

**RA3:** What forms and contents does original research take on these fansites?

As also argued in §4.3.2, if any "research" beyond mere subjective description occurs on the fansites under study, it is of a historiographical nature, in which an impartial account of a phenomenon occurring in the past is assembled from both first-hand and documentary evidence. The primary impartializing feature of these sites is the ability of users to write

over each others' content, making the content, at worst, an inter-subjective description, at least for well-edited pages. Of course, the mechanisms for ensuring the qualifications of the writers, adequate peer review, and that a certain field's reporting standards are upheld are all largely absent, which means that, even if the content resembles an historiographical account, that account is dubious. This section *does not* argue that a form of research that the academie would find acceptable or original does occur on fansites, but merely explores to what degree *research-like* activities are evident from the fansites' contents. No site under study either referenced or ensured standards of methodological rigor, nor combined its research-like activities with an editorial and peer review social infrastructure adequate for considering the sites' products "research" in an academic sense. It is only popular investigation.

Following §3.3, a hermeneutic content analysis was conducted of a random sample of pages from all four sites, with a sample size adequate for principal components analysis (PCA). From that content analysis, six variables were defined and operationalized for this research question, namely: interpretations and opinions, identifies unanswered questions, gives production details, identifies cultural references, discusses biographical or historical context, and discusses critical reception. See appendix A, table A.14 for a codebook. In the following results sub-sections, both descriptive statistics and PCA results will be given for the variables present on each site.

### Descriptive statistics

On all four websites, because all of the variables were measured as presence-absence/binary, only counts and percentages of presences and absences are available for descriptive statis-

tics. Table 4.16 presents the sample sizes (the same as those given in §3.3), counts, and percentages of those variables that were present on each site.

Table 4.16: Descriptive statistics: Original research variables

| | **GateWorld** | **IMDb** | **Wikia** | **Wikipedia** |
|---|---|---|---|---|
| interp | 30 (9%) | 9 (3%) | 21 (6%) | 10 (4%) |
| unansQuest | 20 (6%) | – | – | – |
| prodDet | 34 (10 %) | 34 (11%) | 28 (8%) | 185 (79%) |
| cultRef | – | 14 (4%) | 15 (4%) | 17 (7%) |
| bio | – | 55 (17%) | 5 (1%) | 165 (71%) |
| reception | – | – | – | 52 (22%) |
| n | 346 | 320 | 353 | 234 |

This table shows that, over all of the sites, content that involves research-like activities is present, though relatively rare. The only exception to this are sections containing production and biographical details on Wikipedia. Regarding interpretations and opinions that go beyond mere description, the small and franchise-devoted sites, GateWorld and Wikia, contained the most, and the larger and more generic contained 2-3 times less. Only GateWorld's editors identified unanswered questions in the plot line. Besides the frequency of production details on Wikipedia, the edited sites contained slightly more than Wikia. This shows that, while all of the sites make a small effort to gather production information, this is one of Wikipedia's strengths. All of the sites except GateWorld occasionally had sections listing non-franchise-related cultural references present in the shows, with Wikipedia hav-

ing slightly more than the others. Wikipedia again dominated over the other sites in terms of biographical information, with IMDb a distant second, and small fansites almost none. Finally, only Wikipedia had sections explicitly devoted to the critical reception of individual topics, though one might also find links and references to such information, without discussion, on the "publicity" sub-pages linked-to from many IMDb pages.

In summary, whereas the larger sites dominated the gathering of technical production, biographical/historical, and critical reception information, the smaller sites provided more of their own critical interpretations.

**Modeling results: Principal components analyses**

PCAs were conducted for each website, in order to describe the correlative structures of variables within each site. This allows one to say which research-like sections tended to occur together or not on the pages of each site.

On GateWorld, one component had an eigenvalue above one, and isotropic variance began with component two, so two components will be interpreted. In PC1, unanswered questions and interpretations had similar loadings, and were contrasted with production details. PC2 merely provided the reverse of PC1. This means that unanswered questions and interpretation sections tended to occur together on GateWorld, but rarely on the same page as production details. This finding echoes the production vs. interpretation dichotomy also seen in the descriptive statistics.

On IMDb, three components were worthy of interpretation. In the first, interpretations, production details, and cultural references loaded similarly, and were contrasted with biographical information. This means that quite a few pages contained only biographical

information. Anecdotally, this seemed to be the case for many actor pages. PC2 contrasted cultural references most with biographical, and secondarily with production, information. This suggests that pages containing cultural references have content that approaches the interpretations found on smaller fansites, which seems intuitively correct, as cultural reference content was compiled by fans. And PC3 contrasted cultural references most with production information, which affirms the same conclusion.

Wikia's data contained two dimensions of noteworthy co-variance. In the first, the production and cultural reference variables were contrasted with the biographical and interpretation variables. In the second, biographical information was contrasted with interpretations. These spectra suggest that, within Wikia, the cross-site distinction between technical and interpretive information is not so stark. When production information was given, it was often alongside cultural references, and the same for biography and interpretation sections. This distinction might be due to the first two sections typically being more listy, and the second more prosaic. However, the second PC points to a split between biographical and interpretive sections, possibly because biographies were more common on pages about individuals and groups of people, whereas interpretations were more common on pages about titles, plot lines, and events.

Finally, three dimensions were noteworthy on Wikipedia. The first contrasted cultural reference and production information with interpretations, a distinction also seen in Wikia's PC1. PC2 contrasted production and biographical variables with cultural reference and interpretation variables. This echoes the large vs. small site split seen several times earlier. And, PC3 contrasted interpretation with reception sections, indicating that pages that focused on the public reception of something provided little interpretive critique of their own.

In conclusion, a number of sites echoed the distinction between technical and interpretive information also found in the descriptive results. Additionally, IMDb had many pages containing only biographical or historical information, and its cultural reference sections approached the kinds of user interpretation sections found on smaller sites. Within Wikia, more listy information sections also tended to be separated from more narrative information sections. Also, biographical and historical prose may have been more common on people-oriented pages and interpretive sections more common on plot- or event-oriented pages. Finally, on Wikipedia, those pages that focused on the public (e.g., mass media) reception of something rarely provided interpretative critiques of their own.

### 4.3.4 Objectivity as impartiality

**RA4:** To what extent do accepted descriptions and interpretations exist in fan reviews and analyses on the same topic across the sites?

The question is, since the IQ literature defines objectivity in terms of inter-rater agreement (i.e., agreement between independent experts), and, given some set of common topics to evaluate, to what degree, and about what, do the four website's accounts, as measured by trained coders, agree or disagree? Essentially, trained coders, whose reliability was ensured (§3.3), determined whether a thematic variable was present on a given site, and the agreements between sites were assessed using the most robust agreement measure available. Hence, the analysis process had the following two stages.

First, after the dataset had been created, Krippendorff's $\alpha$ agreement metric was applied to each set of four observations, one from each site, on a number of topics. This metric was

chosen despite Cohen's $\kappa$ (Cohen, 1960) being more common in the biomedical literature that pervades the literature review chapter (e.g., Bernstam et al., 2005), because Krippendorff's measure superseded Cohen's by a decade, and can be shown to be more robust than Cohen's to several common disagreement patterns (Krippendorff, 2004, pp. 246-247). Krippendorff's $\alpha$ is also still the standard agreement metric in content analysis, and has a statistically rigorous formulation (Krippendorff, 2004, pp. 223-256), which has been shown to generalize several earlier metrics intended for specific scenarios, namely: Fleiss' $\alpha$ (Fleiss et al., 1971), Scott's $\pi$ (Scott, 1955), Spearman's rank correlation $\rho$ (Spearman, 1904), and Pearson's intraclass correlation (Pearson, 1901).

In addition to merely being cautious, the use of as general and robust of an agreement measure as possible was desirable because: the thematic variables were not prescribed from the literature, but rather they hermeneutically emerged from the research process, and some of the summed agreements were small (Herring, 2007, p. 4). On the other hand, the samples coded were the exhaustive set of matching pages from the sites, which is unusual in the analysis of Web data, and Krippendorff's measure is robust for small sample sizes and binary data. Finally, for the purposes of this analysis, following Krippendorff's advice (Krippendorff, 2004, pp. 241-243), $\alpha$ values above 0.8 were considered "high" agreement, 0.67-0.79 "medium" agreement, and below 0.67 "low" agreement. The $\alpha$ values were computed with the `kripp.alpha` and `coincidence.matrix` functions in the `concord` package (Lemon and Fellows, 2009) for `R`.

Second, the metrics' values were interpreted for degree to which, and about what, the sites agreed or disagreed. Also, recall from §3.3 that only title and character pages were examined. Finally, more complex analyses (e.g., multi-level regression, multi-level PCA,

etc.) were not done because the samples were much too small and the number of variables too large for most modeling methods, and because a descriptive analysis of agreement metric results was sufficient for answering the research question.

### Title pages

After computing the $\alpha$ values for each pair of sites on each topic, as well as the overall $\alpha$ value of taking all four sites together, the values for each pair of sites and for the sites overall were tabulated, and their descriptive statistics calculated. Table 4.17 presents those statistics.

Table 4.17: Descriptive statistics: Krippendorff $\alpha$ values of title pages, both between pairs of sites and across all sites ($n$ = 32 matching pages)

| | **gw-imdb** | **gw-wikia** | **gw-wp** | **imdb-wikia** | **imdb-wp** | **wikia-wp** | **overall** |
|---|---|---|---|---|---|---|---|
| stdev | 0.26 | 0.24 | 0.99 | 0.25 | 0.2 | 0.18 | 0.15 |
| min | 0 | 0 | 0 | 0 | 0 | 0.43 | 0.18 |
| 25% | 0.17 | 0.48 | 0.4 | 0.34 | 0.26 | 0.56 | 0.38 |
| median | 0.4 | 0.64 | 0.59 | 0.5 | 0.44 | 0.65 | 0.53 |
| mean | 0.38 | 0.56 | 0.71 | 0.46 | 0.39 | 0.69 | 0.5 |
| 75% | 0.6 | 0.71 | 0.71 | 0.61 | 0.55 | 0.85 | 0.64 |
| max | 0.82 | 0.93 | 6 | 0.89 | 0.89 | 1 | 0.75 |
| skew | -0.07 | -0.66 | 5.17 | -0.3 | -0.09 | 0.4 | -0.28 |
| kurt | -1.43 | -0.03 | 28.3 | -0.44 | 0.17 | -1.12 | -0.93 |

In the descriptive statistics, there were three patterns. GateWorld and IMDb (*mean* =

0.38), as well as IMDb and Wikipedia (*mean* = 0.39), agreed the least. GateWorld and the wikis (*mean* = 0.54 – 0.56), as well as IMDb and Wikia (*mean* = 0.46), agreed moderately. And the wikis often agreed (*mean* = 0.69). These figures consistently follow neither editorial nor size divisions. Indeed so much mixture is apparent that perhaps the sites either imitated or incorporated the content of those sites that were least like themselves. The degree of copying vs. content dissimilarity should become more apparent in the following paragraphs, when especially high- or low-agreeing records will be examined more closely. Regarding skew and kurtosis, the most negatively/right skewed were the GateWorld-wiki pairs, and the Wikia-Wikipedia pair was the most negatively skewed, indicating that the GateWorld-wiki pairs usually spanned more of a moderate-to-high level of agreement, and the wikis a more moderate level of agreement, than the means alone would suggest. This only strengthens the idea that sites of different types cross-pollinate their content. The kurtosis figures similarly say that sites of similar editorial types have less definite peaks around a certain agreement level.

Now to examine specific cases of high and low agreement, beginning with high. The only site pairs having high agreement levels were GateWorld-Wikia and Wikia-Wikipedia. Indeed, for most variables, a kind of triad formed, in which all three sites contained title pages having nearly the same themes, namely: military/honor, technology/weapons, slavery, (demon) possession, romance, in-jokes, science/nature, invasion, explosions, cultural homages and references, federal governments, treasure hunting, lying/betrayal, plans for a sequel, cast and crew motivations, cast and crew personal preferences, special effects, and production mistakes. Most of these are the thematic highlights that one might expect from this franchise. Also, several variables were consistently present only between the

wikis, namely travel and medicine, possibly indicating content biases towards those things. Although a close qualitative reading would be required to estimate the exact degree of content copying/sharing between these three sites, the present analysis can say that there were many thematic similarities between their texts. Anecdotally, Wikipedia, which usually had relatively short pages, occasionally appeared to plagiarize the first few sections of introductory text verbatim from Wikia, which usually had longer and more detailed pages. No other such copying behavior was observed, nor did Wikia noticeably either copy or expand upon Wikipedia pages.

In terms of low agreement, IMDb consistently disagreed with the other three sites on those variables that the other three sites often shared. For example, IMDb disagreed with all and exactly the variables present on Wikia in every record. With GateWorld and Wikipedia, IMDb disagreed with all of the standard variables as well as included several production-related variables not present on the other three sites, namely: budget, deleted scenes, filming locations, critical reception, and quotes. This suggests that, on title pages, IMDb followed a markedly different content paradigm from the other three sites, focusing more on production technicalities, awards, and compilation of quotes than on lengthy textual content explicating the themes present in that title. GateWorld also occasionally disagreed with the Wikis on several variables, namely: archæology, linguistics, possession, explosions, research, production budget, deleted scenes, and quotes. This suggests that GateWorld focused more on representations of academia, explosive special effects, and esoteric production details than did the wiki sites.

### Character pages

A comparable table of descriptive statistics for character pages is provided in table 4.18.

Table 4.18: Descriptive statistics: Krippendorff $\alpha$ values of character pages, both between pairs of sites and across all sites ($n$ = 21 matching pages)

| | **gw-imdb** | **gw-wikia** | **gw-wp** | **imdb-wikia** | **imdb-wp** | **wikia-wp** | **overall** |
|---|---|---|---|---|---|---|---|
| stdev | 0.22 | 0.2 | 0.23 | 0.21 | 0.22 | 0.24 | 0.15 |
| min | 0 | 0 | 0 | 0 | 0 | 0 | 0.11 |
| 25% | 0.2 | 0.35 | 0.33 | 0.02 | 0 | 0.37 | 0.21 |
| median | 0.4 | 0.47 | 0.48 | 0.12 | 0.2 | 0.48 | 0.35 |
| mean | 0.37 | 0.46 | 0.44 | 0.21 | 0.21 | 0.44 | 0.35 |
| 75% | 0.56 | 0.63 | 0.55 | 0.42 | 0.35 | 0.55 | 0.47 |
| max | 0.69 | 0.78 | 0.85 | 0.6 | 0.63 | 0.84 | 0.64 |
| skew | -0.44 | -0.44 | -0.27 | 0.56 | 0.4 | -0.58 | 0.12 |
| kurt | -0.95 | 0.18 | 0.07 | -1.35 | -1.34 | -0.23 | -1.2 |

On the character pages, the sites' textual themes disagreed more than they agreed. IMDb's themes agreed the least with both wiki sites' (*mean* = 0.21, in both cases), GateWorld and IMDb agreed moderately (*mean* = 0.37), and GateWorld and the wikis as well as the wikis with each other agreed the most (*mean* = $0.44-0.46$). The highest agreements, around 0.85, were between GateWorld-Wikipedia and Wikia-Wikipedia. This suggests that GateWorld sits on a spectrum between IMDb and the wikis. Regarding skew and kurtosis, similarly, IMDb's relations with the wikis were the only pairs having a positive/left skew, with Wikia's being slightly larger than Wikipedia's. However, the kurtosis of those pairs

was quite flat. This indicates that, though IMDb consistently had low agreement with the wikis, those agreement scores do not settle around a very definite value. By comparison, the other pairs were negatively/right skewed and had more positive kurtosis, indicating a more positive and consistent relationship between them.

The few agreements were as follows. GateWorld and Wikipedia often contained themes of military decoration, superiors, colleagues, biographies, characters' characteristics, medicine, science/nature, romance, and homage. The wikis, by contrast, agreed most on themes of superiors, age, gender, alien contacts, medicine, explosions, and cast and crew motivations. Though difficult to distinguish, GateWorld and Wikipedia's agreements seem more to do with substantive aspects of characters' contexts, backgrounds, and development, whereas the agreements between wikis seems more related to the superficial characteristics that their Infoboxes or introductory/summary paragraphs might share.

At least seven low/dis-agreement patterns existed between the sites, given here in order of how many variables were involved in each pattern, from most to least. In the first, all of the sites except the wikis had difficulty agreeing on the inclusion of complex, substantive, and eventful topics such as biographies, characters' characteristics, science/nature, large explosions (wiping out civilizations), romance, and cast and crew motivations. Perhaps the wikis only did not disagree because of Wikipedia's copying behavior? The second pattern showed disagreement over including more perfunctory themes, such as rank, superiors, colleagues/team, station/post, and technologies used by the character. The wikis still did not disagree, and GateWorld did not disagree with Wikia's inclusion of these themes, though it did with Wikipedia's. Third, most sites disagreed on whether to present three quirky topics: treasure hunting, in-jokes, and cultural homages/references. GateWorld again did not

disagree with Wikia, and also IMDb. Fourth, two variables on medicine and alien contact – alien wildlife and races are often a source of new medical advances on Earth in the shows – were always avoided only by IMDb. Fifth, two variables on fan followings and public critical reception showed disagreements between Wikipedia and the edited sites, though Wikia did not disagree with Wikipedia. Whereas Wikipedia usually devoted a section of text to discussing such issues, the edited sites preferred to link to other fan and media sites, causing them not to appear in this text-only analysis. Sixth, the nicknames and quotes variables showed disagreement between all pairs of sites except GateWorld-Wikipedia. This is strange, because there is no clear connection between the two sites. Finally, none of the sites could agree on how/whether to represent a character's age. Some did it passingly in text, others in Infoboxes, and others in tables and lists.

**Conclusion**

The following are the extents to which accepted descriptions and interpretations existed in fan review and analysis texts on the same topics across the sites.

On the title pages, a triad of sites, including GateWorld and the wiki sites, consistently agreed on which core themes to include in their texts. Anecdotally, the core of this triad seemed to be sites that were smaller and more devoted to the franchise, namely GateWorld and Wikia, with Wikipedia occasionally plagiarizing the highlights of Wikia. The wikis also focused more on themes of travel and medicine than did the other sites. IMDb focused on the compilation of production technicalities, awards, and quotes, whereas the triad of sites more on lengthy textual explications of complex themes. GateWorld also had a particular focus on depictions of academia, explosion special effects, and esoteric production

details.

On the character pages, there existed a spectrum of sites – IMDb :: GateWorld, Wikia / Wikipedia – with GateWorld being thematically closer to the wikis than to IMDb. GateWorld and Wikipedia agreed on including substantive aspects of characters' contexts, and did not disagree on including sections on nicknames and quotes. GateWorld and Wikia did not disagree on including more perfunctory topics, such as characters' ranks and stations; and GateWorld did not disagree with Wikia and IMDb on the inclusion of several fun and quirky topics. The wikis agreed on including cursory/Infobox details, and did not disagree on including either substantive or perfunctory topics. IMDb generally avoided the topics of alien contacts and medicines. The editor-controlled sites differed with Wikipedia on how public/critical reception details should be presented, with the edited sites preferring lists of citations and Wikipedia preferring discussion paragraphs. Finally, none of the sites could agree on when or how to include a character's age.

### 4.3.5 Empiricality as verifiability

**RA5:** What types of evidence are cited by articles on fansites?

This question was answered in the same way as the question of original research in §4.3.3, but on different binary variables from the same dataset (see Appendix A, table A.13). In the following sub-sections, descriptive statistics for these binary variables are presented and interpreted, followed by principal components analyses of the data available for each site.

**Descriptive statistics**

Table 4.19 presents the descriptive statistics available, namely counts and percentages, for these binary variables.

Table 4.19: Descriptive statistics: Evidence variables

| Variable | GateWorld | IMDb | Wikia | Wikipedia |
|---|---|---|---|---|
| local | 339 (98%) | 320 (100%) | 352 (100%) | 234 (100%) |
| gateworld | – | – | 45 (13%) | 60 (26%) |
| imdb | 156 (45%) | – | 32 (9%) | 178 (76%) |
| wikia | – | – | – | 20 (9%) |
| wikipedia | – | – | 41 (12%) | – |
| publisher | 7 (2%) | – | – | – |
| users | – | 61 (19%) | – | 16 (7%) |
| studio | – | – | 24 (7%) | 62 (26%) |
| tvnetwork | – | – | 17 (5%) | 38 (16%) |
| adsite | – | – | 5 (1%) | – |
| sgwiki | – | – | 11 (3%) | 9 (4%) |
| rda.com | – | – | 5 (1%) | – |
| sg1archive | – | – | 1 (0%) | – |
| savesg1 | – | – | 1 (0%) | – |

Continued on Next Page...

Table 4.19 – Continued

| Variable | GateWorld | IMDb | Wikia | Wikipedia |
|---|---|---|---|---|
| fansite | – | – | 8 (2%) | 34 (15%) |
| indexsite | – | – | 4 (1%) | 96 (41%) |
| books | – | – | 18 (5%) | 31 (13%) |
| producer | – | – | 6 (2%) | 27 (12%) |
| DVDextras | – | – | 4 (1%) | 8 (3%) |
| massmedia | – | 32 (10%) | 2 (1%) | 96 (41%) |
| personal | 2 (1%) | – | 1 (0%) | 71 (30%) |
| acad | – | – | – | 2 (1%) |
| n | 346 | 320 | 353 | 234 |

This table shows that the wiki sites generally cited a larger variety and quantity of evidence than did the editor-controlled sites. Other than local references (i.e., references to other pages on the same site), which every site included on nearly every page, very little evidence was cited by the editor-controlled sites. GateWorld, which had pages for characters but not for actors, frequently referred users to IMDb's actor pages. Otherwise, GateWorld only cited a very few publishers' and actors' personal websites. IMDb's only non-self citations were to usernames on its message boards and plot summaries as well as

publicity related citations – to mass media interviews, articles, and the like – on the most successful actors' pages. Overall, GateWorld only used citations either to support quotes and claims about production details or to refer the user to another editor-controlled site. IMDb similarly made little attempt to support the claims on its main pages with primary source citations, ignored the other Stargate sites, and only referred users to large corporate mass media sources.

The wikis, by contrast, both had notably large numbers of citations to the editor-controlled sites, as well as to many types of evidence other than the mass media. However, Wikia cited GateWorld more often and Wikipedia IMDb, indicating a bias among sites in favor of other sites of similar size. Wikia also cited Wikipedia more often than vice versa, perhaps due to Wikipedia's greater popularity. Also, though Wikia had a greater range of evidence types, Wikipedia had considerably higher counts of a subset of those sources also cited by Wikia, namely to: the mass media, index sites (i.e., sites hosting celebrity profiles), cast and crew members' personal sites, the studio's site, the TV networks' sites, non-Stargate-related fansites, non-academic books, and the producer's weblog. Wikipedia also cited two academic works, a physics article in arXiv.org on wormholes (Nandi et al., 2004) and the edited volume of critical readings of themes in SG-1 by Beeler and Dickson (2006). These points suggest that the wikis were considerably more open to any source of evidence than were the editor-controlled sites, though the smaller wiki contained more references to small and obscure sources, and the larger wiki primarily referenced more mainstream sources. The first of these could be due to differences in organizational philosophies, between those that believe that socially emergent structure creates business value vs. restricting users' immediate interface options creating value. The latter observation could

be due to larger organizations feeling more public pressure to provide politically correct content.

**Modeling results: Principal components analyses**

For GateWorld, two principal components had eigenvalues above 1.0 and isotropic variance began with PC3, so three PCs will be interpreted. In the first, the local and publisher variables were set in contrast, indicating that the few pages that did not cite other GateWorld pages probably cited publishers (i.e., were book or comic pages). In the second PC, which accounted for a similar amount of variance as the first, the personal homepages and imdb variables were contrasted with the local variable. This suggests that those pages that cited other sites in regard to actors (i.e., omnipedia character pages) also were somewhat less likely to include self-references back to GateWorld. Finally, in the third PC, the imdb and personal homepage variables were set in contrast, suggesting that, if an omnipedia page linked to an actors' IMDb page, it was less likely to also link to the actor's personal homepage. This is because, typically, only one "played by" link was given on those pages.

As shown in the previous sub-section, IMDb had only three evidence variables and one of them was constant across all pages. This means that only two variables contained variability, making simple pairwise correlation more appropriate than PCA. That correlation was $r = 0.55$, indicating that pages that cited the mass media (i.e., popular cast and crew pages) also tended to have many users discussing and writing about them, which is not surprising. Nevertheless, running a PCA on these two variables, and using page types as a grouping factor in the data matrix's row names, also showed that title pages garnered more discussion than did cast and crew pages. While perhaps also not surprising, this is

interesting to have shown empirically.

Wikia, having many more variables, had seven PCs worth interpreting. As in §4.2.5, the following list contains the sorted loadings for each component, each decreasing from positive to negative.

1. studio gw sgwiki tvnetwork rda.com sg1archive savesg1 adsite producer DVDextras personal wp book massmedia local fansite indexsite imdb

2. personal massmedia fansite producer imdb indexsite tvnetwork wp local adsite DVDextras book rda.com gw studio sgwiki sg1archive savesg1

3. savesg1 sg1archive fansite imdb indexsite sgwiki personal massmedia local wp rda.com producer book studio gw DVDextras adsite tvnetwork

4. indexsite fansite imdb rda.com sgwiki gw DVDextras studio wp local adsite book tvnetwork savesg1 sg1archive producer personal massmedia

5. rda.com sgwiki producer wp massmedia personal local gw imdb fansite DVDextras indexsite studio savesg1 sg1archive book tvnetwork adsite

6. adsite rda.com tvnetwork studio producer imdb indexsite sgwiki fansite massmedia sg1archive savesg1 personal local gw book DVDextras wp

7. local book adsite rda.com imdb producer sg1archive savesg1 massmedia sgwiki personal wp gw studio tvnetwork fansite indexsite DVDextras

PC1 contrasts types of evidence affiliated with the studio (e.g., studio, gateworld, sgwiki, and tvnetworks) with those more affiliated with index sites (e.g., imdb, indexsite,

fansite, and local). GateWorld and the smaller Sgwiki have an obvious history of trying to ensconce themselves within the production fold, in order to have an inside track, and IMDb could be viewed as a high-quality version of a celebrity profile/index site. PC2 contrasts professional sources, such as mass media sites and the homepages of cast members, with obviously amateur sources, such as the now-defunct Save SG-1 fan campaign website. PC3 contrasts the same amateur category with the sites run by TV networks, which tended to be filled with advertisements. PC4 contrasts the same professional and index site categories already described. PC5 contrasts TV networks, advertisements, and the studio with small devotional sites, the epitome of which is rdanderson.com, a "shrine" site made by a fan of the actor Richard Dean Anderson, who also is occasionally an executive producer for the franchise. Anderson has endorsed that fansite as being his official personal website, and became friends with the fan who created and maintains the site. In PC6, advertisement sites are contrasted with more mainstream and documentary sources, such as Wikipedia, official DVD extras, and books. Finally, PC7 contrasts Wikia's self-citations with sources that are especially non-wiki or non-Wikia affiliated, such as DVD extras and index sites.

Wikipedia also had seven PCs worthy of interpretation. As before, their sorted loadings can be found in the following list.

1. imdb personal fansite indexsite book massmedia sgwiki acad user DVDextras wikia tvnetwork studio producer gw

2. wikia tvnetwork producer book massmedia gw personal studio fansite user acad indexsite sgwiki DVDextras imdb

3. sgwiki gw user producer imdb studio wikia tvnetwork DVDextras fansite acad in-

dexsite massmedia book personal

4. acad DVDextras book sgwiki wikia personal producer gw studio imdb indexsite massmedia tvnetwork fansite user

5. wikia tvnetwork acad imdb fansite personal user DVDextras producer gw indexsite massmedia sgwiki studio book

6. fansite wikia indexsite sgwiki personal studio DVDextras book gw tvnetwork acad user producer massmedia imdb

7. personal book user sgwiki studio producer fansite imdb gw wikia tvnetwork DVDextras acad massmedia indexsite

The largest amount of variability found on Wikipedia is a dichotomy between the personal indexing sites by/about actors (e.g., IMDb, which offers casting services to actors, and actors' personal websites) with, like Wikia, more studio-affiliated and general-index sites, such as GateWorld, the producers' blog, the studio's and network's sites, and Wikia. PC2, by comparison, contrasts sources that are more superficial and targeted towards outsiders to the franchise (e.g., IMDb's focusing on studio credits, DVD extras, and index sites) with those that provide more well-documented information targeted towards insiders (e.g., Wikia, the network's and producer's sites, and books about or based on the franchise). In PC3, the same amateur vs. professional dichotomy seen on Wikia is again in evidence. PC4 contrasts both amateur and professional media sources with academic sources, which, interestingly, are set alongside other commentarial sources, such as DVD extras. PC5 contrasts small-scale producers of media, academia among them, with large-scale media pro-

ducers, such as fiction books and the MGM company. PC6 similarly contrasts small media indexers, such as fansites, with large media indexers, such as IMDb and typical mass media sources (e.g., news, interviews, articles, etc.). Finally, PC7 distinguishes professionally made sources into those with content that is relatively deep, such as personal websites and books, versus those that are relatively shallow, such as index sites and the mass media.

**Conclusion**

All sites cited themselves on nearly every page, indicating that they wish to keep users on their site, though none was as devious in this attempt as are most social networking, gambling, and pornography sites. Editor-controlled sites were the least outward-looking, using references only to either support quotations and claims about something or to refer users to another editor-controlled site, often an affiliate. In this sample of pages, editor-controlled sites also never linked to open wiki sites. The wiki sites had greater variety and quantity of sources, including to the editor-controlled sites. Whereas Wikipedia cited a few large/mainstream sources many times, Wikia cited many small sources only a few times. Wikipedia was also the only site to cite academic literature, and that was only two Stargate-related works in the fields of physics and critical media studies.

Additionally, principal components analysis was used to categorize types of sources that were included together on the pages of each site, though this was most fruitful on the wiki sites, which cited a greater variety of sources. GateWorld showed different citation patterns for self-references, references to publishers, and person-specific vs. general indexing sites. On IMDb, it was apparent that cast and crew with many mass media citations (i.e., popular cast and crew) also had many users discussing them. Also, titles were more discussed than

people on IMDb.

On both Wikia and Wikipedia, studio, TV network, and advertisement sources often occurred together, as did sources that were notably indexical, professional, and amateur. However, on Wikia, amateur sites could also be seen as devotional, professional sources could be richly documentary, and self-references were contrasted with references that were very different from the wiki paradigm. On Wikipedia, both indexical and documentary sources could be divided into small, large, and somehow special or thoughtful. Small sites depicted and advertised a richly informative experience targeted towards fans, whereas large sites offered a high-level view based largely on information from the production and marketing companies. Furthermore, like GateWorld, index sites could be personally specific to the person being indexed (e.g., actors' personal homepages). But, unique to Wikipedia, documentary sources could also be academic, which occurred alongside references to other forms of non-academic commentarial literature.

### 4.3.6 Consistency in concepts and styles

**RA6:** To what extent does a consistent set of data fields and stylistic motifs emerge across the sites?

As discussed in §3.3, in addition to studying the variables prescribed by the IQ literature closest to this topic, because this project is the first major study of the IQ of science fiction fansites, additional variables were sought hermeneutically from each site under study, in order to represent the sites as fully as possible. Having collected a fairly exhaustive set of variables in this context, this section's research question could be understood as a

kind of meta-question, seeking the most consistently occurring variables across the sites. Hence, to answer this question, binary incidence matrices were constructed of which variables occurred on which sites and their subsections *to any degree*. Although principal components analyses could be done of every variable type, the small number of site sections/observations and the large number of variables – especially for the short text field variables – would cause the extremes of the latent dimensions to be poorly differentiated, resulting in large numbers of difficult-to-interpret dimensions. However, because there are few observations, merely comparing the column sums of each variable, and interpreting on what types of sites each variable occurred, should be sufficient to answer the question. Variables that were not data fields on the sites' pages (e.g., PageRanks, inlink counts, etc.) will not be presented.

Additionally, most of the variables were not stylistic in nature, as different sites' usage of specific styles was studied in §4.2.6 using CSS. However, one type of variable not presented in that section or elsewhere will be presented here, namely variables describing the types of pages (e.g., episode transcripts, omnipediac pages on characters, etc.) that occurred in the random sample of all pages, discussed in §3.3. These variables are somewhat style-related, in that they describe different page layout types. Though page types have been used in previous sections as an assumed category of analysis, these additional variables should show page types' relative frequencies across the sites.

For a full description and example of each variable, please see appendix A.

**Date variables**

Date variables' occurrences, from most-to-least, are given in table 4.20. Percentages were determined with respect to the number of site sections (e.g., wikia.books, wp.crew), namely 26, because these variables were studied with respect to the all-pages sample (described in §3.3).

Table 4.20: Frequencies: Date variables, listed row-wise in descending order

| created | lastmod | released | airdate.us | production | airdate.syn |
|---|---|---|---|---|---|
| 17 (65%) | 17 (65%) | 9 (35%) | 3 (12%) | 2 (8%) | 1 (4%) |

The most prominent date variables, created and last-modified dates, as seen in §4.3.1, were only present on wiki pages, as editor-controlled sites did not disclose the history of their editorial processes. Release dates were available on an array of page types. GateWorld and Wikia most often used this variable on book, comic, and game pages; Wikia also used it for episode pages; and the two large sites used it only for episode pages. The airdate.us variable, representing an episode's initial broadcast in the US, was, predictably, only available on episode pages, though IMDb did not include it, perhaps because they considered it synonymous with the release date. Production dates were only given on the wiki episode pages, and syndication airdates were only present on GateWorld's episode pages.

### Evidence variables

Evidence variables' occurrences are given in table 4.21. Percentages were determined with respect to the number of sites, four, because these variables were studied with respect to the random pages sample (described in §3.3).

Table 4.21: Frequencies: Evidence variables, listed row-wise in descending order

| local | imdb | gw | wp | wikia | user | studio | tvnetwork |
|---|---|---|---|---|---|---|---|
| 4 (100%) | 4 (100%) | 3 (75%) | 2 (50%) | 2 (50%) | 2 (50%) | 2 (50%) | 2 (50%) |
| adsite | sgwiki | rda.com | savesg1 | fansite | indexsite | book | producer |
| 2 (50%) | 2 (50%) | 2 (50%) | 2 (50%) | 2 (50%) | 2 (50%) | 2 (50%) | 2 (50%) |
| DVDextras | massmedia | personal | publisher | sg1archive | acad | | |
| 2 (50%) | 2 (50%) | 2 (50%) | 1 (25%) | 1 (25%) | 1 (25%) | | |

All sites cited both other pages on their own sites, as well as on IMDb, as evidence. IMDb was the only site not to cite GateWorld, and both of the wikis cited each other. Only the large sites cited user comments, messageboards, and forums as evidence. Most of the range of evidence types, from the studio to personal variables, were only present on the wiki sites. However, GateWorld was the only to cite publishers, only Wikia cited the SG-1 Archive, and only Wikipedia cited academia.

### Link variables

Link variables' occurrences are given in table 4.22. Percentages were determined with respect to the number of site sections, 26.

Links to official sites occurred most consistently. Actor, author, and crew pages pos-

Table 4.22: Frequencies: Link variables, listed row-wise in descending order

| official | review | forum | miscSite | soundSite | videoSite |
|---|---|---|---|---|---|
| 7 (27%) | 4 (15%) | 4 (15%) | 3 (12%) | 3 (12%) | 3 (12%) |
| amazon | photoSite | firstchp | transcript | itunes | showtimes |
| 2 (8%) | 2 (8%) | 1 (4%) | 1 (4%) | 1 (4%) | 1 (4%) |

sessed them on the two large sites, as did video game pages on GateWorld, and title pages on IMDb. The variables ending in 'Site', in addition to showtimes, were present only on IMDb, which separated links to external websites into categorized sub-pages on all buts its character pages. The remaining variables occurred only on GateWorld. Review and forum links were present on all except omnipedia pages; episode pages had links to transcripts, iTunes, and Amazon; and book pages had links to first chapters and Amazon.

**Long text field variables**

Long text field variables' occurrences can be found in table 4.23. Percentages are out of the number of site sections, 26.

The most consistent pattern across the long text field variables was that only one of them, the long description, appeared on IMDb. By far, the most common long text field type was a lengthy (at least a paragraph) summary of the page's topic, such as a person's biography, which were present on all except some of Wikipedia's list, game, and technology pages. Next most common were production detail discussions, which occurred on episode pages of all sites except IMDb, as well as on Wikia's omnipediac pages and Wikipedia's

Table 4.23: Frequencies: Long text field variables, listed row-wise in descending order

| longdesc | production | shortdesc | notes | previously | plot | characterdev |
|---|---|---|---|---|---|---|
| 22 (85%) | 5 (19%) | 5 (19%) | 3 (12%) | 3 (12%) | 3 (12%) | 3 (12%) |
| pubsum | career | personallife | tech | reception | society | analysis |
| 3 (12%) | 2 (8%) | 2 (8%) | 2 (8%) | 2 (8%) | 1 (4%) | 1 (4%) |
| questions | cheats | authorsum | english | german | gameplay | altreality |
| 1 (4%) | 1 (4%) | 1 (4%) | 1 (4%) | 1 (4%) | 1 (4%) | 1 (4%) |

character pages. Short descriptions of the page's topic occurred most on all GateWorld pages except the omnipedia, and on Wikia's episode pages, but not on Wikipedia. Editorial notes were present on GateWorld's episode pages; Wikia's book, episode, and omnipedia pages; and again not on Wikipedia.

With the exceptions of characterdev, analysis, and questions, which occurred on GateWorld's episode pages, as well as the cheats variable, which occurred only on GateWorld's game pages, all of the variables from 'previously' to the end occurred only on wikis. Also, the plot, publisherssummary, authorsummary, english, german, gameplay, and altreality variables occurred only on Wikia, and the reception and society variables occurred only on Wikipedia. Otherwise, the actual variables' patterns were as one would expect, namely: plot-related variables (previously and plot) were present on episode and omnipediac pages; characterdev was on character and peoples pages; publishers' summaries were on book, game, and video pages; technology was on omnipedia and peoples pages; career and personallife were on cast and crew pages; character and episode pages had the reception vari-

able; author, english, and german text summary variables were on book pages; and gameplay was on game pages. Finally, alternate reality notes were only present on Wikia's omnipediac pages, usually describing a character's dealings in an alternate reality.

### Original research variables

Original research variables' occurrences are given in table 4.24. Percentages are with respect to the number of sites, four.

Table 4.24: Frequencies: Original research variables, listed row-wise in descending order

| interp | production | cultref | bio | unansQuest | reception |
|---|---|---|---|---|---|
| 4 (100%) | 3 (75%) | 3 (75%) | 3 (75%) | 1 (25%) | 1 (25%) |

All websites provided interpretation sections on at least some of their pages, going beyond mere description. All except Wikipedia also provided production details gleaned from outside sources. All except GateWorld identified cultural references from outside the franchise that were present in the shows, as well as compiled biographical or historical information about their topics. Finally, only Wikipedia provided discussion of the franchise's critical reception.

### Page type variables

Page type variables' occurrences are given in table 4.25. Percentages are with respect to the number of sites, four.

Being perhaps the core offerings of fansites about a media phenomenon, every site offered episode and character guides, though IMDb gave characters less prominent attention

Table 4.25: Frequencies: Page type variables, listed row-wise in descending order

| episode | char | peoples | place | tech | sci/nature | language |
|---|---|---|---|---|---|---|
| 4 (100%) | 4 (100%) | 3 (75%) | 3 (75%) | 3 (75%) | 3 (75%) | 3 (75%) |
| book | game | actor | crew | cultref | battles | overall |
| 3 (75%) | 3 (75%) | 3 (75%) | 3 (75%) | 2 (50%) | 2 (50%) | 2 (50%) |
| DVD | template | admin | list | datetime | disambig | transcript |
| 2 (50%) | 2 (50%) | 2 (50%) | 2 (50%) | 2 (50%) | 2 (50%) | 1 (25%) |
| review | makingof | awards | comic | author | demographics | |
| 1 (25%) | 1 (25%) | 1 (25%) | 1 (25%) | 1 (25%) | 1 (25%) | |

than actors. Actors and crew were not covered by GateWorld, though the other small fan-site, Wikia, included them. IMDb was the only site not to cover a large block of page types, from peoples to games, showing their much narrower scope on core topics. They and GateWorld also did not cover the variables from cultref to disambig. Though template through disambig are notably wiki page types, the others contribute to the notion that editor-controlled sites have narrower scopes than wikis. Only GateWorld included transcript, review, making-of, and comic pages. Finally, only Wikipedia included pages about awards received by the franchise, or the real-life features and demographics of people, places, and things in the franchise.

### Short text field variables

Short text field variables' occurrences can be found in table 4.26. Percentages are with respect to the number of site sections, 26.

Table 4.26: Frequencies: Short text field variables, listed row-wise in descending order

| | | | | | | |
|---|---|---|---|---|---|---|
| name | birth | name.alt | spouse | ep.key | title | writers |
| 9 (35%) | 8 (31%) | 7 (27%) | 7 (27%) | 7 (27%) | 6 (23%) | 6 (23%) |
| directors | editors | producers | name.birth | name.nick | height | nationality |
| 6 (23%) | 5 (19%) | 5 (19%) | 5 (19%) | 5 (19%) | 5 (19%) | 5 (19%) |
| trivia | filmAct | filmNonAct | publisher | author | occupation | theatre |
| 5 (19%) | 5 (19%) | 5 (19%) | 5 (19%) | 4 (15%) | 4 (15%) | 4 (15%) |
| char.playedBy | planet.hometo | epnumber | gueststars | composer | died | gender |
| 4 (15%) | 4 (15%) | 4 (15%) | 3 (12%) | 3 (12%) | 3 (12%) | 3 (12%) |
| awards | salary | interviews | author.maincover | genres | size | starring |
| 3 (12%) | 3 (12%) | 3 (12%) | 3 (12%) | 3 (12%) | 3 (12%) | 2 (8%) |
| illustrator | cinematography | education | quotes | trademarks | wherenow | articles |
| 2 (8%) | 2 (8%) | 2 (8%) | 2 (8%) | 2 (8%) | 2 (8%) | 2 (8%) |
| pictorials | covers | author.story | distributor | mediatype | pages | isbn10 |
| 2 (8%) | 2 (8%) | 2 (8%) | 2 (8%) | 2 (8%) | 2 (8%) | 2 (8%) |
| isbn13 | issn | sections.content | worksContrib | runtime | era | (ones |
| 2 (8%) | 2 (8%) | 2 (8%) | 2 (8%) | 2 (8%) | 2 (8%) | omitted) |

Since this question is about consistent/repeated data fields, and because there were many short text field variables, for brevity, those variables only occurring in one site's sub-section have been omitted. The most consistently occurring variable was the name of someone. With the exception of Wikia's only having "name" fields for characters and not actors or crew, all person-related pages on the sites had such a field. Everywhere except GateWorld, Birth, alternate name, spouse, key episode(s), birth name, nick name, height, trivia, filmography, and theater (sometimes using the British spelling) fields typically occurred together, on actor, author, character, and crew pages. Nationality fit in this group as well, though only on the wiki sites. Title, writer, director, editor, and producer fields similarly often occurred together, on episode, book, cast, crew, and game pages.

The rest were smaller patterns. Actors who portrayed characters, and the home-planets of characters, were consistently on character pages. Episode numbers used by the studio during production, and guest star lists, were on all episode pages except IMDb's. Died, gender, awards, size, starring, illustrated, cinematography, and education fields were only present on the wikis. Long lists of quotes, trademarks, where are they now, mass media articles, pictorials, and covers only existed on IMDb's "publicity" pages. Finally, print media metadata fields, such as media type, page numbers, ISBN-10, ISBN-13, and ISSN were only present on Wikia.

### Vendor variables

Finally, vendor variables' (i.e., advertisements) occurrences can be found in table 4.27. Percentages are with respect to the number of sites, four.

Notably, Wikipedia pages contained no advertisements, other than fundraising requests

Table 4.27: Frequencies: Vendor variables, listed row-wise in descending order

| | | | | | | |
|---|---|---|---|---|---|---|
| IT | games | scifi | auto | travel | health | fam |
| 3 (75%) | 3 (75%) | 3 (75%) | 3 (75%) | 3 (75%) | 3 (75%) | 3 (75%) |
| contest | acad | fashion | massmedia | amazon | finance | charity |
| 3 (75%) | 3 (75%) | 3 (75%) | 3 (75%) | 2 (50%) | 2 (50%) | 2 (50%) |
| itunes | pet | security | self | rental | furnishings | |
| 1 (25%) | 1 (25%) | 1 (25%) | 1 (25%) | 1 (25%) | 1 (25%) | |

from their founder or board of directors, so the maximum value of table 4.27 was 3. All of the other sites included advertisements for information technology, games, other science fiction and fantasy franchises, automobiles, travel, health, family and romance issues, contests and sweepstakes, academia (often distance learning courses), fashion, and non-science-fiction mass media franchises. Were one to sketch a psycho-social profile of the sites' users from these variables, one might guess: nerdy males, in their twenties, looking for cars and vacations, needing health insurance, looking for dates and college degrees, caring about their appearance, and being more interested in new- and multi-media than traditional print media.

Smaller patterns were as follows. Only the editor-controlled sites had ads for Amazon and retail finance. Though IMDb is an Amazon subsidiary, GateWorld's Amazon (and iTunes) ads might indicate a corporate partnership. Finance ads might indicate that editor-controlled sites are more financially motivated than wikis. IMDb and Wikia had ads for charity groups; since Wikia is a for-profit company, this might again corporate partnerships.

GateWorld also had a number of pet- and security/military ads; the pet ads seemed strangely out of place, though the security and military themes fit in with the themes of the franchise. Only Wikia had ads showcasing their own other sites, probably because they are the only wiki farm (hosting many sites) under study, as well as home furnishings and general retail. Finally, only IMDb advertised media rental services, usually through their Blockbuster affiliate company.

**Conclusion**

A consistent set of data fields existed on the sites to the following extent.

Regarding dates, though all sites provided release dates in some form, only the wikis provided dates about editorial processes as well as production dates, and only small sites provided the release dates of non-episodes.

Regarding sources of evidence, all sites cited themselves and IMDb, though IMDb did not return the favor only to GateWorld, and primarily the large sites cited usernames within page content. GateWorld was the only one to cite publishers in main page content. The wikis had the greatest variety of evidence variables, and Wikipedia was the only one to cite academic sources.

Links to official sites were most common, though mostly on the large sites. IMDb had sub-sections to/for different types of relevant offsite media, and GateWorld was the only site to have review, transcript, forum and vendor links.

Of long description texts, most pages had at least one lengthy summary text, and that was the extent of most IMDb pages' long text fields. Most episode and character pages on the other three sites had production note texts. Wikipedia had no short summaries

of page contents or sections marked as notes, and "previously" paragraphs – which contain the contents of cast-narrated summaries of relevant past episodes that precede each of Stargate's myth-arc episodes – were unique to Wikia. Character development texts were unique to GateWorld and Wikipedia, and analysis, questions, and cheats (for games) texts were unique to GateWorld. A number of other long text fields were unique to wikis.

All sites possessed at least an interpretive level of original research. Additionally, all except GateWorld identified cultural references and compiled biographies/histories of the topic at hand. All except Wikipedia also collected production details, and only Wikipedia included textual discussions of topics' public critical reception.

Regarding page types, the core content of this media phenomenon is probably episode and character pages. There also existed a set of core page types (i.e., peoples, places, technologies, science/nature, languages, books, and games) represented on all sites except IMDb. Several possibly other core types (i.e., cultural references, companies, battles, franchise-overall, and DVDs) were missing from both of the editor-controlled sites. Only GateWorld had transcript, review, making-of, and comic pages, and only Wikipedia had pages for awards received by the franchise and demographics of real-life phenomena mentioned in the franchise.

The personal names of cast, crew, characters, and authors was the most common short text field mentioned on all of the sites. Otherwise, two groups of common short text fields occurred. In the first, information about birth dates and places, names, spouses, key episodes, height, trivia, and past filmography were common to most cast, crew, character, and author pages on all sites except GateWorld. In the second group, title, writer, director, editor, and producer fields occurred on book, cast, crew, and game pages across all of the

sites. Additionally, a person's nationality was only given its own text field on the wiki sites. Actors who portrayed a character, as well as a character's home-planet, were common on character pages. Episode production numbers and guest star lists were on all episode pages except IMDb's, and IMDb was the only site to have publicity pages. Finally, Wikia was the only to include typical print media metadata fields.

Vendor advertisements on the sites were common enough to suggest that the targeted users are technology savvy males in their twenties. A variety of corporate partnerships were also possibly in evidence. Notably, Wikipedia differed from both of these trends by including no advertisements to outside vendors. Therefore, it is difficult to know whether that site targets the same user group as likely do the other three.

### 4.3.7 Advertisements as agenda-loaded

**RA7:** What can be inferred about the vendor affiliations and target audiences of the fansites from their advertisements and user profiles?

This question was studied in two ways. First, the advertisements of different types of vendors (e.g., IT, finance, retail, etc.) were among the variables content analytically coded-for in the random all-pages sample. Examples of these variables can be found in appendix A, and the overall presence of these variables on the four sites was discussed in §4.3.6. Both descriptive statistics and principal component analyses of these variables will be presented for each site.

Second, sorted Infobox fields as well as affiliation/attribute badges (known as "user-boxes" on Wikipedia), which resembled advertisements on user profiles, from 11,713 user

profile pages will be interpreted. Only the wiki sites provided user profiles for public study. GateWorld did not allow users to edit its content, so no user profiles were available in connection with the main site. Though GateWorld's forum did have user profiles, these were not studied because there was no connection between these profiles and the main site's content, and extensive screen scraping would have been necessary to obtain them. IMDb did allow users to edit content, but did not expose user profile information to the general public, making them unavailable for this study.

However, both wikis allowed users to edit all page content, exposed user profiles to the public and did not restrict their use, and provided an API for obtaining user profiles in a structured format (i.e., XML and MediaWiki). Hence, on 6 April 2010, the APIs of Wikia and Wikipedia were used to download all available profiles for users who had edited Stargate-related pages. To obtain this sample, first, the revision lists for all Stargate-related pages were requested from both APIs. Both APIs limited numbers of returned results to 500 for each page, so the most recent 500 were requested for each page. The usernames from each page's revision list were extracted, sorted, and had duplicates and IP addresses (i.e., users without profiles) removed using standard POSIX utilities. Then, the profile pages of each unique user from the resulting list were requested from the APIs. From Wikia, 989 profile pages were available; from Wikipedia, 10,724. This indicates that, though Wikipedia had many fewer Stargate-related pages than Wikia, Wikipedia's pages had many more editors. The Infobox fields and affiliation/attribute badges from these profiles were extracted and sorted using standard POSIX utilities.

### Vendor advertisements

#### *Descriptive statistics*

Table 4.28 contains counts and percentages of vendor variables on each site. Percentages are with respect to the total number of pages on each site, not the total number of ads. Multiple ads could occur on one page.

Table 4.28: Frequencies: Vendor variables, per-site

| Variable | GateWorld | IMDb | Wikia |
|---|---|---|---|
| amazon | 58 (17%) | 168 (53%) | – |
| itunes | 40 (12%) | – | – |
| IT | 164 (47%) | 219 (68%) | 352 (100%) |
| games | 44 (13%) | 9 (3%) | 36 (10%) |
| scifi | 84 (24%) | 66 (21%) | 345 (98%) |
| auto | 17 (5%) | 13 (4%) | 61 (17%) |
| pet | 1 (0%) | – | – |
| travel | 8 (2%) | 32 (10%) | 10 (3%) |
| finance | 29 (8%) | 119 (37%) | – |
| health | 12 (3%) | 297 (93%) | 52 (15%) |
| fam | 2 (1%) | 9 (3%) | 4 (1%) |
| sweepstakes | 7 (2%) | 81 (25%) | 1 (0%) |

Continued on Next Page...

Table 4.28 – Continued

| Variable | GateWorld | IMDb | Wikia |
|---|---|---|---|
| acad | 9 (3%) | 33 (10%) | 7 (2%) |
| security | 6 (2%) | – | – |
| fashion | 40 (12%) | 49 (15%) | 22 (6%) |
| massmedia | 18 (5%) | 232 (73%) | 91 (26%) |
| self | – | – | 353 (100%) |
| rental | – | 65 (20%) | – |
| furnishings | – | – | 126 (36%) |
| charity | – | 24 (8%) | 14 (4%) |
| total pages | 346 | 320 | 353 |
| total ads | 539 | 1,416 | 1,474 |

Overall, from this table, one can see that IMDb and Wikia had many more advertisements than did GateWorld on a similar number of pages. This might be due to IMDb and Wikia being explicitly for-profit companies, and GateWorld being small and largely staffed by volunteers. Whereas GateWorld had relatively low volumes in most of its pages, only rising to a moderate level of IT advertisements, IMDb had many ad categories with mid-to-high counts, and Wikia had several high counts, but was otherwise low-to-medium in most

categories. Hence, it appears that IMDb tried to maintain a more diverse portfolio of ads, whereas GateWorld and Wikia focused on a few key topics.

GateWorld's ad efforts focused (from most-to-least) on IT, science fiction, Amazon, games, iTunes, and fashion, and avoided self-advertisement, media rentals, home furnishings and retail, and charities. IT is central to most science fiction premises, and is possibly more commodifiable and/or reliably lucrative than is media sales. GateWorld had a clear partnership with AT&T. Amazon and iTunes offer media sales, and GateWorld, being small and independent, would have no qualms with including links to both companies, though they are competitors. Games are likely also a core group of GateWorld's demographic, though so many fashion ads was somewhat surprising. Self-advertisements would probably not be appealing to the site, because it does not sell anything directly. Media rentals are also perhaps less appealing to IT-savvy young people than digital downloads, and general general retail and charities would be peripheral.

IMDb's ads focused on health and food, non-science fiction mass media franchises, IT, Amazon (their parent company), and finance, and avoided iTunes (Amazon's competitor), security, non-Amazon self-advertising, security, and retail. Having a scope larger than the Stargate franchise, IMDb's ads clearly addressed a broader audience. However, perhaps some level of content-based targeting steered them towards IT and away from retail. Besides with their parent company, they had several notable partnerships with Assurant Health, Verizon, and Blockbuster.

Wikia focused on advertising other wiki sites in their farm, IT, science fiction, and retail, and avoided Amazon and iTunes, finance, security, and rentals. Wikia also had a clear partnership with Verizon, and seemed to advertise those other Wikia wikis that were

related either to computers or science fiction and fantasy. Besides promoting their own wikis, more attention was clearly given to advertising tangible, sell-able commodities than media.

***Modeling results: Principal components analysis***

The three sites also each possessed between 6 and 10 principal components worth interpreting. These analyses were conducted on each site's content-analytic binary incidence matrix.

GateWorld's sorted loadings are provided in the following list, and interpreted thereafter.

1. amazon itunes scifi acad massmedia sweepstakes gaming IT fashion health pet financial fam auto security travel

2. gaming health massmedia security fashion travel acad fam auto sweepstakes itunes pet financial amazon scifi IT

3. health security acad IT amazon scifi financial sweepstakes massmedia pet auto itunes gaming fam travel fashion

4. financial auto gaming sweepstakes massmedia pet scifi amazon itunes security IT travel health acad fam fashion

5. gaming massmedia IT security amazon itunes fashion health scifi pet fam travel sweepstakes acad financial auto

6. financial fam fashion scifi sweepstakes security travel health acad gaming amazon IT pet itunes massmedia auto

7. sweepstakes scifi travel pet security massmedia IT amazon health gaming fam fashion itunes auto acad financial

8. fam auto security sweepstakes fashion gaming financial scifi amazon IT health itunes pet acad massmedia travel

9. pet acad massmedia fam health IT itunes sweepstakes scifi amazon fashion gaming financial auto security travel

10. massmedia fam health financial travel IT sweepstakes scifi amazon itunes fashion auto acad pet security gaming

As often happens, PC1 contrasts the most frequent variables – in GateWorld's case, Amazon and iTunes, which are advertised on most episode pages – with variables that occur rarely. PC2, in which only the IT variable loads highly in the negative direction, points out that IT ads are rarely combined with other types of ads on the same page. PC3 contrasts safety related variables (health and security) with more frivolous and fun ad types (fashion, travel, family/romance, etc.). PC4 appears to contrast achievement, savings, or status-related variables (finance, auto, gaming) with variables pertaining to one's environment or personal being (fashion, family, health). Returning to the theme of fun, PC5 compares cheap forms of entertainment (gaming, mass media, IT) with expensive forms (auto, finance, academia, sweepstakes, travel). In this context, the finance variable often took the form of ads by online investment brokers, putting it closer to the realm of gambling and

sweepstakes. Similarly, PC6 contrast good vs. bad investments (finance, which could gain money, with automobiles, which generally always decay in value). Relatedly, PC7 compared risky (sweepstakes) variables with more predictable ones (finance, academia, auto). PCs 8 and 9 seem more related to the idea of responsibility. PC8 contrasts two high-loading variables, namely: the variable perhaps closest to fixity and responsibility (family), and the one close to transience (travel). PC9 similarly compares on high-loading variable that relates to personal responsibility (pet) with several related more to impersonal transience (travel, security, auto). Finally, PC10 contrasts two high-loading variables: the mass media industry with the gaming industry.

IMDb's sorted loadings are presented in the following list.

1. amazon rental charity massmedia game health acad fashion fam auto scifi finance travel sweepstakes IT

2. health finance rental acad charity scifi IT massmedia fam sweepstakes travel amazon fashion game auto

3. finance massmedia auto amazon game health IT charity travel sweepstakes rental fam fashion acad scifi

4. travel sweepstakes health charity auto game rental IT amazon finance massmedia acad scifi fam fashion

5. massmedia travel amazon acad sweepstakes fashion fam health IT auto finance rental charity game scifi

6. acad fam charity game auto IT finance rental travel amazon sweepstakes massmedia scifi health fashion

7. finance fashion travel rental acad charity sweepstakes scifi amazon massmedia IT auto game fam health

8. acad IT massmedia health scifi auto game amazon fashion sweepstakes rental charity travel finance fam

IMDb's PC1 contrasted common topics, and in this case the site's affiliates and focus (Amazon, rentals, charities, the mass media), with rarer or perhaps more opportunistic partnerships (IT, sweepstakes, travel, finance). IT may be more opportunistic because the large IT vendors (e.g., HP, Dell, AT&T, and Verizon) seemed willing to place their ads on as many sites as possible, without much allegiance. As with PC6 on GateWorld, PC2 appears to contrast potentially good or small vs. more expensive and frivolous investments (health, finance, rentals, and academia vs. autos, games, and fashion). PC3 similarly follows the investment trend, contrasting the high-loading variable finance, possibly indicating savings, with those more about spending (sci-fi, academia, fashion, etc.). PC4 contrasts themes of risk and donations (travel, sweepstakes, health, charity) with one's personal being or environment (fashion, family, sci-fi, academia). PC5, like PC10 on GateWorld, compares the general mass media with science fiction and games. The sci-fi/game combination will also occur on other sites. PC6 compares the high-loading academia variable, which perhaps indicates more mental or deep activities, with several more superficial or body related variables (fashion, health, sci-fi). PC7 returns to themes of transience and fixity also seen on GateWorld, with more transient or liquidity related variables (finance, fashion, travel,

rentals) contrasted against high-loading fixity or responsibility related variables (health, family). Finally, PC8 compares something akin to individualism vs. collectivism, contrasting many variables related to personal achievement, enjoyment, and health (acad, IT, massmedia, health, scifi, auto) against two high-loading variables pertaining to more social or group achievement (family and finance).

Wikia's sorted loadings are presented in the following list.

1. scifi furnishings charity acad fashion sweepstakes fam auto travel games health massmedia

2. acad fam fashion games massmedia scifi travel health charity sweepstakes auto furnishings

3. games furnishings massmedia sweepstakes charity acad scifi fam fashion auto health travel

4. scifi games health travel sweepstakes furnishings fashion massmedia fam acad charity auto

5. charity scifi health acad games massmedia fam travel furnishings auto fashion sweepstakes

6. sweepstakes health fam charity massmedia acad scifi furnishings travel auto fashion games

Wikia's PC1 also compared common topics with rare ones. In this case, only two topics were most common: scifi and retail/furnishings. This emphasis on retail echoes the

earlier descriptive findings. This is in contrast to the high-loading massmedia variable, which was indeed rare on Wikia. PC2, like the other sites, seems to compare good vs. bad investments, with academia and family considered good investments, and retail and autos being poor. PC3 contrasts more domiciliary topics (non-sport games, retail, massmedia) with more alfresco topics (travel, health, auto, fashion). PC4 compares ads for the sci-fi and game industries with those for the automotive and charity industries. PC5 contrasts the by-now-familiar themes of healthfulness or surety (charity, sci-fi, health) and risk-taking (sweepstakes). Finally, PC6 compares risks that are frivolous (games, fashion, autos) with those that could have large returns (sweepstakes, health, family).

In conclusion, a number of common themes emerged within the sites' vendor variables. All of the sites had some most common theme. For GateWorld, this was Amazon and iTunes downloads, as well as IT ads that dominated pages; IMDb focused on more stable vs. more opportunistic business partnerships; and Wikia on sci-fi and retail. Otherwise, a fairly standard set of vendor categories/themes emerged – most, understandably, related to money – with each site having its own variations on that set. The standard set included the following themes: achievement, savings, or status; spending; one's personal environment; good vs. bad investments; risky vs. predictable behaviors or investments; fixity vs. transience; and the mass media vs. the science fiction and gaming industries, the latter two often being grouped together. Additionally, GateWorld included categories for cheap vs. expensive entertainment. IMDb's ads included mind vs. body and individualism vs. collectivism dimensions. And, Wikia contained indoor vs. outdoor, sci-fi/games vs. automotive and charity, and large payoff vs. frivolity dimensions.

### User profile results

Although one can infer quite a bit about authorial intent and the target audience(s) from the patterns in advertisements (outlined in §4.3.6), user profiles on the wiki sites may further elucidate these issues, by exploring the audience members' self-representations. In this sub-section, patterns within the infobox fields and userbox badges on user profiles in the samples discussed in the introduction to this section will be presented. The exact fields and their contents vary enough, and the datasets are small enough, that interpretation of the sorted fields would be most efficient and effective. Obviously, these results represent only that small number of users who chose to create portfolios and to report personal information, and are only accurate to the degree that they told the truth, though there is little reason to believe that people would be deceitful, excepting a small number of obviously impossible values (e.g., height=70 meters), which were probably intended as jokes.

#### *Wikia*

Unlike Wikipedia, infobox fields were used considerably more often than badges by Wikia's users to indicate their preferences and affiliations. The only badges occasionally used by Wikia's users were those indicating which Stargate series the user preferred. Infoboxes, by contrast, contained many of the following fields about users: birth date, birth place, eye color, gender, hair color, height, hobbies, occupation, and town or country of residence.

Beginning with birth and residence information, 40 users reported a *birth date*, having a mean of 1988 (i.e., 22 years old), a median of 1990, a min of 1968 and a max of 1995. Birth dates' standard deviation was 4.83, with a decided right skew (-2.07), and fairly high

kurtosis (6.20). This suggests that users may have been in their low twenties and upper teens. Twenty eight *birth places* were reported. Ten were in the UK (two from London; two from Scotland; two from Wales; one each from Liverpool, Northampton, and Oxford; and one only specified England). Five were in the USA (Georgia, Missouri, New Jersey, West Virginia, and Wisconsin). Two were in Germany (Darmstadt and Seligenstadt). Two were in the Netherlands (Heythuysen and Roosendaal). Two were in Sweden (Sundsvall and Trollhättan). Finally, one was in each of the following: Argentina (Buenos Aires), Finland (no city specified), France (Rouen), Malaysia (Johore), New Zealand (no city specified), Romania (Bucharest), and Russia (Novosibirsk). *Places of residence* followed much the same distribution, with the losses of France, Germany, Malaysia, New Zealand, and Russia. Additions to the previous list included: Austria (city unspecified); Gloucester, UK; Meerssen, Netherlands; Michigan, USA; and Singapore (city unspecified). Hence, these users of Wikia's Stargate pages claimed to be primarily from northern Europe and the middle-to-eastern USA, with a few outliers in eastern Europe, South America, and Southeast Asia.

Twenty six users specified an *occupation*. Sixteen identified themselves as students, one as a college student, and one had "just finished Uni." Three identified as unemployed. Two were computer programmers. And one of each were the following: an "analyst," a "technician," and a "sous chef."

Regarding *hobbies*, 19 users specified one or more. Twelve stated the obvious, that Stargate or general science fiction were their hobbies. Four identified sports or other physical activities (karate, badminton, fishing, pool/billiards, and jogging). Three identified reading as one of their hobbies. Two named specific media franchises, namely: LOST, Star Wars,

and House M.D. Finally, two said "computer/video games," and two identified the Internet and "doing computer stuff." Hence, in addition to computer hobbies, watching science fiction, and reading, Wikia users may have engaged in several common outdoor and indoor physical past-times.

Finally, users reported a number of physical features. Regarding *gender*, 53 users identified as male, and only two as female. *Height* values ranged from 1.64 - 1.91 meters (5.38 - 6.27 feet) and 5 feet 7 inches to 6 feet 7 inches (1.7 - 2 meters), which are consistent with average male heights in the Western and Asian regions. *Eye colors* were, from most-to-least common: brown, blue, hazel, green, and dark brown. *Hair colors* were: brown, black, blond, dark brown, and "reddish blonde." Of these, particularly the gender figures are consistent with previous inferences.

***Wikipedia***

Wikipedia users provided considerable numbers of both userbox badges and infoboxes in their profiles. Indeed, so many userbox badge types existed (i.e., 10,786 different types) that it would be impossible to name them all. Therefore, this analysis will describe the top 10 most frequently occurring badges, listed in table 4.29.

The first of these badges indicates that the user speaks English. The second, that the user is male. The third outputs the Google logo, alongside the text "This user uses Google as a primary search engine." The fourth outputs a picture of a police officer, alongside the text "This user is a recent changes patroller." The fifth refers to a "user who makes useful incremental edits without clamouring for attention" (Wikipedia, 2010b). The sixth outputs the ubiquitous no-smoking sign, and indicates that "This user does not smoke."

Table 4.29: Top 10 userbox badges: Wikipedia

| Count | Userbox name |
|---|---|
| 280 | en |
| 241 | UBX/male |
| 134 | Menasim/Userboxes/User Google |
| 107 | wikipedia/RC Patrol |
| 98 | wikipedia/WikiGnome |
| 90 | Ginkgo100/Userboxes/User non-smoker |
| 72 | wikipedia/Administrator |
| 69 | Feureau/UserBox/SimpsonsExcellent |
| 69 | Feureau/UserBox/FuturamaGoodNews |
| 68 | Menasim/Userboxes/User Writing |

The seventh indicates that the user is a Wikipedia Administrator. The eighth and ninth indicate that the user is a fan of The Simpsons and Futurama TV shows, produced by Matt Groening. Finally, the tenth prints a picture of a quill and ink, alongside the text "This user enjoys writing." In addition to confirming that some editors of Wikipedia's Stargate pages speak English and are male, they also provide several interesting cultural associations, namely: the use of Google instead of other large Web search engines, the involvement in either administrating or policing Wikipedia, being non-smokers, and enjoying Groening's work. Judging from DVD extras and behind-the-scenes footage viewed by the researcher, many of these cultural associations appear to also be true of many members of the Stargate

cast and crew.

Regarding infoboxes, and again beginning with birth and residence information, 80 users provided *birth date* information. These Wikipedia's users were older and more varied in age than Wikia's, with mean at 1983 (27 years old), median at 1985, min at 1953, and max at 1995. With a standard deviation of 8.84, skewness of -1.29, and kurtosis of 1.5, most of these Stargate-related Wikipedia users were in their mid-twenties, the peak was not particularly high, and span of ages was quite wide.

Sixty four users identified a *birth place*, and 83 provided a *place of residence*. Forty one were born in the USA (primarily California, Ohio, Connecticut, Illinois, Louisiana, Massachusetts, Nebraska, New York, and Texas). Six were born in the UK (one each in Newcastle, North Yorkshire, "North East," Scotland, Wiltshire, and West Midlands). Six were born in Canada (three in Alberta, and one each in Manitoba, Ontario, and Saskatchewan). Three were born in Australia (two did not specify where, and one in South Australia). Two were born in the Netherlands (Dongen and Schagen). The rest were individuals, with each being born in a different country, namely: Bangladesh, Brazil, China, Tanzania, New Zealand, and Norway. Regarding places of residence, 54 lived in the USA (11 did not specify where; four each in New York and Texas; three each in California, Maryland, Ohio, Virginia, and Washington; two each in Indiana, Kentucky, and Louisiana; and the rest in a variety of states). Otherwise, nine lived in Canada (four in Ontario, three did not specify, one in Newfoundland and Labrador, and one in Quebec). Six were in the UK (two in Scotland, and one in each of Greater Manchester, North East England, Southwest England, and one unspecified). Three were in Australia (South Australia, New South Wales, and one unspecified). Three were in the Netherlands (two in South Holland and one in Noord-

Brabant). Two were in the Philippines' Metro Manila area. The rest were individuals each living in a different one of the following countries: Mexico, New Zealand, Norway, Russia, Tanzania, and the United Arab Emirates. These figures suggest that Wikipedia's Stargate editors may have been more often located in the USA, Canada, and Australia than they were on Wikia. There is also a broader range of the USA represented. Otherwise, northern European countries were still well in evidence, as was Southeast Asia and Oceania, though eastern Europe and South America were less so, and several Middle Eastern and South Asian countries were added.

Wikipedia provided information on users' educational degrees as well as occupations. *Educationally*, two editors had business degrees, and the rest had one of the following: computer science, journalism, law, mechanical engineering, philosophy, physics, political science, and Russian. Educational institutions were spread quite thinly across the aforementioned geographical regions. *Occupationally*, 16 users were students at some level, seven held academic positions, six were in medicine, five were in IT, five were programmers, four were fiction authors, three were in entertainment, three were in the food industry, two were consultants, two were lawyers, two were musicians, two were photographers, two were in retail, two were unemployed, and the rest were individuals in one of the following fields: art, civil service, homemaking, journalism, politics, sales, or security. This shows that, though were many students, as on Wikia, these Wikipedians were also made up of working professionals in a variety of fields.

In addition to hobbies and interests, Wikipedia profiles also included information on users' political and religious affiliations, and relationship statuses. *Politically*, five users called themselves Democrats, four Independents, four Libertarians, three Conservatives,

three Liberals, three Republicans, two Moderates, one Anarchist, and one Green Party supporter. This shows that these users may have been fairly evenly split across many political philosophies. *Religiously*, 10 users identified as Atheists, eight Agnostics, five Roman Catholics, three unspecified Christians, two Methodist Christians, two Humanists, and the rest one of the following: Bahá'í, Eastern Orthodox Christian, Episcopal Christian, Evangelical Christian, Lutheran Christian, otherwise Protestant Christian, Seventh Day Adventist Christian, Confucian, Jewish, Sunni Muslim, Taoist, or Unitarian Universalist. This suggests that most of these users either eschewed religious affiliations or ascribed to one of the prominent branches of Christianity in the USA.

Regarding *hobbies and interests*, besides the large number of people interested in science fiction and Wikipedia, three identified interests in mass media franchises, three in music, and two in religions. The rest showed interest in one of the following: animal/human rights, comics, computers, engineering/science, the military, politics, railroads, sports (volleyball and running), or Web design. As with the userbox badges, many of these topics are prominent in the Stargate franchise's content.

Regarding *relationship statuses*, 17 users identified as single, three married, two "cohabitating," two in a relationship, two engaged, and one in a same-sex marriage. This suggests the common trope among nerdy online communities (e.g., Slashdot), that nerds are usually single young men. Though the age results presented earlier show that the range was wider than just young men, the Infobox *gender* fields confirm that most of these users identified as male. Fifty three users reported being male, and only five reported being female. Six users reported having heterosexual *sexual preferences*, and two reported being homosexual. Most users reported having brown or blue *eyes*, had brown or black *hair*,

were 5 feet 9 inches to 6 feet 5 inches in *height*, and 148-265 pounds in *weight*. Three users also reported their Myers-Briggs *personality types*, namely: ISTJ, INTJ, and ISFJ. These figures are also consistent with a common stereotype about members of the western male population who are most likely to be drawn to science fiction and to spend time editing articles online, namely that they are introverted, judgment-oriented, heterosexual, and over-weight.

### *Conclusion*

The findings from these user profiles enriched speculations about the sites' users made in other analyses in this chapter. In addition to the stereotype of nerdy young men, Wikians more often identified themselves as British or otherwise European than American, and mentioned a range of physical past-times in addition to science fiction. However, evidence also supported the stereotype that they are young, and the finding from §4.3.1 that most Wikians were students.

Stargate-interested Wikipedians, by contrast, more often identified themselves as being located in the USA or Canada, as being older on average (mid-to-upper twenties), and as spanning a broad range of ages. They also claimed to have been educated in a variety of fields, and often to be professionally employed. The identified hobbies and interests were more often aligned with those appearing in the Stargate shows, and there were often indicators of their being active participants in the Wikipedia community. Politically, these users claimed to fairly evenly span the range of philosophies, though they most often identified with being agnostic, atheistic, or affiliated with one of the few most common Christian denominations in the USA. Finally, the personal characteristics identified by most

of these users aligned with the stereotype of being single, male, heterosexual, introverted-judgmental, and over-weight. This stereotype is also well-personified in the Eli Wallace character who is central to the currently airing Stargate Universe series.

### 4.3.8 Recommendations as ratings: amateurs vs. professionals

**RA8:** How do fan ratings compare with editor and Nielsen ratings on GateWorld, as well as across the sites on similar topics?

First, though this question asks for analysis of the two Nielsen ratings variables present on GateWorld, this was not done, because the copyright legality of GateWorld's publicly republishing these data could not be satisfactorily established. The page on GateWorld's website that explains the site's use of Nielsen ratings (GateWorld, 2010b), as well as those episode pages on the site that list Nielsen ratings, did not display a Nielsen copyright, which that company usually requires of its licensees (cf. IMDb, 2009g). When pressed on the issue, GateWorld's owner, Darren Sumner, said the following: "The info from past seasons has come from a variety of sources.... These include network press releases, trade publications, *Stargate* [original asterisks] production personnel, SciFi Wire (the network's Web site, which until the end of 2008 published weekly SCI FI, Channel and syndication Top 10 lists), and insiders with access to Nielsen Media reports who communicated with GateWorld directly" (personal communication, 28 April 2010). Without access to the documents or testimonies referenced by Mr. Sumner, or to a copyright attorney who would be able to establish that republishing those documents' contents on the Web is legal, this study was reluctant to use the Nielsen data contained in GateWorld.

Additionally, an Account Executive at Nielsen, Carly Litzenberger, was contacted regarding a license to legally obtain and analyze these two variables. In addition to ensuring legality, this would have offered the benefit of there being no missing data in the dataset, which was not true for GateWorld's piecemeal Nielsen data. However, the choice was made not to license these data from Nielsen for the following reasons. First, the company wanted $1,050 for the mere 311 data points on two variables. Second, because these two variables contain Nielsen's estimates of the viewing size of episodes when first broadcast on both cable and in syndication, and since examination of publicly available Nielsen data (e.g., in USA Today) reveals that viewership is usually higher for premier and finale episodes than normal episodes, the role that these variables would play in an analysis of fan ratings could be somewhat replaceable by binary variables, based on publicly available episode broadcast dates (e.g., from Wikipedia), indicating whether an episode was a premier, finale, or normal episode. This was done below.

Third, the quality of Nielsen's data is questionable. The data originate from a supposedly "representative" sample of households which own televisions (TheNielsenCompany, 2010), though they do not explain how the sample's representativeness is ensured. The company also can only survey, even by automated means, approximately 25,000 households per day (TheNielsenCompany, 2010), despite the SCI FI Channel (the TV network airing SG-1 seasons 6-10) reaching an estimated 88 million households between 2005-2006 (GateWorld, 2010b). Also, Nielsen's cable ratings only reflect households both that subscribe to the TV network airing Stargate and that had their TVs turned on when Stargate was airing, not households that were necessarily watching Stargate (GateWorld, 2010b). Finally, the error rates / confidence intervals for Nielsen's measurements are unknown.

This question implies several sub-questions and suggests several analyses. First, the user and editor ratings variables on GateWorld, which only occurred on episode pages, require description as well as the finding of the correlation between them. Second, Wikia was the only other site to provide/allow ratings on pages, these were only user ratings, and they existed for page types other than episodes. Therefore, the user ratings for each page type need description. Correlational analyses between page types would not be possible, because each page type includes different pages/observations. Finally, the GateWorld episode rating variables' observations need to be matched with the episode rating data available from Wikia, and correlations and latent variables sought between both sites' variables.

An exhaustive sample of ratings data was obtained from GateWorld. The data were obtained via screen scraping, and were also manually checked against the website's pages by the researcher. Wikia's user rating data were obtained by querying the API for all votes for the entire Stargate wiki. No upper limit to the number of allowed results was specified in the API documentation, so large values of the maximum search result setting (wklimit) were tried until no additional votes were returned by increasing that setting. This approach resulted in 792 votes being obtained, which was probably an exhaustive sample. The numbers and types of pages with ratings were as follows: 146 episodes, 142 characters, 116 places, 108 technologies, 89 peoples, 77 miscellaneous, 49 ships, 17 actors, 3 books, 3 games, and 2 videos. The Wikia votes were obtained in XML format along with their page titles, which was parsed using standard POSIX utilities.

This section's interpretation of fan ratings as indicating preference about an episode (rather than about the webpage itself) was most straightforward for Gateworld, where editors made obvious that they were rating the episode and not the page, and where each user

poll asked "How would you rate [episode name]?" Though Wikia's rating function asked users to "Rate this article:," examination of especially low-rated pages revealed nothing abnormal about them; indeed some were as lengthy and detailed as highly rated episode pages. Therefore, it was assumed that Wikia users were rating the episode rather than the article.

Finally, as with the study of user profiles and targetted marketing in §4.3.7, though this section's findings do not directly describe webpage IQ, they aid in characterizing the pages' usual authors, from which one can infer which types of pages such authors would be most likely to produce and consume.

**GateWorld**

Descriptive statistics for the GateWorld ratings variables are presented in table 4.30.

Editor ratings were given in terms of 1-4 stars/integers, and could include half/0.5 star values. Fan ratings were the averages of fans' rating the episode with integers from 1-10. Though the number of fans who voted on each episode (n) was available by parsing a separate "view results" page, this was not done, because even unpopular episodes received 1,000 or more votes (i.e., large samples), and because it would have involved additional data collection and parsing. The site also appeared only to track unique IP addresses, so the same users could have voted multiple times either by using different computers not located behind a common router or by requesting a new IP address from their Internet service provider and re-voting. The first of these, because it is easier, is perhaps more likely (e.g., if someone voted from home and work).

From this table, one can see that there were a number of missing/NA values for each

Table 4.30: Descriptive statistics: GateWorld episode ratings ($n$ = 316)

| | **editors** | **users** |
|---|---|---|
| %NA | 0.11 | 0.11 |
| sdev | 0.69 | 0.98 |
| min | 1 | 4.53 |
| 25% | 2 | 7.47 |
| median | 2.5 | 8.14 |
| mean | 2.7 | 8.02 |
| 75% | 3 | 8.81 |
| max | 4 | 9.83 |
| skew | 0.16 | -0.73 |
| kurt | -0.59 | 0.33 |

variable, because ratings were not listed on every episode page, though editor and user ratings were always either present or absent together. Editors' ratings had a nearly symmetric and quite flat curve around a median of 2.5 stars, with a slight left skew. This shows that editors did not center on only a few scores, though did have a slightly negative bias. Users' ratings were considerably more biased towards high scores, with average values around 8 of 10. A slight peak existed, with a slight right skew.

The pairwise correlation between editor and user ratings was $r$ = 0.51 ($n$ = 316). This shows that editors and users often, though not always, rated episodes similarly.

### Wikia

Descriptive statistics for the Wikia user ratings of different page types are presented in table 4.31.

Table 4.31: Descriptive statistics: Wikia user ratings

| | **actors** | **books** | **episodes** | **games** | **chars** | **misc** | **people** | **places** | **ships** | **tech** | **videos** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| sdev | 1.79 | 1.5 | 0.82 | 1.15 | 1.08 | 1.18 | 1.18 | 1.06 | 1.18 | 0.96 | 1.06 |
| min | 1 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 3.5 |
| 25% | 1.5 | 4.25 | 4 | 3 | 4 | 4 | 3 | 4 | 4.17 | 4 | 3.88 |
| median | 5 | 5 | 4.88 | 4 | 5 | 5 | 4.44 | 5 | 5 | 5 | 4.25 |
| mean | 3.68 | 4.25 | 4.4 | 3.33 | 4.31 | 4.22 | 3.99 | 4.33 | 4.38 | 4.36 | 4.25 |
| 75% | 5 | 5 | 5 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 4.63 |
| max | 5 | 5 | 5 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| skew | -0.79 | -2 | -1.58 | -1.73 | -1.78 | -1.35 | -1.01 | -1.74 | -1.99 | -1.64 | N.E.O. |
| kurt | -1.38 | 4 | 2.71 | N.E.O. | 2.52 | 0.42 | 0.01 | 2.27 | 2.92 | 2.2 | N.E.O. |
| n | 17 | 4 | 146 | 3 | 142 | 77 | 89 | 116 | 49 | 108 | 2 |

Perhaps most noteworthy about these figures is their consistency across the different page types. Despite varying numbers of observations, Wikia users consistently gave most pages a 4 out of 5 rating. Standard deviations were all around 1-1.5, skews were all slightly or moderately negative/right-skewed, and most kurtosis was between 2-3. These results are generally consistent with GateWorld's users giving higher-than-moderate scores, suggesting that average fans, since they are fanatical/enamored, may have an overall positive bias. However, this was less pronounced when users rated actors and peoples. Kurtosis and skew figures were less pronounced, indicating more willingness by fans to give both higher and lower scores to certain people(s) than to other aspects of the franchise. Also, recall that the abbreviation "N.E.O." is used for cases when Not Enough Observations were available to

calculate a statistic. Though the samples were exhaustive, one should always be wary of findings with particularly small sample sizes.

### GateWorld and Wikia

On only the episode pages, because GateWorld only provided ratings for those pages, after manually matching Wikia episode page names with GateWorld pages, pairwise correlational and principal component analyses were done.

Pairwise correlations between the GateWorld variables and the Wikia variable were as follows, omitting missing values as before: GW.editor::Wikia $r = -0.1$ ($n = 124$) and GW.user::Wikia $r = 0.07$ ($n = 123$). These figures show that Wikia users' ratings were nearly unrelated to GateWorld users' ratings, and that Wikia users' ratings were mildly dis-associated with GateWorld editors' ratings. This suggests that GateWorld editors, GateWorld users, and Wikia users belong to three different rating groups, and that the Wikia user group behaves somewhat anti-thetically to GateWorld's editors.

In order to further characterize these different groups, each episode was dummy coded for whether it was a season premier, season finale, a mid-season finale (i.e., when airing, each season went on hiatus at least once for several weeks or months in the middle), mid-season premier, or normal episode. These binary incidence variables were combined into a matrix with the rating variables, which was subjected to principal component analysis, in order to find latent dimensions among them.

Six dimensions were worth interpreting. In PC1, GateWorld editors, and to a lesser degree GateWorld users, gave the highest ratings to season and mid-season premiers, as well as to season and mid-season finales to a lesser degree. This suggests that GateWorld's

editors, and to a lesser degree regular users, were most often '*high anticipation viewers*,' namely viewers that rated highest whatever the studio released as a premier. More effort may have been invested into these well-marketed episodes by the studio and/or GateWorld fans' anticipation could have biased their opinions. This also suggests that GateWorld's editors were the most either influenced or affiliated with the studio's agenda.

In PC2, GateWorld users rated normal episodes more highly than did GateWorld editors, contrasted with Wikia users rating mid-season finales and premiers highly. The first of these findings suggests that regular GateWorld users often came in search of non-premier/finale episodes, possibly seeking a more *connoisseuring* experience in more obscure episodes. Wikia users preferring mid-season finales and premiers could indicate a syndication-related pattern. TV networks often air Stargate re-runs one season at a time (GateWorld, 2010c). Hence, syndication viewers – who are also probably more casual viewers than either high anticipation or connoisseur viewers – may be less likely to keep track of season-level premiers/finales, because the plots of those episodes typically span seasons. TV networks may also schedule mid-season finales/premiers at times when more syndication viewers are likely watch them. Therefore, most Wikia users may be casual *syndication viewers*.

Finally, PC3 and PC5 showed that some GateWorld and Wikia users rated season and mid-season finales highly, but not season or mid-season premiers. Similarly, PC6 contrasted some Wikia users' rating season premiers highly against GateWorld users and editors preferring mid-season finales and, to a lesser degree, premiers. PC4 merely contrasted the mid-season finale and premier variables. These results suggest that small numbers of GateWorld and Wikia users may be high anticipation viewers, like GateWorld editors.

Such viewers most often preferred finales – which are typically suspenseful, ending in "To be continued" – and less often premiers, though a small number of Wikia users preferred season premiers. Hence, most of these viewers may be *thrill/suspense seekers*, with only a few wanting to know the ending.

**Conclusion**

Descriptively, whereas GateWorld editors' ratings were balanced or slightly negatively biased on average, both GateWorld and Wikia users' ratings were clearly positively biased on average. Wikia users were also more willing to give a variety of ratings to pages about actors and groups of people than to other page types.

Regarding pairwise correlations, GateWorld editors' and users' ratings were moderately positively correlated ($r = 0.51$, $n = 316$), GateWorld users' ratings were slightly positively correlated with Wikia users' ratings ($r = 0.07$, $n = 123$), and GateWorld editors' ratings were slightly negatively correlated with Wikia users' ratings ($r = -0.1$, $n = 124$). The first of these findings shows that GateWorld editors and users often, though not always, rated episodes similarly. However, Wikia users' ratings were largely dis-associated with either GateWorld group. None of these rating groups' behaviors were strongly similar.

Finally, combining the three rating variables as well as binary variables indicating whether an episode was a season or mid-season premier or finale, a principal components analysis revealed three categories of viewership. GateWorld editors were most often high anticipation viewers, namely viewers that rated highest whatever the studio released as a premier. Though quite a few GateWorld users were also high anticipation viewers, most were more connoisseuring, preferring normal episodes that were neither premiers nor fi-

nales. In addition to a few high anticipation viewers, Wikia users were most often casual syndication viewers, judging by their preference for mid-season premiers and finales as well as several known business practices of TV networks. Finally, among those few GateWorld and Wikia users who were high anticipation viewers, most preferred finales over premiers, possibly indicating that they were thrill/suspense seekers.

### 4.3.9 Recommendations as ratings: fan ratings vs. IQ factors

**RA9:** To what extent do high or low fan ratings correlate with any other quantitative IQ characteristics of the fan literature available for each episode?

Essentially a regression question, this question asks one to relate the GateWorld and Wikia rating variables, as dependent variables (DVs), to all other IQ variables available, as independent variables (IVs), for those two sites separately. As discussed in §3.4.4, the goal of these regressions is to be able to characterize the IQ features of pages that also tend to contain high ratings, not to assert that certain ratings were provided by users or editors because of the presence or absence of certain IQ features on those pages. Such an assertion may be true for fan ratings, as fans may take into account the IQ of a page, though this was not likely the case for this research context, for the reasons discussed in §4.3.8.

The following considerations were made, when choosing how to construct these regression models. The two websites (GateWorld and Wikia) had to be modeled separately, both because their pages contained different IQ features and measurements thereof, and because Wikia also contained ratings for topics other than episodes. Since every page on GateWorld that contained an editor rating also contained an average user rating (i.e., there

were no missing data between those variables), and since all of their distributions were continuous and multivariate normal enough not to seriously violate standard assumptions, canonical correlation (CCA) was employed. The binary premier and finale variables defined in the previous section (§4.3.8) were not appropriate for inclusion among the DVs, because they do not represent fan ratings. (Also, if it had been contextually appropriate to mix binary and continuous DVs, normal CCA would not have been an appropriate model choice.) Nevertheless, because the binary variables were easily available, and this is an exploratory project, they were included among the IVs of the analyses of both GateWorld and Wikia's episode pages.

Wikia's only ratings variable, however, was both ordinal and partially discrete, making both exponential re-expressions and continuous regression models inappropriate. Because the decimal values in the variable were the result of averaging what had originally been integer-precision votes from users, the decision was made to round the data back to integers, so that a polytomous, ordinal logistic regression (POLR) model would be appropriate. Also, as shown in §4.3.8, because sufficient numbers of observations existed for episode and each type of omnipediac pages on Wikia, in order to obtain as fine-grained of results as possible, the user ratings for each of those page types were modeled separately.

Regarding model-building, an iterative stepping-up approach was employed, whereby optimal models were constructed by adding one IV, re-calculating the model parameters' significance and $R^2$ values, and only keeping a variable in the model if it consistently met the standard $p < 0.05$ significance level. Only the most optimal models will be presented in the following sub-sections. Also, recall from §3.4.4 that all continuous variables in all linear models done for this dissertation were re-expressed, centered, and scaled, in order to

improve normality.

### GateWorld

#### *Canonical correlation*

The procedures in, and format of, this section are largely based on Tabachnick and Fidell (2007, pp. 567-606). The functions used were `cancor` in the standard `R stats` package, as well as the `cancor2` function by Habing (2005), which uses equations from Mardia et al. (1979) and produces output similar to the `SAS PROC CANCOR` package, used by Tabachnick and Fidell.

Multivariate normality was confirmed by plotting pairs of canonical variates, visually looking for violations of normality, linearity, and homoscedasticity (e.g., points not falling around the origin, or tight bunching of points around the origin), as well as looking for obvious outliers, following examples by Habing (2005). As with all models in this dissertation, qq-plots of each IV were exponentially transformed towards normality. Bivariate boxplots were also used for finding outliers, which were removed. Additionally, a Mahalanobis distance plot of the data matrix was examined for linear conformance with the $\chi^2$ distribution, which is characteristic of multivariate normality (Everitt, 2005, p. 10).

While most of the variables were normal enough after transformation, several of the lexical variables were nearly singular (e.g., word and character counts of the same text). These multicollinearities were assessed through examination of all pairwise correlations between IVs, and, for groups of variables with high correlations, only one was kept in the analysis. The following IVs were removed; the variable that was kept in their place

is in parentheses: author.teleplay (author.story); word average lengths (Coleman-Liau); sentence average lengths (Fry); and paragraph average lengths, word counts, and sentence counts (character counts). In this way, assumptions regarding within-set multicollinearity were met.

In the analysis, there were two DVs (the rating variables) and 53 IVs (48 IQ variables + 5 binary premier/finale variables). The first canonical correlation was 0.61 (37% overlapping variance); the second was 0.58 (33% overlapping variance). With all canonical correlations included, $\chi^2(106) = 337.01$, $p < 0.001$, and with the first canonical correlation removed, $\chi^2(52) = 324.28$, $p < 0.001$. Hence, all of the two available pairs of canonical variates accounted for the significant relationship between these two sets of variables.

Table 4.32 presents model results for these two pairs of canonical variates. Shown in the table are correlations between the variables and the canonical variates, standardized canonical variate coefficients, within-set variance accounted for by the canonical variates (percent of variance), redundancies, and canonical correlations. Also, recall from §3.4.4 that a cutoff correlation/loading of ±0.32 will be used in presenting and interpreting table 4.32.

Table 4.32: Canonical correlation results: GateWorld

| | **First Canonical Variate** | | **Second Canonical Variate** | |
|---|---|---|---|---|
| | *Correlation* | *Coefficient* | *Correlation* | *Coefficient* |
| **Ratings set** | | | | |

Continued on Next Page...

Table 4.32 – Continued

| | First Canonical Variate | | Second Canonical Variate | |
|---|---|---|---|---|
| | *Correlation* | *Coefficient* | *Correlation* | *Coefficient* |
| Editor | -0.67 | -0.01 | -0.74 | 0.07 |
| User | -0.98 | -0.05 | | |
| **Percent of variance** | 0.71 | | 0.28 | Total = 0.99 |
| **Redundancy** | 0.26 | | 0.1 | Total = 0.36 |
| **IQ set** | | | | |
| Season premier | -0.34 | -0.01 | -0.32 | 0.02 |
| Season finale | -0.37 | -0.02 | | |
| Normal episode | 0.41 | -0.01 | 0.31 | 0.01 |
| Review links | 0.33 | 0.02 | -0.53 | 0.02 |
| Question sentences | | | -0.31 | 0.01 |
| Auxiliary verbs | | | -0.31 | 0.03 |
| External links | | | 0.37 | -0.06 |
| Inlinks | -0.51 | -0.02 | -0.41 | 0.02 |
| Validation errors | | | 0.35 | -0.01 |
| **Percent of variance** | 0.03 | | 0.05 | Total = 0.08 |
| **Redundancy** | 0.01 | | 0.02 | Total = 0.03 |
| **Canonical correlation** | 0.61 | | 0.58 | |

The variables in the ratings set that were correlated with the first canonical variate were negative of editor and user ratings. Among the IQ variables, negatives of season premier, season finale, and inlinks, as well as positives of normal episode and review links correlated with the first canonical variate. As in principal components analysis, the signs of both sides of the variate pairs can flip arbitrarily and independently of each other. So, the magnitudes and canonical correlations are all that one can compare across the variate pair, not the signs. Hence, the first pair of canonical variates indicate that user, and to a lesser degree editor, ratings were moderately positively correlated (0.61) with a spectrum contrasting, on the one hand, normal episodes and review links with, on the other hand, inlinks, season finales, and season premiers. That is, many users and some editors gave high ratings either to/on pages having many inlinks and being about season premiers and finales, or pages about normal episodes and possessing a link to an editorial review. The first of these would be pages about episodes that were highly anticipated by the general public, and the latter pages about more obscure episodes, interest in which could be called connoisseuring or perhaps fanatical.

The second canonical variate showed that editor ratings were also moderately positively correlated (0.58) with a spectrum of, on the one hand, external links, HTML validation errors, and normal episodes, and, on the other hand, a review link, many inlinks, season premiers, questioning sentences, and auxiliary verbs. That is, editors usually gave high ratings either on lengthy non-premier/finale pages for which they themselves probably connoisseuringly/fanatically produced most of the content, or on pages about highly anticipated premiers, which they wished to critique.

### Wikia

This section presents the results of polytomous ordinal logistic regressions (POLR), predicting user ratings from IQ characteristics, on episode and omnipediac page types. The models were calculated using both the `lrm` function from the `R Design` package (Harrell, 2010), following examples by the UCLAAcademicTechnologyServices (2010), as well as the `polr` function from the `R MASS` package (Venables and Ripley, 2010), following examples by Quinn (2010). In general, `lrm`, which contains its own calculations, was used first, to automatically create cut-points in the continuous latent variable observed in the data. The results were confirmed against the `polr` function, which uses the `as.ordered/factor` function in the `base` package as well as the `glm` function in the `stats` package. All results matched closely, using this approach.

Also, in addition to the usual checks and transformations for multivariate normality and homoscedasticity, before any of the following analyses proceeded, the frequencies of each level of responses (i.e., how many times users rated episode pages with a '1') was checked, to ensure that none had zero or one values. No page types had zero values, though the miscellaneous and ship pages had a number of '1' value on the low scores and small values in the middle scores. Therefore, those page types will be excluded from the analysis.

Furthermore, the proportional odds (i.e., parallel slopes) assumption underlying ordinal logistic regression was tested for each IV on each level of the DV by using the `qlogis`, `as.factor`, and `glm` functions, following UCLAAcademicTechnologyServices (2010). This involved constructing a table of linear predictions from a logit model, used to model the probability that $y$ is greater than or equal to a given value for each level of $y$, using each

value of one predictor variable at a time. Essentially, this produces the predicted values one would find from regressing each level of each IV on each level of the DV. The parallel slopes assumption was confirmed by subtracting each pair of predicted values for each possible value of the same IV, to make sure that the slopes remained nearly constant. Inconsistent slopes would mean that the effect of different levels of an IV could differ across different levels of the DV. All of the IVs reported in the following sub-sections obeyed this assumption.

Finally, although directly interpreting the magnitudes of the coefficients in the following models, which represent ordered log odds of variously re-expressed variables, would be difficult, as with all other models in this dissertation, the relative magnitudes of coefficients and their signs should be sufficient for grouping similar variables together and exploratively answering the research questions.

***Episode pages***

With 135 observations, and a $\chi^2$ likelihood ratio of 42.52 on 13 degrees of freedom, the model that included all levels of $y$ was significant at $p < 0.001$, compared with a model having no predictors, and had a pseudo-$R^2$ of 0.32. This indicates that, though IQ variables capture/explain a highly significant amount of the variance in user ratings of episode pages, a considerable amount remains unexplained.

Twelve IQ variables and the season premier variable were found to be significant, using the Wald z-test, namely: whether the episode was a season premier (coefficient=2.04, $p < 0.05$), whether the episode's number (2.76, $p < 0.01$) and/or story writer (-2.93, $p < 0.01$) appeared in the page's Infobox, to-be verb count (0.67, $p < 0.05$), preposition count (-0.39,

$p < 0.05$), count of sentences beginning with interrogative pronouns (1.05, $p < 0.01$) and conjunctions (-0.84, $p < 0.05$), link count (-0.34, $p < 0.05$), count of links to PDF files (1.02, $p < 0.01$), count of production notes (0.68, $p < 0.01$), count of listed releases of the same title available in different media (-1.18, $p < 0.05$), the Coleman-Liau readability metric (0.29, $p < 0.01$), and inlink count (0.24, $p < 0.01$).

This means that fan ratings of episode pages were higher when: the episode was a season premier, fans had gone to the trouble of including the episode's number, many to-be verbs and interrogative sentence beginnings were present, many PDF files were linked-to, many production notes were given, many inlinks were present, and the page's text received a higher than average Coleman-Liau readability score, a metric which prioritizes texts having many characters but few sentences or words (i.e., having a few lengthy sentences and lengthy words). However, fan ratings were lower when the page: listed the story's author, contained many prepositions and conjunctive sentence beginnings, contained many links, and listed many alternate media.

In the previous section's second analysis of Wikia (§4.3.8), which characterized users' ratings in terms of amounts of variance accounted-for rather than extremity of ratings, users who preferred season premiers accounted for only a small fraction of all user rating behavior. Hence, those few users must have had the most positive scoring bias, rather like if GateWorld editors' and users' behaviors were combined. Though not identical, such premier pages on Wikia shared a similar IQ profile as those premier pages preferred by GateWorld editors, namely: many inlinks, lists or links to supporting materials, many interrogative sentences, and lengthy texts. This regression analysis also allows the most user-avoided episode pages to be characterized, namely those with run-on sentence constructions and

unusual story authors (i.e., not the series' main writers).

### *Omnipediac pages*

The analyses of omnipediac pages will be presented together, because of those page types' similarity of purpose. Overall, the character and people models were of similar significance, as were the place and technology models. The character and people models had $R^2$ values between 0.27-0.38, at $p < 0.001$; the place and technology models had $R^2$ values between 0.13-0.18, at $p < 0.05$. This seems a fairly intuitive division between people and things. Continuing with this dichotomy, the clearest division between scores on these pages was that place page scores were most associated with word usages, whereas technology page scores were most associated with content characteristics. Character and people page scores were a mixture of both.

Character pages received the highest scores when many textual characters, sentences beginning with pronouns, JPGs, and notes were present. Lower scores were present on character pages having many words, passive sentences, sentences beginning with subordinate conjunctions, and high Coleman-Liau scores. This suggests a somewhat opposite pattern from episode pages. Here, fans seemed to be seeking less verbose trivia notes and photos of their character of interest, rather than description.

People pages also received high scores when they contained many textual characters, though having many to-be verbs, lengthy sentences, and a high Fry readability score, which prioritizes long sentences, had more of an effect. With respect to content, the effect of JPGs was as positive for these pages as for character pages, and discussions of peoples' alliances provided an additional boost to scores. Scores were lower on pages that had many short

sentences, many name and alternate name fields, many "distinctions" (i.e., characteristic qualities) fields, and many language fields. That is, unlike character pages, lengthy discussion was rewarded and listy presentation was not.

On place pages, high scores went to pages with many sentences, especially when sentences were particularly short, and with many sentences beginning with articles. However, having many passive sentences could negate all of these positive effects. Hence, place pages were most rewarded when they offered terse information about a place, rather than extended description.

Finally, technology pages received higher scores when they offered technical information, such as about the height or controls of a vehicle, and lower scores when offered more classificatory information, such as a vehicle's class or role.

**Conclusion**

From these analyses, and the related ones in the previous section (§4.3.8), three profiles of user and editor rating behaviors emerged, and were found to be associated with various IQ factors, on GateWorld and Wikia. GateWorld's editors and a few GateWorld and Wikia users were "high-anticipation viewers," most preferring pages about season premiers, with many inlinks, links to supplementary material, and interrogative texts. However, most GateWorld users, and occasionally editors, were more "connoisseuring viewers," preferring normal episode pages with links to editorial reviews and lengthy editorial content. Most Wikia users followed a third profile, possibly indicative of "syndication viewers," where mid-season premiers and finales were preferred by most, and exciting season finales preferred by some. Most high ratings on Wikia's omnipediac character, place, and technology

pages were also given on pages containing much trivia, tersely descriptive, and technical information. Hence, casual syndication viewers may have been most interested in quick summaries of topics of interest to them. However, highly rated omnipediac people pages were the opposite, with lengthy texts that included images and discussed those peoples' alliances.

### 4.3.10 Inlinks and PageRank

**RA10:** To what extent do pages with high PageRank values or counts of inlinks correlate with other IQ characteristics?

Another question of regressing some variable against page IQ characteristics, in this case, only inlink counts were continuous, making standard multiple regression (e.g., IWLS) appropriate. Had the full range of PageRanks (0-10) existed on the sites, standard multiple regression may have been appropriate. However, they never exceeded 5. This, in combination with the ordinal and integer/discrete nature of PageRanks, makes polytomous ordinal logistic regression (POLR) again most appropriate. All procedures and functions for checking model assumptions, for transforming variables, and for computing the models were the same as in the POLR portions of §4.3.9.

#### Inlinks

Many IQ variables predicted inlink counts on these sites. To facilitate interpretation, the effects have been organized into content-, word usage-, link-, and authorship-related types. Within each type, each variable's effects/occurrences for different sites have been combined and sorted from those causing the most positive to the most negative effect on inlinks.

***Content-related***

JPGs had highly positive effects on the inlink counts of Wikia and Wikipedia crew pages. Hence, on both wiki sites, crew pages with photographic images were more cited from off-site. On GateWorld episode pages, JPGs also had a small positive effect, though JPGs in addition to a forum link decreased inlinks by more than having JPGs alone. Forum links were present on all Atlantis and SG-1 episode pages, but not Universe or Infinity. At the time of data collection, Universe was highly anticipated, but had not yet begun airing episodes, and Infinity, a children's cartoon that lasted only one season, was deemed by GateWorld's editors to contain "non-canon" plots not worthy of extensive documentation. Hence, the two active series had somewhat lower inlink counts than did the few pages dedicated to either anticipated or obscure series, suggesting that GateWorld may have been considered by the outside world as more the place to go for information about the latter types of series. Finally, JPGs had a small negative effect on Wikia actor page inlinks, except when combined with the hair color infobox field, which had a small positive effect. The JPG-haircolor combination was usually only present on the pages of the most famous actors, suggesting that the least notable of actors (i.e., with the fewest inlinks) may have been more likely to have JPG photos, except when the actor was very famous, in which case their physical features would also usually be described by fans.

Release dates only predicted inlinks on GateWorld book pages, generally having a positive effect, unless the page's text also contained many particularly short sentences, which somewhat negated the effect. Most book pages had release dates, explaining the first effect, and the few without release dates were non-fiction guide books, the pages of which

primarily listed their contents.

Author fields and long descriptions were predictive of inlinks on GateWorld comic pages. The presence of story, maincover, and color authors each decreased inlink counts, though a page listing both a color author and having a long textual summary had somewhat higher counts, though a long textual summary alone also lowered counts. Comics about the original film usually received more inlinks, did not have long descriptions, and color author rarely occurred with a long description. Comics about the television series (SG-1 and Atlantis) usually did have long descriptions, often with a color author field. Hence, although movie comic pages, which were primarily listy, usually received more inlinks, a few popular television comic pages arose to the movie inlink level and were likely to identify the author who colored them.

Pages with fields identifying who played a character and, to a lesser degree, the race who occupied a planet, had increased inlink counts on GateWorld omnipedia pages. However, pages with fields identifying the home-planet or ships of a race had fewer inlinks. Since each of these fields can only occur on one type of omnipedia page, one can infer that character and place/planet pages were more often cited than race or ship pages.

Biography texts were associated with increased inlinks on Wikipedia crew pages, but decreased on Wikia crew pages. Wikia crew pages that specified both a writer and director also had more inlinks. This suggests that, in documenting crew members, Wikia may have been more listy and Wikipedia more textual.

Of the many infobox fields unique to Wikia's omnipediac pages, fields relating to the heading council of a race were most predictive of inlinks. Namely, if their legislative branch, dates restored or re-organized, era of existence, or state religion were listed, inlinks

were higher; and, if their date established, people's currency, people's language, executive branch, and date fragmented were listed, inlinks were lower. This suggests that, of all the possible omnipediac pages on Wikia, users on other sites found pages containing fields about governing bodies to be the most worth linking to, with pages containing fields about legislation, re-organization, and religion being most interesting, and fields about executives, establishment, language, currency, and dissolution least interesting.

Wikipedia's crew pages had a number of variables predicting their inlinks. The largest boosts to inlink counts were a field giving the crew member's occupation and lists of their productions, followed by paragraphs of text describing their early life and spouse. The largest detriments to inlink counts were an image of the crew member, a link to their website, and listy information about their birth date, birth place, and nationality. This suggests that Wikipedia crew pages that many people cited were characterized by substantive information about the person's work and personal life, whereas less oft-cited pages provided more superficial information, which referred the user elsewhere.

Finally, Wikia actors pages had slightly more inlinks when an actor's nicknames were present and slightly fewer when long lists of trivia were present. Also, Wikia book pages had slightly fewer inlink counts when an ISSN was present.

***Word usage-related***

The majority of differences related to word usage were between crew pages on the two wikis and comic pages on GateWorld.

Crew pages on both wikis showed the most extreme inlink effects due to page length. Such pages with many characters had considerably fewer inlinks on both wikis, though

pages with many words had considerably more. Having many sentences was only of benefit to Wikipedia's crew pages, and was of great detriment to Wikia's. Also, having many characters was of moderate benefit to GateWorld's comic and omnipedia pages. This suggests that those who cited the wikis' crew pages did not simply prefer longer texts, but rather texts with many words and, in Wikipedia's case but not Wikia's, many sentences. By comparison, those who cited GateWorld's comic and omnipedia pages were less discerning.

Regarding sentence length, Wikia's crew pages had more inlinks when they had many short sentences and fewer inlinks when many long sentences. Wikipedia's crew pages, by contrast, had the most inlinks when their sentences either very short or long, and had fewer when their sentences were less extreme in either direction. GateWorld's comic pages similarly had more when sentences were very long, and book pages more when sentences were very short. A clear finding from this is that those who cited Wikia's crew pages wanted them to have short sentences, but those who cited Wikipedia's crew pages wanted more variety.

Questioning sentences were associated with the greatest inlink counts on Wikipedia's crew and episode pages, the latter in association with conjunctions, and the fewest on GateWorld's comic pages, though added length or passive sentences could somewhat mitigate that effect. These findings are consistent with the earlier notions that Wikipedia's crew and GateWorld's comic pages provide more complex and lengthy language. Similarly, passive sentences were negatively associated with inlinks on both wikis' crew pages, though somewhat positively on GateWorld's comic pages. Passive sentences were perhaps more indicative of a general documentation style, which produces lengthy text with typical sentences and word usage, something that both crew pages would avoid, but that the comic

page might not.

In terms of general word usage variables, to-be verbs were associated with many inlinks on both wiki sites, as were prepositions on Wikia crew pages. Wikipedia's moderately inlinked crew pages also had many conjunctions and nominalizations, and Wikia's crew pages many pronouns. Negatively, Wikipedia crew pages with many prepositions and pronouns had fewer inlinks, as did Wikia crew pages with many conjunctions. GateWorld's low-inlinked comic pages also had many nominalizations. Many auxiliary verbs were also slightly associated with low inlink counts on both wikis' crew pages. With the exception of to-be and auxiliary verbs – which were favored and neglected, respectively – these results primarily suggest that those who linked to the two wiki sites preferred different writing styles: Wikia many prepositions and pronouns, and Wikipedia many conjunctions and nominalizations.

Sentence beginnings were most associated, and usually negatively, with inlinks on GateWorld's comic pages. Besides Wikia crew pages having many inlinks when sentences began with pronouns, GateWorld comic pages had fewer inlinks when sentences began with pronouns, conjunctions, prepositions, or articles; though inlinks were slightly higher when sentences began with interrogative pronouns. This Wikia result is consistent with the previous paragraph. The GateWorld result shows that the site's inlinking users' preferred writing style centered more around sentence beginnings than middles or endings, preferring sentences beginning interrogatively.

Finally, regarding readability metrics, the Lix metric was associated with the most inlinks on Wikia crew pages, the ARI with moderate inlinks on Wikipedia episode pages, and the Fry with slightly more inlinks on GateWorld omnipedia pages. Wikipedia crew

pages had moderately fewer inlinks when Lix and PageRank scores were high, and GateWorld omnipedia pages had slightly fewer when both Fry and PageRank scores were high. The Lix and ARI metrics prioritize having many words, thus making its results correlated with those of the word count variable presented at the beginning of this sub-section. Fry prioritizes sentences and syllables, making it closer to the character length results reported above. That having a high PageRank would decrease inlink counts is an interesting finding that will be more deeply explored in the follow sub-section, as well as in §4.3.10. It suggests that the Wikipedia crew and GateWorld omnipedia pages to which many popular sites (e.g., the mass media) link may differ from those to which many sites in general (e.g., many fans' sites) link. As has been shown several times, fans can be much more connoisseuring than corporations.

***Link-related***

Considerably more variety in page types, than in the previous sub-section, were evident in the ways that non-inlink types of links were associated with inlinks.

External links, which are links going from the originating site to another site – as opposed to internal links, which go to a destination within the same site – were most associated with inlink counts on GateWorld episode pages. Having many external links was associated with moderately lower inlink counts, lower still when a page's PageRank was also high, though this was mitigated slightly when a long descriptive text was also present. This echoes the last finding from the previous sub-section, that pages that are preferred by, and in this case possibly linked back to, highly popular (i.e., high PageRanked) sites may have fewer overall inlinks. That this was mitigated by the presence of lengthy text is

also consistent with the earlier finding that fans (i.e., general inlinkers) come looking for substantive content.

However, having a high PageRank alone was usually positively associated with inlink counts. On GateWorld episode, comic, and omnipedia pages, IMDb character pages, Wikipedia list pages, and Wikia omnipediac pages, having a higher PageRank occurred alongside more inlinks. Only on Wikipedia crew pages were high PageRanks alone associated with fewer inlinks. These patterns indicate that, in most cases, having a higher PageRank probably equates with a general notion of social popularity, not just popularity among popular sites. That is, most inlinkers also gravitate towards the popular sites, in the rich-get-richer pattern (i.e., preferential attachment, cumulative advantage, etc.) commonly associated with the world's use of large, centralized search engines (Barabási and Albert, 1999). Wikipedia crew pages' antithesis to this pattern might be due to those pages having been taken over by powerful studios or producers, who point their high PageRanked sites to their own Wikipedia pages and provide little of substantive interest on those pages to fans.

Finally, GateWorld episode pages also evinced a unique linking pattern. Pages with links to transcripts or reviews were associated with more inlinks, but having links to the forum, especially if the page also contained many JPG images, were associated with fewer inlinks. Photographic JPG galleries and the forum are each presented as separate sections of GateWorld, whereas reviews and transcripts are embedded within the conceptual architecture of the episode pages. If fans were looking for the galleries or forums, which were not studied for this project, they might link directly to those sections, rather than going through the episode pages. Hence, only enriched textual content may increase inlinking

traffic to the episode pages.

### *Authorship-related*

This sub-section addresses associations between inlinks and the numbers of unique authors and revisions to pages, variables which are presented/studied on their own in §§4.4.3 and 4.4.4. As discussed further in those sections, author and revision counts were available only for the wiki sites.

The most prominent pattern for these variables was on the crew pages of Wikia. There, the combined presence of many authors, many revisions, and if the crew member had been the writer of many productions, was associated with a large number of inlinks. The long writer list alone also implied many inlinks. However, not having either this combination or a long writer list implied many fewer inlinks than average. This suggests that a class of prolific crew members existed that inspired many users to actively participate in constructing their Wikia pages. However, though a crew member may be prolific and attract either many authors or revisions, all three must exist in order to attract many inlinks from the outside world. The pages that qualify are those of the TV franchise's main creators, producers, and writers (i.e., Robert C. Cooper, Joseph Mallozzi, Andy Mikita, Paul Mullie, Martin Wood, and Brad Wright). The page about the director of the original film (Roland Emmerich), who is not involved in the TV franchise and has moved on to other topics, did not qualify.

Wikipedia crew pages, on the other hand, saw more inlinks when more authors were present, and fewer when long writers lists were present. This suggests a simpler relationship between participating authors and inlinks, and affirms again that those who link to

Wikipedia articles seem to prefer less listy content.

Smaller patterns also existed on other Wikia pages. On actor, book, and omnipediac pages, more revisions generally meant slightly higher inlinks. This is a similarly simple participation-equals-inlinks pattern, as in the previous paragraph. On episode pages, having both many authors, revisions, links overall, and production notes meant slightly higher inlink counts. However, if production notes were replaced with sentences beginning with prepositions, slightly fewer inlinks occurred. Although this difference is not great, it supports the idea, seen throughout this dissertation, that, unlike Wikipedia, Wikia users were more interested in lists of esoterica or desiderata than in textual portraiture.

### PageRanks

In this sub-section, the results of polytomous ordinal logistic regressions will be presented, having pages' PageRank (PR), rather than inlink count, as the dependent variable and all available IQ variables as possible independent variables. Like the previous sub-section, to facilitate interpretation, the effects have been organized into types, and the same types as the previous section were used.

#### *Content-related*

The presence of both JPG and GIF images were strongly associated with Wikia book pages having high PRs, and JPGs were associated with slightly higher PRs on Wikipedia actor pages, though JPGs meant slightly lower PRs on GateWorld episode pages. Not many book pages had images on Wikia, however those that did (e.g., of their covers) typically had higher PRs, indicating greater popularity and possibly a connection to publishers' com-

mercial sites. The same appeared true for actor pages on Wikipedia. On GateWorld, JPGs could be an indicator of editor preference, which §4.3.8 showed not to be synonymous with fan preference, potentially lowering their PRs.

Author fields implied lower PRs on several types of GateWorld pages, namely color authors for comic pages and excerpts authors for episode pages. These author fields occurred most on niche pages, and rarely on popular titles, confirming that PRs indeed indicate the most mainstream of content.

Writing credit fields, listing what the writer has written, implied slightly higher PRs on Wikia crew pages, and considerably lower PRs on GateWorld episode pages. As the most successful writers are the most likely to have credit fields, those pages' PRs should be higher. In contrast to the author pattern in the previous paragraph, most episode pages listed their writers. Hence, the issue is not one of obscurity, but of ubiquity. Since writers were listed even on many mundane/low-PR pages, overall, they were more associated with low PRs than high PRs.

Though with different variables, episode pages across the sites had a number of similar PR effects. On Wikipedia, captions implied high PRs; on GateWorld, guest star and question lists; and on Wikia, lengthy cast lists. Lower PRs were associated with, on Wikia, episode numbers and lists of media formats in which the episodes were available for purchase; and on GateWorld, production information. Hence, on popular episodes, users and editors had typically gone to the trouble of providing captions for images, full lists of cast members, and important plot questions. Less popular episodes provided more trivial information, in which the general public/consumer would be unlikely to be interested.

Similarly, omnipedia-like pages on GateWorld and Wikia manifested variables related

to PRs, and most were positively related. The following variables were associated with higher PRs on Wikia: listing many alternate reality events, giving Earth's interest in a planet or people, listing the status of an action or people, giving an image of the subject matter, listing the target of an attack, the presence of many discussant notes and trivia items, listing the power source of something, and listing someone's military rank. Also, on GateWorld, listing the race who uses a certain ship as well naming the home-world of a race implied higher PRs. The only variable associated with lower PRs were listing the galaxy in which an event occurred. Most of the variables associated with higher PRs regard people, and occasionally technologies, whereas the one lower PR-related variable most often regards planets. Hence, most mainstream users/linkers of the omnipediac pages were probably most interested in the peoples and technologies depicted in the franchise, and not in the locations or astronomical context of the episodes.

Staying on the theme of people, a number of other people-related variables were associated with PRs across different sites and sections. On IMDb's character pages, listing many quotes was associated with higher PRs than average. On Wikipedia, crew pages giving a person's title and textually describing their early life, as well as actor pages giving textual descriptions of career details and listing their theatrical appearances, also had higher PRs. High-PR Wikia actor pages similarly listed birth information. The only negative associations were on Wikia omnipediac pages, where giving textual biographical details lowered PRs. In addition to the recurring theme of Wikipedia pages providing more textual information and Wikia more listy information, these findings suggest that quotes were a primary feature of popular IMDb character pages, and that, like IMDb, Wikipedia often lists theatrical appearances for popular actors.

Finally, summary texts implied higher PRs on Wikipedia episode pages, but lower on Wikia episode pages, though Wikia omnipediac pages with summaries of past events were had higher PRs. While again generally confirming that popular Wikipedia pages were more text-affirming, popular Wikia pages were willing to tolerate some text, usually when it either recounted the historical dealings of characters or peoples, or the capabilities of ships and technologies.

***Word usage-related***

Regarding text lengths, PRs were higher on GateWorld episode and omnipedia pages having many words, and lower on episode pages having many characters. This suggests that popular GateWorld pages had many short words, and less popular pages longer words. Shorter words are more indicative of texts intended for the general public, who have a fairly low average reading level (i.e., around an $8^{th}$ grade level; U.S.DepartmentOfEducation, 2003). Longer words may be used in more obscure pages, perhaps intended for the more highly educated nerd population typical of science fiction fansite users (cf. §4.3.7).

In terms of sentence lengths, GateWorld book pages with many relatively short sentences had higher PRs, as did, do a lesser degree, Wikia book and omnipediac pages. Oppositely, Wikia episode pages with many very long sentences had higher PRs. Finally, Wikia actor pages with many relatively long sentences had lower PRs. This suggests that, with the exception of episode pages, most page types on the smaller sites were more popular when they had short sentences. This is consistent with the previous paragraph's observation that the general public has a low reading level. Though the episode page exception is somewhat surprising, such pages are the most likely to contain lengthy textual plot descriptions and

interpretations, which often contain lengthy sentences.

Crew pages did not dominate general word usage variables with PRs, as they did with inlinks. However, to-be verbs did imply high PRs on Wikipedia actor pages, as they did inlinks on Wikipedia crew pages. Nominalizations and conjugations also had a similar positive association with PRs as they did with inlinks, though, in this case, with Wikia book, actor, and omnipediac pages. Wikia actor pages words of average length (in terms of characters) also implied high PRs. Low PRs were found on Wikia actor pages with words of average length (in terms of syllables), on Wikipedia actor pages with many passive sentences, on GateWorld omnipedia pages with many prepositions, and on GateWorld book pages with many to-be verbs. These low-PR patterns bear no resemblance to the inlink findings. Rather than concluding that inlinking fans prefer different general writing styles, these findings suggest the highest PR pages on Wikia and GateWorld follow a writing style more like that used on the Wikipedia pages preferred by more general inlinkers, namely having: many to-be verbs, many nominalizations and conjunctions, few passive sentences, and few prepositions. This would be a more complex writing style than was present on most Wikia pages.

Regarding sentence beginnings, the most pronounced patterns were that IMDb title pages with many article beginnings had much higher PRs, but GateWorld episode pages with many prepositional beginnings had much lower. Other effects were less pronounced. Wikia episode and GateWorld omnipedia pages' sentences also began with many articles and had somewhat higher PRs. Wikia omnipediac and Wikipedia actor pages also began with many either pronouns or interrogative pronouns, respectively, and had slightly higher PRs. However, Wikia book pages' sentences began with many pronouns and had moder-

ately lower PR scores. Articles and pronouns are again often indicative of simple sentence constructions, and prepositions more complex when used in sentence beginnings. Hence, one would expect the general public to prefer the simpler beginnings and avoid more complex ones, as was the case here. Interestingly, more complex beginnings appeared to be expected on Wikia book pages, perhaps because their text was more literary.

Finally, readability variables were most polarized on GateWorld episode pages, where having high ARI scores implied much higher PRs, and high Kincaid scores much lower PRs. High ARI scores also meant slightly higher PR scores on Wikia omnipediac pages, and Fry scores slightly higher PRs on IMDb character pages. High Fog scores also implied moderately lower PRs on Wikia episode pages. ARI prioritizes high word and/or character counts, Kincaid and Fog high syllable and/or word counts, and Fry high sentence and syllable counts. Therefore, since word counts figure into both ARI and Kincaid/Fog, pages of these types with high character counts tended to have higher PRs, and with high syllable counts lower PRs. This pattern, manifest on GateWorld and Wikia, is somewhat contradicted by the small relation between Fry and IMDb, indicating that IMDb may follow a somewhat different pattern. One might interpret from these patterns that higher syllable counts could indicate lengthier words, which would probably be less preferred by the generic public, though character counts would remain high even with many small words. Longer sentences and more syllables on IMDb could mean that popular sites prefer the more verbose IMDb character pages, though such pages were typically so listy that not much text would be required.

### *Link-related*

Many fewer link- and authorship-related effects were related to PRs than to inlinks.

External links on GateWorld omnipedia pages implied higher PRs, rather than the negative relationship that GateWorld episode guides had with inlinks. This was probably because external links on omnipedia pages were most often either to IMDb or to actors' personal sites, and were only present for the most successful actors. Hence, those pages should have been more likely to have popular sites link to them.

On IMDb character pages, and Wikipedia list and crew pages, having many inlinks also meant having a higher PR. This suggests that most of those inlinks were from other large/high-PR sites, rather than from many small fansites. One should also notice that both IMDb and Wikipedia are large sites, making them the probable reference targets of other large corporate and non-corporate sites.

Finally, unlike with inlinks, links to both the forum and transcripts on GateWorld's episode pages consistently meant higher PRs. Forum links were present on every SG-1 and Atlantis episode page, and transcripts were available for every SG-1 and Atlantis episode only of certain seasons. That higher PRs are related to this pattern, and somewhat higher to forum than transcript links, suggests that the SG-1 and Atlantis episode pages were generally popular, certainly moreso than Infinity or Universe, and, by virtue of their near constancy, these two link variables tracked fairly well the higher SG-1 and Atlantis PR scores.

### *Authorship-related*

Both Wikia and Wikipedia actor pages had much higher PRs, when pages had more authors. Unlike inlinks, no relationship existed between numbers of revisions and PR values on any site. Since these are actor pages, and actors often have a stake in having reputable Web presences, one might wonder whether actors and their companies edit these pages, such that the heavier their involvement the higher the page's PR. It might also just be that more popular actors' pages tend to receive more attention from both fans and other popular sites.

## Conclusion

For this conclusion section, the many findings pertaining to this research question have been organized into topical groups, which will be presented from most-to-least general.

### *General trends*

Overall, pages with high inlink counts or PageRanks had more substantive content, and those with low inlink counts or PageRanks had more obscure content. On Wikipedia crew pages, those pages with many inlinks were characterized by substantive information about the person's work and personal life, whereas less oft-cited pages provided more superficial information, which often referred the user elsewhere. Wikia actor pages received more inlinks, when an actor's nicknames were present, and slightly fewer when long lists of trivia were present. On episode pages with high PageRanks on both Wikipedia, Wikia, and GateWorld, users and editors had typically gone to the trouble of providing captions for images, full lists of cast members, and important plot questions. By comparison, less popular episodes provided more trivial information, in which the general public/consumer

would be less likely to be interested.

As another prime example of substantive content on popular pages, on both wiki sites, crew pages with photographic JPG images were more cited from off-site. In a more nuanced pattern, on Wikia actor pages, the least notable of actors (i.e., those both having the fewest inlinks and who were known by the researcher to be less famous) were more likely to have JPG photos, possibly in an attempt to promote themselves. However, when an actor was very famous, they would have a JPG in addition to descriptions by fans of their physical features (e.g., hair color).

On omnipediac pages, standard popular and substantive content manifested in pages about individual people and technologies being the most linked-to, and pages about logistical context being less. On GateWorld omnipedia pages, character and place/planet pages usually had more inlinks than race or ship pages. Also, on both GateWorld and Wikia omnipedia pages, those pages depicting peoples and technologies had higher PageRanks, whereas those giving locations or astronomical contexts of the episodes had lower. Finally, on Wikia omnipediac pages, those pages containing fields about governing bodies received more inlinks. Specifically, pages containing fields about governments' legislation, re-organization, and religion received the most, and containing fields about executives, establishment, language, currency, and dissolution the least.

The most obscure pages had the following patterns. On GateWorld episode and comic pages, author fields occurred most on pages with low PageRanks. However, although movie comic pages usually received more inlinks, a few popular television comic pages arose to the movie inlink level and were likely to identify the author who colored them. On Wikia book pages, slightly fewer inlink counts occurred when an ISSN was present. Also, on

GateWorld's book pages, most pages had release dates, and the few without them were non-fiction guide books to the franchise. Finally, on GateWorld's episode pages, the two active series (at the time, SG-1 and Atlantis) had somewhat lower inlink counts than did the few pages dedicated to either anticipated or obscure series (Universe and Infinity). Many of these points suggest that GateWorld may have been considered by the general public to be the place to go for obscure information about the franchise.

### *Complexities*

Though generalization is difficult, Wikipedia was usually more textual and substantive, and Wikia usually contained more lists and esoterica. The distinction between being more textual or list-oriented was evident in the way in which Wikipedia and Wikia documented crew members and episodes, as well as actors on Wikia. However, Wikia was willing to tolerate some text on its pages, usually when either recounting the historical dealings of characters or peoples, or the capabilities of ships and technologies. Also, Wikipedia and IMDb contained what might be called substantive lists, which provided more than biometric descriptions of phenomena, as often occurred on Wikia, and which were central to certain popular page types on those sites. Namely, pages containing long lists of quotes were highly PageRanked on IMDb's character pages, as were lists of theatrical appearances popular actor pages on Wikipedia. Finally, evidence existed on Wikia actor, book, episode, and omnipediac pages that its users were more interested in lists of esoterica, vaguely reminiscent of role-playing games such as Dungeons and Dragons, than in texts.

The lengths and language styles of texts on the wikis were also distinguishable both from each other and from texts on GateWorld. Regarding length, those who inlinked to

both wikis' crew pages did not simply prefer longer texts, but rather texts with many words and, in Wikipedia's case but not Wikia's, many sentences. By comparison, those who cited GateWorld's comic and omnipedia pages preferred length by any measurement. Passive sentences were perhaps more indicative of a general documentation style (cf. §4.2.5), which produced lengthy text with typical sentences and word usage, something that both wikis' crew pages would probably avoid, but that the GateWorld comic page might not.

Regarding language styles, on a general level, most highly PageRanked episode and omnipedia pages (e.g., on GateWorld) had many short words, and lower PageRanked pages longer words. Shorter words may be more indicative of texts intended for the general public, who have a fairly low average reading level (i.e., around an $8^{th}$ grade level; U.S.DepartmentOfEducation, 2003). Consistent with this is that, with the exception of episode pages, most page types on the smaller sites (i.e., GateWorld book, and Wikipedia actor, book, episode, and omnipediac pages) had higher PageRanks when they had short sentences. Similarly, on IMDb title, GateWorld episode and omnipedia, and Wikia book, episode, and omnipediac pages, PageRanks were higher for pages containing many sentences with simpler beginnings. Interestingly, more complex beginnings appeared to be expected on Wikia book pages, perhaps because their texts were more literary. Also on Wikipedia crew pages, those that contained more variety in word lengths were more highly inlinked. Basically, though popular pages generally had substantive and lengthy content, that content was usually written for digestion by the general public, whereas less popular as well as more literary pages were written more for connoisseuring fans (echoing §4.3.8).

The most highly regarded writing styles on the wikis and GateWorld can also distinguished further. Those who linked to the two wiki sites' crew pages were shown to pre-

fer different word usages: Wikia many prepositions and pronouns, and Wikipedia many conjunctions and nominalizations. Highly inlinked GateWorld comic pages were also shown to center most around interrogative sentence beginnings. Finally, the highest PageRanked pages on Wikia and GateWorld followed a writing style more like that used on the Wikipedia pages preferred by more general inlinkers, namely having: many to-be verbs, many nominalizations and conjunctions, few passive sentences, and few prepositions. This would be a more complex writing style than was present on most Wikia pages.

***Evidence of manipulation***

Although PageRanks were usually comparable to inlinks, there was evidence suggesting preferential attachment, manipulation of pages by their stakeholders, and favoritism among large sites.

On most page types (i.e., GateWorld comic, episode, and omnipedia; IMDb character; Wikipedia list and crew; and Wikia omnipediac), having a higher PageRank equated with a general notion of social popularity (i.e., high inlinks), not just popularity among popular sites. That is, most inlinkers also gravitated, or were redirected, towards highly PageRanked pages, in the rich-get-richer pattern (i.e., preferential attachment, cumulative advantage, etc.) commonly associated with the world's use of large, centralized search engines (Barabási and Albert, 1999).

Some Wikipedia crew pages' negating of this pattern, having high PageRanks but few inlinks, might be due to those pages having been taken over by powerful stakeholders, who point their highly PageRanked corporate sites to their own pages and tightly control those pages' contents. Evidence of similar manipulation existed on both wikis' actor pages. On

Wikia crew pages, a class of prolific crew members existed that inspired many users to actively participate in constructing their Wikia pages. However, though a crew member may be prolific and attract either many authors or revisions, all three factors had to exist in order to attract many inlinks from the outside world. The reason may be simple coalescence around popular people, or some level of orchestration may exist among fans and/or the franchise's stakeholders.

Finally, the especially high PageRanks of IMDb character pages and Wikipedia list and crew pages, compared with their inlink values, suggested that those pages may be common reference targets for other large sites, possibly indicating favoritism or even deals between these large organizations and others.

### 4.3.11 Link analysis

**RA11:** Which pages on the fansites contain the most links, and where do links on pages usually go?

These questions imply what Schneider and Foot (2005) called Web sphere analysis, also called link analysis (Herring, 2007), a branch of Web content analysis in which link types, destinations, and origins are counted and those counts reported directly. In this section, for each website, content analytic results will be presented for internal and external links as well as internal and external inlinks. For the purposes of this project, an *internal link* refers to a link originating from the page being examined, and going to another page on the same website. An *external link* originates from the page being examined, but goes to a page on a different website. An *internal inlink* originates is a link from another page on the website

under study, and goes to the page being examined. And, an *external inlink* originates from a different website, but goes to the page being examined. Though the research question does not ask for inlink results, they will be reported regardless, because inlink information was collected as part of answering the research question in §4.3.10, and because inlinks are a standard part of link-based content analysis.

Links were automatically parsed and counted from each site's sampled pages using standard POSIX utilities. As also said in §3.3, inlink records for each page were obtained from Yahoo!'s free Site Explorer API (Yahoo!, 2009), and were automatically parsed and counted using standard POSIX utilities.

For each website, link frequencies will be presented first, followed by inlink origins and link destinations. A summary of patterns within and across the site sections is provided in this section's conclusion.

**GateWorld**

For each sub-section of GateWorld, table 4.33 presents counts of total links, external links, internal links, total inlinks, external inlinks, and internal inlinks. Averages (means) of each link type per page are given in parentheses.

Also, one might expect the numbers of internal links in tables 4.33 through 4.36 to be equal, or at least similar, to the numbers of internal inlinks, because the internal links on one page are the internal inlinks on another. However, this is not the case. This could be because inlink data were obtained from Yahoo!, but internal link data were obtained by using POSIX utilities to count instances of relative/internal links in either the HTML or MediaWiki markup language. Yahoo!'s index of these pages' links could have been

incomplete or outdated. Also, the POSIX parsing and counting scripts were written to find anything that could have possibly been a link. Hence, syntax errors or improperly used link tags occurring on sites could have been collected as links, possibly inflating the counts. The researcher attempted to mitigate such errors by manually reviewing all links collected, and did not find many erroneous records requiring deletion.

A third possibility is that the data are accurate, and that the unevenness represents interlinking between sections. For example, GateWorld's omnipedia pages having more internal inlinks than internal links, and the opposite pattern on episode pages, could mean that many episode pages link to omnipedia pages, but the reverse is not true. Though not within the scope of this project, a network analysis of link structures on the pages of these websites could be interesting future work. A sense of the most common network structures should become evident from the analyses of link origins and destinations given throughout this section.

Finally, for GateWorld and IMDb, which only provided their data in HTML format, the internal navigation link counts were removed from the results reported in this section. This makes their data more comparable to the wiki sites, which provided their data without the HTML template. Supplementary link counts (e.g., a list of links to similar pages, located within a page's content body) were not removed, because they occurred in both the editor-controlled sites' HTML and the wikis' page markup.

On GateWorld, the pages containing the most total/overall links (not inlinks) on average were episode and book pages, with comic and video game pages having moderate overall links, and omnipedia pages having few. On all pages, the majority of links were internal, with small numbers of external links following the same order of prominence as for total

Table 4.33: Frequencies: Links on GateWorld

| | **books** | **comics** | **episodes** | **omnipedia** | **videogames** |
|---|---|---|---|---|---|
| total links | 8,490 (133) | 2,626 (97) | 58,778 (163) | 23,162 (9) | 795 (80) |
| external links | 581 (9) | 135 (5) | 6,800 (19) | 908 (0) | 52 (5) |
| internal links | 7,909 (124) | 2,491 (92) | 51,978 (144) | 23,162 (9) | 743 (74) |
| total **in**links | 3,590 (56) | 639 (24) | 20,578 (57) | 49,971 (20) | 318 (32) |
| external **in**links | 1 (0) | 0 (0) | 668 (2) | 189 (0) | 16 (2) |
| internal **in**links | 3,589 (41) | 639 (24) | 19,910 (55) | 49,782 (19) | 302 (30) |
| pages (n) | 64 | 27 | 360 | 2,562 | 10 |

links. Inlinks also followed the same general pattern, though omnipedia pages's counts were large enough to be called moderate. External inlink counts were as small or smaller than were external link counts.

These findings show that internal linking activity is considerably more common on GateWorld than is external, though enough external links exist for most pages that each page could either contain or receive several. That episode pages would contain the most links is not surprising, as such pages are the cornerstone of GateWorld's content. The abundance of links on book pages, which are fewer in number and much more peripheral on the site, is somewhat surprising, until one realizes that every book page lists links to every other book page on the site. The same is true for comic and video game pages. Episode pages only linked to the other episodes in the same season. Regarding external links, episode pages, by far, linked to the outside world most often, followed by omnipedia,

book, comic, and video game pages. This indicates that GateWorld does not try to trap users. Finally, the relatively sparse amount of external inlinking shows that these pages make more reference to other sites than other sites do to them. This suggests that the outside world primarily links to GateWorld for its episode pages, secondarily for the omnipedia, and tertiarily for its video game pages.

In the rest of this sub-section, interpretive summaries will be provided of link destinations and inlink origins on each of GateWorld's sections. These summaries are based upon automated counts of unique internal and external links and inlinks for each site section, sorted in order of most-to-least frequent. The complete lists are not being included because, as shown in the tables 4.33 - 4.36, many of the lists contain thousands or tens of thousands of entries. Typical link counts will be given in parenthesis for every link pattern mentioned.

Links from book pages on GateWorld most often went to the GateWorld homepage (390 times across all pages), followed by links to GateWorld's Facebook page (130 times), forums (113 times), Youtube page (65 times), a sci-fi news aggregation site (Scifi Stream; 65 times), their podcast on iTunes (65 times), and vendors for buying media (i.e., Amazon and Big Finish audio books; 1-12 times). Links also often (on 64 pages) went to other book pages, as supplementary navigation on every page. Next were links to omnipedia pages (1-6 pages each), in all categories of the omnipedia, as well as links to individual episodes. Links to individual forum threads occurred on several pages (1-2 pages each). Finally, links to individual pages on authors' webpages appeared on the appropriate book's page. Regarding inlinks, most, again, came from other internal book pages' supplementary navigations. Links from GateWorld's own editorial interviews and news posts accounted

for 63 internal inlinks, each occurring in 1-6 pages. Eighteen omnipedia pages linked to the book pages, each no more than once. Twenty two comic pages linked to the book pages, also each only once. Finally, the one external inlink came from a forum page on Wikia's Stargate wiki.

Comic pages also had many (162) links to the homepage, Facebook page and forums (54), GateWorld's Youtube and Twitter accounts (27 each), iTunes podcast (27), as well as supplementary navigation on every page to all other comic pages. Otherwise, an assortment of links to omnipedia and episode pages accounted for 112 internal links. The rest were links to images of either the covers or individual pages of comic books (1-2 links for each image). Such images were presented one-per-page and were separate from the site's photo galleries. Regarding inlinks, in addition to the usual supplementary navigation, every comic image page linked back go that comic's main page. No external inlinks were reported by Yahoo!.

Episode pages also generally shared the homepage, Facebook, social networking, podcast, Amazon, and supplementary links with the previous pages, suggesting that these are a kind of informal/organic navigation structure on each page. Between 108-186 links went to the omnipedia pages of the main characters and races, and 1-86 links went to each of an assortment of mostly character and race omnipedia pages. Between 1-80 links went to IMDb for each main and guest cast members appearing in episodes; such links only went to IMDb's name pages, none to title or character pages. The majority of the remainder went either to other episode pages, episode review or transcript pages, image galleries (e.g., for promotional and screen capture images), to related GateWorld news blog posts (35), or forum threads. A handful of external links went to the producer's weblog, cast and crew

interviews with the mass media, small fansites (e.g., rdanderson.com), free episode downloads on the studio or network's site, and one Wikipedia article (on Lagrangian Points).

Regarding inlinks to episode pages, the majority came from either each season's index page, other episode pages, or the omnipedia pages of the main characters and races. Links from GateWorld interview and news pages were moderately common (1-16 times each), as were links from less common omnipedia pages (1-10 times each). External inlinks were dominated by fans' personal sites or forums, with most sites accounting for around 20 inlinks. Ten inlinks came from episode and actor pages on Wikia. Seven came from episode and list pages on Wikipedia. Judging from the top level domains in the fansites URLs, they originated in the following non-US countries: Australia, Belgium, the Czech Republic, France, Germany, Hungary, Russia, and the United Kingdom.

Omnipedia pages primarily linked to other omnipedia pages, and contained no navigation structures, as their pages were part of an HTML frameset. The most popular topics and people accounted for as many as 500 links each. Next most common were links to episodes and films, each accounting for as many as 75 links. Third most common were links to the IMDb pages of actors who played certain characters, accounting for as many as 5 links each (i.e., some actors played multiple characters, especially when in costume or when providing voice-overs). Some character pages linked to actors' personal websites, instead of IMDb. Finally, 22 linked to articles on the Encyclopedia Mythica (Lindemans, 2010), a folklore site to which many academians have contributed. Regarding inlinks, the majority came from other omnipedia pages. Those that did not, of which there were 189, came from small fansites or forums.

Finally, video game pages, in addition to the usual supplemental links, primarily linked

to screen capture images from games. Like the comic book images, each image was given its own page, and was separate from the site's photo galleries. Regarding inlinks, as with comics, most came either from other game pages or their image sub-pages. Forty links came from GateWorld interview or news pages. All 16 external inlinks came from either personal fansites or small forums.

### IMDb

Table 4.34 presents link counts and means for IMDb.

Table 4.34: Frequencies: Links on IMDb

| | **characters** | **titles** |
|---|---|---|
| total links | 1,899 (90) | 5,951 (186) |
| external links | 22 (1) | 487 (15) |
| internal links | 1,877 (89) | 5,464 (171) |
| total inlinks | 165 (8) | 2,683 (84) |
| external inlinks | 1 (0) | 2 (0) |
| internal inlinks | 164 (8) | 2,681 (84) |
| pages (n) | 21 | 32 |

On IMDb, title pages contained more of all types of link than did character pages, and the link vs. inlink ratio was similar for both page types. Although both page types contained long lists of links as their primary content, title pages are at the core of what IMDb has always been (i.e., an index of movie titles), making their content considerably

more developed than on character pages. Title pages linked to external sites considerably more than did character pages, indicating both that title pages are more developed and that IMDb does not try to trap users on its site. Very few other sites linked to these pages, though the samples were small. However, even in these samples, the expected result, that title pages would attract more inlinks because they are more established/mature, is visible.

Non-navigation links on character pages were primarily to title and name pages of the franchise's main episodes and actors, with each title accounting for 1-11 links in the dataset and each name 1-3. All character pages also linked to multiple sub-pages (e.g., biography and quotes) of themselves, provided an "update" button link to submitting page revisions to the editors, and linked to at least one advertisement hosted by DoubleClick. IMDb also had a partnership with Hulu, such that 2 pages linked to episodes hosted/sponsored by Hulu within IMDb's interface. Inlinks to character pages came primarily from title and name pages, accounting for 109 and 36 links, respectively. Nineteen came from other character pages. The one apparently external inlink actually came from a domain outside either the IMDb.com or Amazon.com domains, but which redirected back to IMDb, meaning there were actually no external inlinks found to the sampled character pages.

Title pages' non-navigation links were also primarily to the name pages of the main actors (46-51 times) as well as to the SG-1 general title page (38 times). All pages also contained a link for submitting update requests to the editors for the page's content. Other prominent name, title, and character pages followed in frequency (each 1-30 times). As with character pages, many of the links to title pages were to sub-pages of the current title page. Links to certificates (i.e., movie association ratings, such as PG for "parental guidance" in the USA) were also quite common (each 1-19 times), as were links to Help pages

(1-23 times). Next most common were links to keyword/tag index pages from the genre folksonomy (1-17 times). Some of the keywords (e.g., Drama) were always capitalized, probably came from a controlled vocabulary, and may have been applied by the site's editors, whereas others (e.g., surrealism) always appeared lowercased, seemed to be emergent, and were probably applied by the site's users. Finally, links to TV listings, users' reviews of titles, Hulu-sponsored episode videos, pages showing promotional images, links to buy media at Amazon or rent it at Blockbuster, and to DoubleClick ads were often unique, though 1-2 of each were on every page. Regarding inlinks, most came from name pages (1-26 each), secondarily from title pages (1-18 each), and tertiarily from character pages (1-9 pages). Links from keyword or Hulu video pages were more rare (1-4 each), as were from TV listing pages (1 each). The two external inlinks both came from individual fans' sites.

Additionally, and only for the title pages, four pages contained a section called "External links" in their supplementary navigation, on the left-bottom of the page. That section contained the possibility of linking to sub-pages, given the following labels: showtimes, official sites, miscellaneous sites, photograph sites, sound clip sites, and video clip sites. On most pages, all of these links were colored grey, indicating that they were not functional. However, when a link was blue, it went to a page containing a simple bulleted list of links to external sites, deemed by either the site's users or editors to be of that type. The nine links found on the four pages with one of these sub-page links being active have been included in the external links count in table 4.34. Three of those links went to sites indexing brief episode summaries and cast lists; two to sites indexing picture and media galleries of television shows; one to a site providing earnings figures for the first Stargate

film; one to the network's Stargate page; one to a small fansite; and one to a fan's personal message board system.

### Wikia

Table 4.35 presents link counts and means for Wikia.

Table 4.35: Frequencies: Links on Wikia

| | **actors** | **books** | **comics** | **crew** | **episodes** | **games** | **omni** | **videos** |
|---|---|---|---|---|---|---|---|---|
| total links | 3,596 (10) | 9,326 (29) | 369 (15) | 743 (12) | 27,561 (74) | 318 (20) | 94,929 (32) | 1,059 (42) |
| external links | 542 (2) | 32 (0) | 0 (0) | 3 (0) | 1,136 (3) | 66 (4) | 930 (0) | 7 (0) |
| internal links | 3,054 (8) | 9,294 (29) | 369 (15) | 740 (12) | 26,425 (71) | 252 (16) | 93,999 (32) | 1,052 (42) |
| total inlinks | 4,246 (11) | 4,100 (13) | 179 (8) | 1,106 (18) | 26,676 (72) | 236 (15) | 51,901 (18) | 473 (19) |
| external inlinks | 39 (0) | 84 (0) | 0 (0) | 1 (0) | 353 (1) | 2 (0) | 1,012 (0) | 0 (0) |
| internal inlinks | 4,207 (11) | 4,016 (12) | 179 (8) | 1,105 (18) | 26,323 (71) | 234 (15) | 50,889 (17) | 473 (19) |
| pages (n) | 373 | 325 | 24 | 60 | 371 | 16 | 2,975 | 25 |

On Wikia, episode pages were the focus of both types of internal linking, followed distantly by video pages. Omnipediac and book pages gave more internal links to other pages, though crew and omnipediac pages received more. Actor and crew pages contained more internal inlinks than outlinks. Links to external sites were most common on episode, omnipediac, and actor pages. Inlinks from external sites were most common on omnipediac and episode pages.

These findings suggest that episode pages were probably considered by the Wikia users to be the core of page creation and citation activity on the site. Video pages were more thoroughly linked than on previous sites – comparably to omnipediac, crew, and book pages – possibly suggesting their especial importance to these users. Actor and crew pages'

being referred to more than linked from indicates that those pages are, to some extent, informational dead-ends, containing information that does not always link back to the franchise (e.g., cast and crew personal life information). Somewhat confirming this, actor, episode, and omnipediac pages were the most common to link to the outside world, though crew pages rarely did. The external inlinks show that the outside world most often comes to Wikia for omnipediac information, moderately for episode information, and rarely for book and actor information. That is, Wikia's main function for the international Stargate fan community is as a repository for encyclopedic information about the franchise.

Internal links on actor pages most often went to general pages about the series (11-158 times each). However, unlike the editor-controlled sites, the next most common links were to pages about the countries of the production companies (i.e., Canada and the USA; 25-59 times), the actor list/category page (22 times), and military and civilian ranks/titles (e.g., doctor, colonel, major; 1-15 times). Then came links to popular episode (1-8) and character (1-6) pages. Pages summarizing everything that happened in a given year were next (1-5), followed by category pages for actors with the same last name (e.g., Adams; 1-4). Finally, links to pages listing all appearances by a given character as well as those providing metadata for a multimedia file were fairly numerous, but were present on no more than one page. External links overwhelmingly went to the Wikipedia homepage (50 times). Next most common were links to the Wikia wikis of related franchises (i.e., Battlestar Galactica, Sanctuary, Star Trek; 1-7 times). Actor pages on TV.com and MovieTome were each linked to 1-3 times, and on IMDb and Wikipedia 1-2 times. Links to actors' personal sites were relatively rare and never occurred on more than one page.

Regarding inlinks on actor pages, most came from the actors category/listing page (50-

190), followed by from popular episode pages and season-wide episode listing pages (9-31). Many also came from pages describing other franchises (e.g., Star Trek and The X-Files; 1-15). Pages describing magazines and guidebooks came next (1-5), followed by users' personal Talk pages (1-2). External inlinks were evenly divided between, on the one hand, personal fans sites and forums, and, on the other, sites that aggregate news about celebrities. No external inlink occurred more than once.

Book pages also most prominently linked to general, location, and rank/title pages. Prominent actor, character, author, and crew pages were linked-to directly (1-80 times each), rather than through categories. Publisher pages came next (1-72), followed by date-event list pages (1-10), book title pages (1-4), magazine pages (1-4), and DVD collection guides (1-2). Links to episode pages were also relatively rare (1-5). The remaining internal links went to miscellaneous omnipediac pages or file metadata pages. Ten external links went to authors' personal sites. Eight went to Wikipedia, namely: the homepage, a publisher's page, a book title's page, and an omnipediac page. Eight links went to publishers' sites. Four went to sites about role-playing books. And, two went to the Star Trek Wikia wiki.

Regarding internal inlinks on book pages, the majority came from category/list and date-event pages (1-90), followed by popular author, publisher, and character pages (1-61). Fewer links came from popular episode pages (1-7), and fewer still from other individual book and magazine pages (1-5). The remainder were from omnipediac, user profile, and talk pages. External inlinks most often came from GateWorld's forums (3-4), as well as other personal fan sites and forums (1-2). Fourteen also came from Qwika – a wiki search engine – and the Swiss Open Directory Project's result pages.

Comic pages' links most often went to pages about Avatar Press (20 times), a comic publisher, as well as pages about prominent comic authors (1-14). The usual array of popular character, series, and date-event summary pages were fairly common (1-12). Marginally common were links to other comics and books (1-4). File metadata, category/list pages, and omnipedia pages were the least common. No external links were present from comic pages. Internal inlinks primarily came from the comic books category page (23), Avatar Press and popular author pages (1-10), popular book pages (1-5), and date-event list pages (1-5). Rare inlinks came from user and talk pages, season episode list pages, category pages listing unpublished books. No external inlinks were present. Hence, Wikia's comic pages were highly insular.

Crew pages most often linked to the franchise's creator (13), main writers and producers (1-7), and their most famous creations (1-9). Otherwise, only the most popular actors received scant linking (1-2). Two of the external links went to crew members' IMDb pages, and one went to the crew page of Jeff Woolnough on the Battlestar Galactica wiki, who directed several early episodes for both franchises. Internal inlinks came most from the writers and directors category pages (21-28). Otherwise, season lists (1-10), other crew pages (1-7), and episode pages (1-5) were most common. Links from cast and guidebook/magazine pages also occasionally rarely occurred. The one external inlink came from the Spock/Intelius people search page on Will Meugniot, the director of Stargate Infinity.

Episode pages most often linked to general series pages (219-432), the page about the studio (404), pages describing the main characters (1-244), and pages about the main crew (1-96). Omnipediac pages about places (e.g., Earth; 1-331), peoples/races (1-263), date-event lists (1-204), and ships (1-79) were also common. Links to PDF files occurred 166

times, and to Flash files 153. Actor pages were moderately common (1-64), as were pages about technologies (1-35). Pages about Star Trek were fairly common (1-16), as were links to other episode pages (1-8). The usual mixture of file metadata, user, talk, and category pages comprised the remainder. External links most often went to fansites hosted on AOL (9-48), followed by the producer's weblog (1-5), and GateWorld's news and episode pages (1-3). Transcript pages on the Dave.tv received 171 links, 143 on the SG-1 Solutions wiki, and 15 on the MGM site. Episode guides on the Syfy and SkyOne TV networks also received 64 unique links.

Internal inlinks to episode pages primarily came from pages summarizing a major plot arc (1-291), from pages listing a major characters' appearances (1-264), or from general series (1-197) or category (1-178) pages. Crew pages also often linked to episode pages (1-73), as did character (1-60) and date-event summary pages (1-41). Other episode pages inlinked moderately (1-33), as did season list pages (1-30). A mixture of omnipedia page types (1-7), user and talk (1-2), and magazine/guidebook pages (1-2) filled the remainder. External inlinks most often came from a personal Italian fansite (1-11), Squidoo sites on Atlantis and SG-1 (7-8), other personal fansites and forums (1-4), seven links from the Star Trek Wikia wiki, and 29 from the Hungarian Wikipedia. Judging from the domain names, fansites linked to these pages from the following non-US countries: China, Croatia, Czech Republic, France, Hungary, Ireland, Italy, Latvia, Poland, Russia, Spain, and the UK. This suggests a primarily European user base for these pages, consistent with the user profile results in §4.3.7.

Game pages internally linked to the games category pages (3-8), popular publishers (1-5), and developers/authors (1-4), in addition to the usual links to general series (1-7),

TV network (1-5), and studio (6) pages. Many links went to peoples pages (1-7), few to characters (1-2), and none to cast or crew, possibly because peoples/races are a better premise for a game than are notable individuals. External links most often went to official developer sites (38 links), followed by gaming magazine interviews with game designers (14 links). Fansites specializing in collecting information about one or more games were next (10 links). Two links went to the site of a commercial vendor of trading cards, and two went to the Stargate SG-1 role-playing game page on Wikipedia. Internal inlinks most often came from the games and role-playing books categories (4-9 times), followed by pages about planets, peoples, and characters featured in games (1-4). Several also came from other game pages (1-2), as well as pages about game publishers (1). The two external inlinks came from ads hosted on the ProjectWonderful advertising service.

Omnipediac pages internally linked most to the usual, and fairly uniform, mixture of pages describing the series overall as well as the most prominent characters, peoples and places (1-2,012). Links to science/nature topics (e.g., supernova, wormhole, vacuum energy, time dilation, etc.) were also fairly common (1-991), as were pages about ships (1-232) and date-event summary pages (1-185). Technology page links were somewhat less common (1-120), as were links to episode pages (1-56). Pages about languages, cultures, religions, and real-world cultural references (e.g., references to other franchises, food and drink, etc.) received relatively few links (1-13). External links most often went to the Wikipedia homepage (3-80) or to one of 298 Wikipedia pages, often about non-Stargate topics. The studio's homepage also received 17 links. Otherwise, 144 links went to an assortment of pages on the GateWorld omnipedia, 43 went to mostly character pages on the SG-1 Solutions wiki, 19 went to character profile pages on the SyFy Channel's website,

and 18 to Kate Ritter's tribute site to actor/producer Richard Dean Anderson. The majority of the remainder went to IMDb cast/crew pages or small fansites.

Internal inlinks to omnipediac pages most often came from category pages listing characters, planets, and peoples (1-184), followed by notable pages in each of those categories (1-155). The distribution of internal inlinks from other pages was highly similar to the distribution of internal links, given in the previous paragraph. External inlinks came most often from a game company that is trying to develop a Stargate-based space-shooter game (154 links). However, as with the episode pages, the bulk of links came from personal fansites and small forums. Additionally, 23 links came from character pages on the English Wikipedia; 16 came from actor and episode pages on the Star Trek Wikia wiki; 12 from the Czech Wikipedia; seven from the El Salvadorian Wikipedia; two from the Hungarian Wikipedia; and one each from the French, Italian, and Polish Wikipedias. Other non-US top level domains represented included: Canada, China, Germany, Iceland, Ireland, Japan, the Netherlands, Romania, Russia, South Africa, Sweden, Switzerland, and the UK. Though still prominently European, this is a somewhat more geographically diverse list than was present for the episode pages.

Finally, video pages internally linked most often to crew pages (1-47), followed by cast pages (1-17). Date-event pages were moderately common (1-9), as were video collection and DVD release pages (1-3). Episode page were rarely linked-to more than once, though quite a few were present. Of the seven external links, two went to the DVD collection's official website, two to GateWorld news pages, one to the SyFy Channel's site, one to the Wikipedia page describing the SyFy Channel, and one to a partwork/magazine publisher in the UK. Internal inlinks most often came from the DVDs and videos category pages (5-

13), as well as the pages listing the contents of DVD collections (1-3). Character, episode, people, and place pages provided the remaining inlinks. No external inlinks were present.

### Wikipedia

Table 4.33 presents link counts and means for Wikipedia.

Table 4.36: Frequencies: Links on Wikipedia

| | **actors** | **authors** | **crew** | **episodes** | **games** | **general** | **lists** | **omni** |
|---|---|---|---|---|---|---|---|---|
| tot. links | 28,580 (77) | 930 (44) | 3,243 (61) | 5,999 (128) | 166 (55) | 2,047 (256) | 3,610 (63) | 6,245 (173) |
| ext. links | 910 (3) | 69 (3) | 120 (2) | 227 (5) | 0 (0) | 0 (0) | 0 (0) | 60 (2) |
| int. links | 27,670 (75) | 861 (41) | 3,123 (59) | 5,772 (123) | 166 (55) | 2,047 (256) | 3,610 (63) | 6,185 (172) |
| tot. inlinks | 24,906 (67) | 997 (48) | 4,081 (77) | 818 (17) | 0 (0) | 701 (88) | 1,819 (32) | 2,574 (72) |
| ext. inlinks | 2,917 (8) | 288 (14) | 348 (7) | 41 (1) | 0 (0) | 11 (1) | 53 (1) | 102 (3) |
| int. inlinks | 21,989 (59) | 709 (34) | 3,733 (70) | 777 (17) | 0 (0) | 690 (86) | 1,766 (31) | 2,472 (69) |
| pages (n) | 371 | 21 | 53 | 47 | 3 | 8 | 57 | 36 |

On Wikipedia, general pages took the prominent place held by episodes on the other sites. General pages on Wikipedia are essentially larger episode pages that summarize an entire series (e.g., SG-1 or Atlantis). Though the other sites had pages about entire series, they are more like episode lists or blogs than lengthy summary documents, so were not collected and will not be compared with Wikipedia's general pages. Omnipediac and episode pages contained many links to other pages, though crew and omnipediac pages received more. Actor pages were also prominent in both types of internal links on the site. Only actor, author, and crew pages had comparable internal link and inlink counts; other page types linked out considerably more than in. These three page types also contained the most inlinks from the outside world, and, along with episodes, linked most to the outside

world. This suggests both that pages about real-life people (not characters) had a non-franchise linking component, that episode pages referred often to outside sources, and that the outside world primarily links to Wikipedia's information about real-life people. Other than pages about people, omnipediac, list, and episode pages were the most linked-to by outsiders.

Internal links on Wikipedia actor pages most often went to the general series pages (SG-1 335 links, Atlantis 170 links), followed by pages on the United States and its actors (1-192 links each); Canada, British Columbia, Vancouver, and Canadian actors (66-145 links each); and other science fiction franchises filming in Vancouver (e.g., The X-Files, The Outer Limits, and Smallville; 1-64). Next most common were links to pages describing "[year] in film" (1-47), television networks (1-44), and franchises and filming locations more vaguely related to Stargate (1-35). Quite rare within these links were those to actual Stargate actors or titles (1-9). Few external links repeated. Rather, the individual pages of other sites received many links to specific pages. Unique IMDb names pages, for example, were linked-to 325 times, actors' official websites 92 times, TVToMe actor profile pages 27 times, TV.com person pages 20 times, All Movie Guide person pages 17 times, TV Guide person pages 10 times, GateWorld news pages eight times, and Stargate or related Wikia actor pages six times. The rest were links to interviews with actors and to fans' personal sites.

Countries (top level domains) outside the US to which external links went were as follows: Canada, France, Germany, Hungary, Japan, Nauru (an island nation in Micronesia, in the South Pacific), New Zealand, Tokelau (a South Pacific territory of New Zealand), and the UK. As with Wikia, this list evinces a European user presence, though also an Ocea-

nia presence, though one should note that top level domains can be registered by anyone, regardless of their physical location. Obscure domains, especially, are often purchased by western companies and individuals, either because .com was not available or because an obscure top level domain sounds/looks better with their primary domain name (e.g., bit.ly is located in New York City, not Libya). A manual search and inspection of the WHOIS record for each domain, which was beyond the scope of this project, would have been required, in order to verify the actual (i.e., to the degree that the WHOIS record is not falsified) location of the domains.

Internal inlinks on actor pages most often came from Stargate-related character, people, or place pages, or lists thereof (1-75). Less common were links from episode or year-in-film pages (1-15), as were links from related franchises and fan conventions (1-9). As with regular actor links, a long tail existed, containing at least one link from almost anything vaguely related to Stargate or Canadian science fiction. The vast majority of external inlinks came from personal fansites, small forums, and small sites listing celebrity profiles. However, 22 Wikimedia (Wikipedia's multimedia storage division) pages, mostly about actors and the US Air Force, linked to actor pages, as did 65 Wikiquote (Wikipedia's quotations division) pages, mostly about either Stargate or other franchises' episodes. Thirty Star Trek Wikia wiki pages about actors linked in, as did 15 actor pages on the Stargate Wikia wiki. Non-US countries from which the inlinks apparently originated included: Belgium, Bulgaria, Canada, China, Czech Republic, the EU, France, Germany, Greece, Hungary, Ireland, Italy, Japan, Laos (in Southeast Asia), Latvia, Lithuania, the Netherlands, Peru, Poland, Romania, Russia, Samoa (in the South Pacific), Slovenia, Spain, Sweden, Switzerland, the UK. As with external links, this list is primarily European, though includes more

of Eastern Europe, and different areas of East Asia and Oceania.

Author pages' internal links most often went to the pages of famous authors who have either worked on, or produced work related to, Stargate (e.g., Martin Day and Paul Cornell; 1-15), to category pages listing various types of writers and novelists (1-8), and to general pages about Stargate's series and similar franchises (1-10). Links to pages about publishers were also fairly common (1-5), as were links to related books, magazines, and comics (1-3). Rare were links to notable characters and themes in books (1-2). External links most often went to either authors' official sites (20 links) or sites posting interviews with authors (20 links). Twelve links went to companies with which authors were affiliated, eight went to small fansites, five went to author pages on the Internet Speculative Fiction Database (ISFDB), three to the Star Trek Wikia wiki, and one to an author's IMDb page. Internal inlinks on author pages most often came from pages listing types of Stargate literature (1-11), pages about publishers (1-10), or pages about notable books (1-10). Links from other author pages were infrequent (1-2), as were links from episode and other types of pages. External inlinks, again, came most from small fansites. Those that did not were 21 links from author pages on the ISFDB and 18 links from authors' official sites. Non-US countries represented in the top level domains included: Australia, France, Germany, Italy, Japan, the Netherlands, Poland, Samoa, Slovenia, Spain, and the UK.

Crew pages' internal links, after the usual references to general pages (13-66), most often linked to the most notable crew members or lists thereof (1-15), followed by their most notable works, awards, and affiliated cities (1-8). Year-in-film, production companies, and networks were next most common (1-4), and these patterns continued into ever-more obscure references. External links most often went to IMDb name pages (42 links total),

followed by personal profile/index sites (20), crew members' official sites (16), personal fansites (15), interviews on mass media and smaller sites (14), studio websites (8), TV.com person pages (3), the Star Trek Wikia wiki homepage (1), an ISFDB name page (1), and an All Movie Guide name page (1 link). Internal inlinks most often came from character (1-18 links each), episode and character list (1-18), and title (1-14) pages. Other crew pages linked to crew pages moderately (1-4), as did technology pages and other franchises' pages (1-3). External inlinks were again dominated by small fansites. Only five Wikiquote pages, two pages on the producer's blog, one IMDb name page, one Stargate Wikia wiki page, and one Star Wars Wikia wiki page did not fit that description. Non-US countries present in fansites' top level domains included: Australia, Denmark, Finland, France, Greece, Italy, Japan, Latvia, the Netherlands, New Zealand, Poland, Slovenia, Spain, Tonga, and the Ukraine.

Episode pages most often internally linked to the primary cast and crew members' pages (1-108), followed by general and network/studio pages (1-69). The page describing GateWorld was also prominently linked-to (48 links), as was IMDb (21 links) and Wikia (20 links) to a lesser degree. People pages came next (1-17 links each), followed by episode list pages (1-15), and pages about production locations and topics (1-6). References to similar franchises and cultural phenomena were common in the long tail (1-3). External links most often went to either the network or studio's episode pages (79 links), followed by quotes from episodes on Wikiquote (42), IMDb title pages (41), screenplays at Dave.TV (22), Wikia season or episode pages (15), GateWorld season pages (13), TV.com episode pages (10), The Numbers (a box office data site) title pages (2), a Rotten Tomatoes (a movie trailer and review site) title page, an All Movie Guide movie page, and an SG-1 Solutions

1 title page.

Internal inlinks to episode pages were most frequently from the SG-1 general page (6 links total), the major character and people pages (1-5 links each), and episode list pages (1-5). Technology and ship pages were moderately frequent inlinkers (1-3), as were other episode pages (1-3). Seven external inlinks came from episode pages on Wikia, two from the SG-1 page on Wikiquote, and one from a list of former Stargate cast and crew on the Battlestar Galactica Wikia wiki. Otherwise, small fansites and forums abounded, originating from the following non-US top level domain countries: Hong Kong, New Zealand, and Sweden. This is the first appearance of inlinks from Hong Kong, though not out of line with the East Asia trend.

Games pages' internal links – within the usual array of links to the general and studio pages, main characters, peoples, and places (1-11) – uniquely linked to pages about game development companies (1-5) as well as platforms/systems on which video games can be played (1-3). No external links were present, and no internal or external inlink records were returned by Yahoo!, probably due to the small number (3) of pages involved.

General pages most often linked to the Wikipedia page about GateWorld (56 links), followed by pages on each series within the franchise (15-39 links each), the studio (31), the franchise's main crew members (1-23), and the franchise's most notable technologies and alien races (1-16). Primary characters and their actors came next (1-9), followed by notable episodes (1-5), affiliated production and distribution companies and locations (1-5), and connections to other franchises and popular culture (1-4). No external links were present. Internal inlinks most often came from other general pages about series (1-7), followed by pages about the main characters, actors, races, and technologies (1-6). Links

from episode and ship pages were also fairly common (1-4), though links from year-in-television, related franchise, and cultural reference pages were either rare or rarely repeated (1-2). External inlinks, with the exception of one blog post from a novelist author, all came from small fansites or forums. The only non-US top level domain was Canadian.

List pages also most often linked to the page about GateWorld (37), followed by general series pages (1-36), major cast and crew pages (1-36), major alien race pages (1-34), and TV network and studio pages (1-20). Episode pages were linked-to moderately (1-10), as were pages about science/nature topics (1-8) and other list pages (1-8). Author, ship, technology, and IMDb and Wikia pages received minor linking (1-4). No external links were present. Internal inlinks most often came from peoples, technology, and character pages (1-16). Links from episode, cast member, ship, and other list pages were also common (1-13). Of the pages about the other three sites under study, only the page on GateWorld linked to list pages (13). These patterns continued into the long tail. External inlinks, as usual, primarily came from small fansites and forums. The only exceptions were two links from forums on the SyFy TV network's site. Non-US top level domains present in the links included: Belgium, Canada, the Czech Republic, Germany, Hungary, Russia, and Slovenia.

Finally, omnipediac page internal links most often went to the general series pages (87-176), the main character and crew pages (1-81), the GateWorld page (65), and TV network and alien race pages (1-60). Actor and episode pages were only moderately common (1-21), as were ship, science/nature, and technology pages (1-17). Obscure examples of these categories, as well as cultural references, made up the bulk of the distribution's tail. External links most often went to character or race pages on the Stargate Wikia wiki (23), followed by character and homepages of the SyFy network and studio (22), race pages on

the GateWorld omnipedia (6), and race pages on the SG-1 Solutions wiki (2). Individual links went to character pages on Wikiquote and TVtropes.org; title pages on IMDb and TV.com; a page on Archive.org translating the Gospel of Matthew into the Alteran language from Stargate; and two small fansites. All of these links appeared to be in the United States.

Internal inlinks to omnipediac pages most often came from alien race, ship, and technology pages (1-24), followed by character pages (1-22). Though the Atlantis page linked to many other omnipediac pages (24), this was not true for most general series pages (1-9). The pages of cultural references that appeared in Stargate and linked to omnipediac pages made up much of the long tail (1-2). Of the external inlinks, 60 came from small fansites and forums, 17 from season pages on Wikiquote, 11 from forums on xboxelite.com, eight from forums on londonfetishscene.com, and six from forums on mmorpg.com. Non-US countries represented in the top level domains included: Argentina, Australia, Bosnia and Herzegovina, Israel, Montenegro, the Netherlands, the Ukraine, and the United Kingdom. This suggests greater user concentration in Southeastern Europe than for previous page types.

**Conclusion**

In order to summarize the complex and large amount of link variability introduced in this section, the findings and counts were distilled into two matrices, one for internal (i.e., within-domain) links and one for external (i.e., between-domains) links, tabulating both the links and inlinks of each site's sub-sections with respect to the categorical types of link origins and destinations that emerged throughout this section. Each of these matrices was

subjected to column-wise principal components analyses (PCA), in order to address the first half of the research sub-question. Then, the link and inlink columns of both matrices were divided into separate matrices, transposed, and subjected to PCA, so as to answer the second half of the research question. The results of these six PCAs are interpreted in this conclusion.

### *Internal linking*

The most frequent source of internal linking on these sites were inlinks to Wikipedia episode and general series pages; both inlinks and links to/from Wikia omnipediac pages; and links from GateWorld episode pages, Wikipedia game and omnipediac pages, and Wikia episode pages. Whereas the asymmetry of links to Wikipedia episode and general pages indicates that they were the focal points of much unidirectional attention within Wikipedia, the symmetry of links to/from the Wikia omnipediac pages suggests that those pages formed a strongly connected group/clique, which was probably core to user activity on that site. The other pages mentioned asymmetrically provided more links to other pages than they received.

Sorting the site-link combinations by most overall links vs. inlinks shows that GateWorld omnipedia pages both received the most inlinks and were a strongly connected group. Wikipedia actor, GateWorld episode and book, Wikia game, and Wikipedia omnipediac pages also received considerable attention. By contrast, Wikipedia actor, crew, episode, and list pages linked to many pages of other types, as did crew and video pages on Wikia. Wikia actor pages also formed a fairly strong clique. On the contrary, the most infrequent link-page combinations were both links and inlinks on GateWorld video game

pages; inlinks to Wikia comic and book pages; inlinks to Wikipedia author pages; links from GateWorld book pages; and links from Wikia video pages.

IMDb pages had a markedly different linking structure than did links coming out of many wiki pages. IMDb character pages formed a clique, and were associated with inlinks to IMDb title pages. By contrast, Wikipedia actor and author pages, as well as Wikia actor, comic, and game pages, all linked out strongly to other page types.

Omnipediac pages' links were often associated with list, general, and book pages, and contrasted with actor, game, and video pages. Links from Wikipedia's list and general pages, as well as inlinks to its omnipediac pages, were associated with GateWorld omnipedia page links and inlinks, and less prominently with GateWorld book page inlinks. As an aside, book page links were, as one might expect, usually similar to author and comic links, namely: author links and inlinks on Wikipedia, comic links and inlinks on Wikia, and book inlinks on Wikia. In contrast to the list-general-omnipedia pattern, a clique of video game pages on GateWorld was associated with a Wikia game clique, inlinks to Wikipedia and Wikia actor pages, links from Wikipedia game pages, and inlinks to Wikia video pages. A clique of crew pages on Wikia was also associated with a links coming from video pages on Wikia.

Within book pages, there also exist linking structures associated more with games vs. more with videos, probably indicative of the difference between guidebooks for role-playing games vs. for DVD and special features collections. The associations were clearest in Wikia book pages' links' similarities to GateWorld video game links and inlinks, as well as GateWorld book pages' links' similarities to Wikia video inlinks. Less pronounced associations also existed between Wikia book links and GateWorld book inlinks with Wikia

video page links, and also between Wikia game inlinks, GateWorld video game links, and GateWorld book page links.

Turning now to the same dataset analyzed in terms of link origin and destination types (i.e., row-wise), *internal links* (i.e., those going to other pages on the same domain) most often went to omnipediac pages about places, peoples, characters, science and nature topics (e.g., wormholes), and overall pages about individual series. They went least often to pages that acted as gateways/boundaries for browsing away from the fansite, namely: pages about actors, and pages describing Wikia, IMDb, overall plot arcs, and production companies. The next most prominent internal linking pattern was to review, news, transcript, and episode pages. Links to one of these page types was usually associated with links to the others as well. This pattern was in contrast to links going to pages about crew members, authors, game developers, and pages describing what Wikia and IMDb generally have to offer.

Within the popular places-peoples-characters pattern, there also existed a contrast between technology and ship pages – which were linked-to similarly as were pages about filming locations, the studio, authors, TV listings, and the episode review-news-transcript pattern – versus pages about games, books, videos, and other franchises. This appears to be another type of core vs. periphery distinction, with technology and production-related omnipediac pages being nearer the core of the network, and pages about non-canon (i.e., not created by the main Stargate producers) or non-Stargate materials nearer the periphery. Similarly, links to comic pages are associated with book pages, and contrasted with list, forum, and about-GateWorld pages, suggesting a second level periphery, beyond (semi-)professional media based on Stargate, for links to pages either about or containing ama-

teur/fan activity.

*Internal inlinks* also most often came from people, place, ship, technology, character, and science and nature omnipediac pages. The least often came from pages about books, publication or production companies, authors or game developers, or user profiles; actor pages were not as infrequent of inlinkers as the link-basis analysis might have suspected. This book-related pattern was also contrasted against links from episode, crew, plot/theme, general, and actor pages, suggesting a similar overall structure at the core of the inlinking network as had the out-linking network discussed two paragraphs above. This indicates that the core of these internal linking networks is fairly strongly connected.

A third frequent structure can also be differentiated from a combination of the book and episode patterns, containing links from game, gallery, news, and video pages. This too combines two page types prominent in the out-linking network (i.e., games and news). Consistent with the previous analysis, and perhaps with intuition, comic, book, and news page inlinks also followed moderately similarly patterns, as did game, video, and gallery pages.

Finally, gateway/boundary pages – including pages about actors, other franchises, general series, lists, cultural references, user profiles, and, to a small degree, crew members – formed the periphery of the inlinking network.

### *External linking*

Links from these sites to external domains most often went to the GateWorld omnipedia, followed by Wikipedia actor, episode, and science and nature topics pages. This suggests that these pages are hubs of attention across all four fansites. Least frequent were links

to cast, crew, or authors' official personal webpages, Wikia actor pages, and several mass media indexing sites, namely All Media Guide and TV.com, suggesting that these receive little inter-site attention. After GateWorld and Wikipedia, links to pages on commercial vendors' sites were most frequent: Youtube, iTunes, Facebook, Amazon, and IMDb title pages. The mostly young male users of these websites are clearly in tune with the Social Networking Site and Web 2.0 generation. Also having similar linking patterns as IMDb title pages were GateWorld episode pages as well as pages on the TV networks' and studio's sites.

Separately from the commercial pattern, a nearly as prominent linking structure existed to small fansites, the weblog of one of the Stargate producers (Joseph Mallozzi), a Stargate wiki site that was too small to be studied for this dissertation (SG-1 Solutions), and episode transcripts and screenplays on Dave.tv. Were further research to be done on this franchise, those sites should be among the first considered.

Also separate from the previous commercial pattern, and related to the mass media indexing sites, was a contrast between links to several site-affiliated advertisers and to mass mass media indexing sites. Links to both Hulu sponsored videos and DoubleClick ads, both of which displayed their content within their parent site, followed a similar pattern. However, this pattern was different from links to (esp. commercial) indexing sites, such as: the Internet Speculative Fiction Database, interviews hosted on (mass) media sites, and official company websites (e.g., production and publishing companies). The embedded-advertisement linking pattern was also fairly similar to links to Amazon, as was the indexing-site pattern to links to Twitter and the GateWorld forums.

*Inlinks* from other domains to the four sites under study most often came from the Wikia

wikis of other franchises (most often Star Trek and Battlestar Galactica), from mass media indexing sites, and from users located in Romania and Switzerland. Although inlinking from other franchises and index sites makes intuitive sense, why Romanian and Swiss users would be such prolific linkers is a mystery. By contrast, the GateWorld forums, Hulu videos on IMDb, Wikia episode pages, and users in New Zealand provided the fewest inlinks to these sites, indicating little effort on their parts to link to the four sites. Users' countries are much more prominent in these results than in the previous results in this conclusion section, because many external inlinks came from small fansites and forums, most often located in Europe, East Asia, and Oceania.

The next most common inlinking pattern came from Wikipedia omnipediac pages, official company and persons' sites, TV network sites, the producer's weblog (to a small degree), and users in Russia, Slovenia, Canada, and the Netherlands. That the large and official sites would link to the fansites under study also makes intuitive sense, though the affiliation between eastern and northern European users with such sites is also mysterious. This pattern can be contrasted to links from the producer's blog (to a great degree), Wikipedia episode and list pages, Wikia episode pages, the Wikiquote site (to a small degree), and users mostly from continental or eastern Europe (i.e., Spain, Italy, Belgium, Hungary, the Ukraine and Czech Republic, Australia, and Poland). This suggests that the producer's weblog more resembles, at least in its linking behavior, a wiki episode or list page than a corporate site. Why different areas of Europe would resemble this style more than the other is again unknown.

Finally, the linking structures of non-English Wikipedias (i.e., 32 Hungarian pages, 13 from the old Czechoslovakian top-level domain, seven El Salvadorian, three Spanish, two

Polish, two Russian, one Italian, and one French) most resembled those of users from, in terms of greatest-to-least resemblance: China, Ireland, and Italy. By contrast, small fansites' linking structures most resembled those of users from the Ukraine, Greece, Belgium, New Zealand, and Sweden.

At a high level, these sites exhibited a strongly connected, multi-core network. The most frequently linked peripheries contained professional and commercial organizations, and the less frequently wikis and fansites. Northern European users' inlinks to the sites under study more resembled commercial organizations' inlinks, whereas continental and eastern European users' inlinks more resembled wikis and fansites.

### 4.3.12 Conclusion

Section 4.3.1 showed that Wikipedia's pages were the freshest overall, being updated quarterly, on average, whereas Wikia's were updated every two thirds of a year. Wikipedians updated pages in an ongoing manner, whereas Wikians updated pages in periodic frenzies, which were the result of university students taking semester breaks. Wikipedia pages were an average of 3.5 years old, meaning that most were created towards the end of Stargate SG-1 and the middle of Atlantis. Page creation on Wikia also followed a cycle, with users being more likely to make small changes over semester breaks and to take the time to create new pages towards the end of the summer. Page creation volumes on both sites peaked between 2005-2007, and have since been gradually decreasing, despite the release of DVD movies and the new Universe series.

Whereas the larger sites dominated the gathering of technical production, biographi-

cal/historical, and critical reception information, the smaller sites provided more of their own critical interpretations (§4.3.2). IMDb had many pages containing only biographical or historical information, and its cultural reference sections approached the kinds of user interpretation sections found on smaller sites. On Wikia, list-oriented information sections tended to be separated from more narrative information sections. Also, biographical and historical prose were more common on people-oriented pages, and interpretive sections more common on plot- or event-oriented pages. Finally, on Wikipedia, those pages that focused on the public (e.g., mass media) reception of something rarely provided interpretative critiques of their own.

On pages about individual titles (e.g., episodes and films), a triad of sites, including GateWorld and the wiki sites, consistently agreed on which core themes to include in their texts (§4.3.4). The wikis focused more on themes of travel and medicine than did the other sites. IMDb focused on the compilation of production technicalities, awards, and quotes, whereas the triad of sites more on lengthy textual explications of complex themes. GateWorld also had a particular focus on depictions of academia, explosion special effects, and esoteric production details. On pages about characters, there existed a spectrum of sites, with GateWorld being thematically closer to the wikis than to IMDb. GateWorld and Wikipedia agreed on including substantive aspects of characters' contexts, and the wikis agreed on including cursory/Infobox details. The editor-controlled sites differed with Wikipedia on how public/critical reception details should be presented, with the edited sites preferring lists of citations and Wikipedia preferring discussion paragraphs.

Editor-controlled sites were the least outward-looking, using references only to either support quotations and claims about something or to refer users to another editor-controlled

site, often an affiliate (§4.3.5). The wiki sites had greater variety and quantity of sources. Wikipedia was the only to cite academic literature, namely two Stargate-related works in physics and critical media studies. In general, the small vs. large website dichotomy between these sites paralleled the old academic emic vs. etic dichotomy, where small sites depicted and advertised a richly informative experience targeted towards fans, and large sites offered a high-level overview, based largely on information from production and marketing companies.

All sites provided release dates in some form, though only the wikis provided dates about editorial processes and production dates, and only small sites provided the release dates of non-episodes, such as books (§4.3.6). Links to official sites were most common, though mostly on the large sites. Most pages had at least one lengthy summary text, and that was the extent of most IMDb pages' long text fields. Most episode and character pages on the other three sites had production note texts. All sites possessed at least an interpretive level of original research. Additionally, all except GateWorld identified cultural references and compiled biographies/histories of the topic at hand. All except Wikipedia also collected production details, and only Wikipedia included textual discussions of topics' public critical reception.

The core content of this media phenomenon is probably episode and character pages. The omnipediac pages of all sites except IMDb also covered a core set of topics; pages on cultural references were unique to the wikis; and transcript, review, and making-of pages unique to GateWorld. Information about birth dates and places, names, spouses, key episodes, height, trivia, and past filmography were common to most cast, crew, character, and author pages on all sites except GateWorld. Title, writer, director, editor, and pro-

ducer fields occurred on book, cast, crew, and game pages across all of the sites. Vendor advertisements on the sites were common enough to suggest that the targeted users are technology savvy males in their twenties. A variety of corporate partnerships were also in evidence, on all sites except Wikipedia, which was funded only via donations and did not advertise.

IMDb and Wikia had many more advertisements than did GateWorld, on a similar number of pages (§4.3.7). Having a broader audience than Stargate fans, IMDb tried to maintain a more diverse portfolio of ads, whereas GateWorld and Wikia focused on a few key topics. All of the sites had some most common advertising theme. For GateWorld, this was Amazon and iTunes downloads, as well as IT ads that dominated all of the ad positions on pages where they appeared. IMDb focused on stable vs. more opportunistic business partnerships. Wikia focused on sci-fi and retail. Otherwise, a fairly standard set of vendor categories emerged.

User profiles confirmed that, in addition to the stereotype of nerdy young men, Wikians were more often British or otherwise European than American, and had a range of physical past-times in addition to science fiction. However, the stereotype of them being young (low twenties) was confirmed, as was the finding from §4.3.1 that most were students. Stargate-interested Wikipedians, by contrast, were more often located in the USA or Canada, were older on average (mid-to-upper twenties), and spanned a broad range of ages. They were also educated in a variety of fields, and were often professionally employed. Their hobbies and interests were often more aligned with those appearing in the Stargate shows, and they were often active participants in the Wikipedia community. A broad range of political philosophies was in evidence, and most were either agnostic, atheistic, or affiliated with one

of the common Christian denominations in the USA. Finally, the personal characteristics of most of these users aligned with the stereotype of being single, male, heterosexual, introverted-judgmental, and over-weight.

The analyses in §§4.3.8 and 4.3.9 revealed three profiles of user and editor rating behaviors on GateWorld and Wikia episode and omnipediac pages, which were found to be associated with various IQ factors. GateWorld's editors and a few GateWorld and Wikia users were "high-anticipation viewers," most preferring pages about season premiers, with many inlinks, links to supplementary material, and interrogative texts. However, most GateWorld users, and occasionally editors, were more "connoisseuring viewers," preferring normal episode pages with links to editorial reviews and lengthy editorial content. Most Wikia users followed a third profile, possibly indicative of "syndication viewers," where mid-season premiers and finales were preferred by most, and exciting season finales preferred by some. Most high ratings on Wikia's omnipediac character, place, and technology pages were also given on pages containing much trivia, tersely descriptive, and technical information. Hence, casual syndication viewers may have been most interested in quick summaries of topics of interest to them. However, highly rated omnipediac people pages were the opposite, with lengthy texts that included images and discussed those peoples' alliances.

Pages with high inlink counts or PageRanks typically had more substantive content, and those with low inlink counts or PageRanks more obscure content (§4.3.10). On episode pages with high PageRanks on both Wikipedia, Wikia, and GateWorld, users and editors had typically gone to the trouble of providing captions for images, full lists of cast members, and important plot questions. By comparison, less popular episodes provided more

trivial information, in which the general public/consumer would be less likely to be interested. On omnipediac pages, standard popular and substantive content in pages about individual people and technologies was the most linked-to, and pages about logistical context was less. Linking patterns on the most obscure pages suggested that GateWorld may have been considered by the general public to be the place to go for obscure information about the franchise.

In more subtle findings, Wikipedia was usually more textual and substantive, and Wikia usually contained more lists and esoterica. Wikipedia and IMDb contained what might be called substantive lists, which provided more than the biometric descriptions often found on Wikia. Wikia's lists were vaguely reminiscent of role-playing games, such as Dungeons and Dragons. Most highly PageRanked episode and omnipediac pages had many short words, suggesting that they were written by/for the general public, and lower PageRanked pages longer words, suggesting they were written more for connoisseuring fans. The highest PageRanked pages on Wikia and GateWorld followed a writing style more like that used on the Wikipedia pages preferred by more general inlinkers. Finally, although PageRanks were usually comparable to inlinks, there was evidence suggesting preferential attachment, manipulation of pages by their stakeholders, and favoritism among large sites.

Finally, regarding linking patterns, each of the sites showed evidence of strongly connected, multi-core networks of pages (§4.3.11). The closest peripheries to the cores typically contained professional and commercial organizations, and the farther peripheries wikis and fansites. Northern European users' inlinks to these sites more resembled the inlinks from commercial organizations, whereas continental and eastern European users' inlinks more resembled those from wikis and fansites.

## 4.4 Representational IQ: Completeness

### 4.4.1 Author's agendas and disclaimers

**RC1, RC2, & RC6:** To what extent does each site detail its ownership, sources of funding, and affiliations? To what extent does each site detail its purpose, primary interest, organizational type, and location? Finally, how do copyright statements and disclaimers differ across wiki and edited fansites?

These questions, like in §4.2.1, all require qualitative examination of a common set of webpages, or sections thereof, that were either few in number or were repeated verbatim across every page on each site. Such pages included: the copyright footer on every page, sites' History and About statements, policies about user contributions, privacy and advertising policies, and staff directories. The following sub-sections will address these questions, with respect to each site, to the degree possible from each site's documentation.

#### GateWorld

According to the footer on every page, "'Stargate' and all related characters and images are the property of MGM Television Entertainment." Following this was a statement that all content was copyrighted by "GateWorld LLC. All rights reserved." On the site's history page (GateWorld, 2009d), Darren Sumner was identified as the organization's founder, and a story was told of how the site progressed from a hobby of his, while studying for a bachelors degree in journalism at an unnamed university in Chicago, to an organization staffed by volunteers that has endeared itself with the studio. "From its humble beginnings," the page said, "GateWorld has grown into one of the Web's largest and most popular Stargate

sites, with hundreds of thousands of visits every month. Even the cast and producers of Stargate visit the site from time to time, and some have contributed through interviews, live chat events, and original articles. In 2004, MGM [the studio] sent Darren to Vancouver, B.C. to tour the sets and interview members of the Stargate Atlantis cast!" Nevertheless, the page also claimed that "The site remains independently owned and operated, thanks to generous support from fans." Regarding the site's primary interest, the history page also said the following: "While the episode guide remains the core of the site, we have expanded to news, an encyclopedia of the Stargate universe, licensed products like books and comics, interviews, editorials, episode transcripts and photo galleries, fan fiction, and much more."

On the staff page (GateWorld, 2009e), Mr. Sumner was identified as the "Owner and Managing Editor." A staff directory was also given, listing two co-/assistant editors, one forum manager, who also edit the fan fiction section, one server administrator, one graphic designer, four writers, and seven forum moderators. The advertisement policy page (GateWorld, 2009a) claimed 1.2 million visitors come to the site each month (n.b., this is considerably larger than the figure on the history page), identified advertisements as a source of funding for the site, and directed interested marketing campaigners to the Gorilla Nation online advertising sales representation firm.

Regarding the site's interest, the news contribution page (GateWorld, 2009b) also identified three categories of fan-made content that the editors were interested in publishing, namely: news, features, and opinions. News was defined as a 200-500 word article on a topic somehow "worthy of its own stand-alone story." Suggested topics were quotes from interviews, the latest novels or games, conventions, or an event to which GateWorld's editors could not devote enough manpower. Features were defined as 500-1,000 word articles

that contextualize an ongoing event. For example, "When Atlantis [the series] came to an end we ran a 5-part series on the top 5 episodes for each main character. ... It doesnt need to be breaking news to be worth publishing...." Finally, opinions were defined as 800-1,500 word articles that provide commentary on the production choices of the franchise. As the page said, "This is going to be the hardest category in which to get something published, since it shows up on the GateWorld home page and appears (even though it has your name on it) to represent the opinions of our site. So for the time being, make your argument balanced and level-headed and stay away from anything terribly controversial."

Only one page was given to describing the disclaimers and privacy policies of the site (GateWorld, 2009c). The page began with a general statement of organization purpose, being a "news and entertainment site devoted to exploring and enhancing the science fiction fandom experience." After assuring the reader that "All personal information you submit will be held in strict confidence," the page listed a number of affiliations, and data sharing policies, with other organizations, including: email hosting services via Everyone.net, merchandise purchases through Yahoo! Stores, unnamed contest and sweepstakes companies, DoubleClick ads, unnamed third-party targeted ad campaigns, and law enforcement agencies. While assuring the reader that GateWorld only intentionally passes purchase transaction records to the companies involved in completing the transaction, and does not store more on the user's computer than a login and state-keeping cookie for the GateWorld forums, it acknowledged that DoubleClick and third party vendors may include additional cookies and tracking technologies in their ads, which the site cannot control. Finally, as is true of all the sites under study, users under 13 were prohibited from posting content to any part of the site, without their parents' permission.

**IMDb**

For all but the most determined users, IMDb left its ownership vague. The words "An amazon.com company," on the footer of every page, were the only indication of ownership on most pages. In addition to listing the countries of several of their international subsidiaries in the footer of the homepage (i.e., Germany, Italy, Spain, France, and Portugal), the only indication of their location was on the copyright page (IMDb, 2009d), where a P.O. Box in Seattle, Washington, USA was given. To find their own account of themselves, one had to search for the knowledge base article entitled "What is the Internet Movie Database" IMDb (2009l), and click a parenthetical link, "see history." On that page (IMDb, 2010b), following a welcoming letter from the founder and CEO, Col Needham, a lengthy account was given of IMDb's early pages as a textual list distributed via USENET newsgroups from 1990; its hosting on early Web servers at Cardiff University around 1993; the founders' decision to incorporate in 1996, in order to seek funding via advertising, licensing, and partnership deals; and their acquisition by Amazon.com in 1998. The history page also linked to a page (IMDb, 2010a) listing annual messages sent to the site's top contributors, which were a kind of informal annual report, giving usage statistics, the year's product improvements, and plans for the future.

Despite one Help page claiming that "we just love movies" (IMDb, 2009l), many other pages on the site made clear that, as a subsidiary of a large corporation, their arguably primary interest was in making money, preferably via larger contracts than most individual consumers would spend. On IMDb (2009l), the following was claimed: "We are some of our site's most hardcore users. Our managing director claims to have seen over seven

thousand movies. Most of our people could write or win a movie trivia game show. ...we're just a bunch of hardcore movie fans who still can't get over the fact that we're getting paid to keep improving this tool we use so much for our own pleasure." Nevertheless, a host of other pages (e.g., IMDb, 2009b, 2009c, 2009g, 2009k) made clear that the company was only interested in licensing contracts worth above $15,000; in advertising budgets over $10,000; that their content can only be re-presented for free when in both personal and non-commercial circumstances; and that only the limited data on the FTP site are allowed for non-commercial purposes. (See §3.2.3 for the degree to which this dissertation has used their data. The brief quotes from their pages in this section are understood to be Fair Use.)

Their copyright and disclaimer pages also contained the most restrictive terms, and vague affiliations, of any of the sites studied. On a Help page entitled "How/where you get your information? How accurate/reliable is it?" (IMDb, 2009f), only vague sources and affiliations – with "industry," "visitors like you," on-screen credits, press kits, official biographies, autobiographies, and interviews – are listed. They were also sure to include that they were not liable for anything posted on their site. Their copyright and privacy policy pages (IMDb, 2009d, 2009h) similarly, essentially, said that they own everything on the site that is not owned by one of their partners (i.e., partners' intellectual property is respected more than that of the users' who contributed content to their site), and that they may partner with whomever they choose. Furthermore, in addition to prohibiting screen scraping and commercial use without their permission, using HTML frames to include their site within pages on other sites, mentioning their name or trademarks in another site's HTML meta tags, and inlinks or comments that they find to be somehow offensive were prohibited.

Regarding privacy, all of the personal information fields required of users were listed, they used both cookies and server logs to track users, and they used email response requests to judge newsletter readership. As did GateWorld, though they claimed not to intentionally send users' personal information to third parties, they acknowledged that third party ads may track users. However, they did acknowledge intentionally share users' personal information with others under the following circumstances: with "affiliated businesses we do not control" (i.e., businesses with which they partner to provide a single service), with agents (i.e., "companies or individuals [employed] to perform functions on our behalf"), for making promotional offers to targeted users, during "business transfers" (e.g., mergers and acquisitions), with law enforcement agencies, and under special circumstances with the user's consent.

**Wikia**

As a corporation that, to some degree, respects "free culture" – a generalization of the free/libre and open source software movement – Wikia was an interesting mixture of corporate proprietary/secretive and open source philosophies. Generally, the company itself was relatively proprietary in its inner workings, and profited through advertising and investment capital, though all of the wiki content it hosted was free and open. On the About page (Wikia, 2009a), their stated organizational purpose and interest were to offer a "consumer publishing platform where millions of passionate fans come to discover, create and share a shocking abundance of information on thousands of topics." Jimmy Wales, also a Wikipedia founder, and Angela Beesley were identified as the founders, with Gil Penchina as CEO. Their headquarters were said to be in San Francisco, California, USA.

Similarly, their Advertising and Press pages (Wikia, 2009b, 2009i) identified by name those staff members in charge of various roles, and even included their personal email addresses, though no physical contact information was to be found on their pages. (A street address was given in their WHOIS records. Though not said on their pages, since their copyright attorney was listed as being in St. Petersburg, Florida, USA, where Wikipedia used to be, it should be safe to assume that they moved to San Francisco with Wikipedia, and may not yet have fully integrated with the local legal community.) The Hiring page (Wikia, 2009h) additionally gave a glimpse into their organizational structure, listing open positions for a community manager, marketing associate, Web developer, and gaming community intern. That page also revealed that they have an office in Poznań, Poland, with openings for a Web developer and project manager, both of whom were required to have near-native English ability.

Finally, in addition to a few investors (i.e., Bessemer Venture Partners and Amazon) and an acquisition (i.e., Grub, a distributed Web crawler) listed on the Press page, and though many of the older links on that page were broken, one link still successfully went to the "Spring 2009 Update" (Wikia, 2009k). That page provided an informal annual report, much like IMDb's annual messages to top contributors, giving: site descriptive statistics, the insight that the company should become profitable for the first time in 2009, and their accomplished and planned interface and infrastructural improvements.

Regarding disclaimers (Wikia, 2009g), like IMDb, they too claimed not to be liable for anything on their sites, though they courteously warned the user that content on the sites may be offensive, and they mandated that all content on all of their wikis use the free/open Creative Commons Attribution-Share Alike License 3.0 Unported (CC-BY-SA).

These disclaimers were from the company's central wiki; no more specific disclaimers were found on the Stargate wiki. The Licensing page (Wikia, 2009d) merely went over the terms of the CC-BY-SA license, though, unlike IMDb, also emphasized that all content is owned by the users who created/posted it, not by the company.

The Privacy page (Wikia, 2009j) listed user information they required (only a username, password, and birth date), noted that other optional fields (e.g., email address) may be made public or be used for targeted advertising, and noted that all wiki admin/creation details are public. The site used cookies and server logs to track users, and said that they were willing to share any user information they have with "our subsidiaries and affiliated companies, contractors, and vendors," as well as law enforcement and internal security. They were also willing to share "with third parties aggregated, non-personal information, such as the number of new user registrations over a specific time period or the number of users who edited a particular wiki."

Similarly, from the Terms of Use page (Wikia, 2009l), they were not willing to notify users of changes to the Terms; users were forbidden from posting content that interfered with the display or functioning of ads; they had a lawyer dedicated to prosecuting copyright infringements according to the Digital Millennium Copyright Act; they were not liable for any damages beyond what users actually paid for the service (i.e., usually nothing, or no more than $1,000); and they made no claims of the error-freeness, accuracy, reliability, or satisfactoriness of either their sites' content or service infrastructure.

**Wikipedia**

The pages cited in this section come mainly from the Wikimedia Foundation, of which Wikipedia is one of its projects. For example, when Wikipedia linked to its privacy policy, the link went to the Foundation's site.

Wikipedia (and the Foundation) attempts to be an entirely open source organization. As a Web-based non-profit organization – having both minimal staff, a high degree of distribution of labor among the general public, accountability to government and other powerful donors, and no profit motive – Wikipedia apparently made every effort to be as transparent as possible. Complete lists of advisory board (WikimediaFoundation, 2009a) and board of trustees (WikimediaFoundation, 2009b) members were available, usually giving biographies and photos of each person, their responsibilities, and often their personal contact information. Additionally, the complete bylaws of the Foundation were available (WikimediaFoundation, 2009c). A complete annual report (WikimediaFoundation, 2010) was also available, giving: the organization's mission statement, letters from the directors/chairs, their location (and the history of their recent move from Florida to California), an organizational chart, financial figures audited by a certified public accountant (CPA), the next year's annual plan and projected revenue, aggregated site and user descriptive statistics, summaries of services and social initiatives, a summary of their infrastructure, the work done by local chapters around the world, lists of notable benefactors, a staff directory, and a P.O. Box in San Francisco. Many of these sections also had their own pages (e.g., the Benefactors, Financial Statements, and Fundraising pages), providing more details than were in the annual report.

The interests and actions of both the committees and user population of Wikipedia were also documented at length. An FAQ (WikimediaFoundation, 2009e) provided high-level answers to questions of organizational purpose and interest, ownership, funding/spending, and Wikipedia's relation to Wikia. On the Current events, Press room, and Our projects pages (WikimediaFoundation, 2009d, 2009h, 2009i), lists of major new donations and initiatives were listed, and contact information was provided for the media in 25 different regions of the world. There also existed pages for each local chapter (WikimediaFoundation, 2009g). The Resolutions page (WikimediaFoundation, 2009l) gave a list of motions approved by the various boards and committees. The Staff page (WikimediaFoundation, 2009m) listed executive, technological, program-related, financial, administrative, legal, usability, and strategic planning staff, as well as gave both the Foundation and usability project's organizational charts. Finally, the Values page (WikimediaFoundation, 2009o) gave paragraph-length introductions to the Foundations' mission statement, free/open culture, attempts to have the highest infrastructural availability possible, independence from all other organizations by using only donations for funding, commitments to ethnic openness and diversity, striving for the greatest transparency possible, and considering the community to be the organization's greatest asset.

Regarding disclaimers, a General Disclaimer page (WikimediaFoundation, 2009f) stated that all content is owned by, and liable to, its creators. A special resolution also existed (WikimediaFoundation, 2009k), saying that all projects must either use free licenses or include a rationale, called an Exemption Doctrine Policy, for using non-free content, as well as to replace that content, if free content of equivalent educational value ever became available. Policy pages existed for conflicts of interest, gifts, licensing, non-discrimination,

user and staff privacy, donor privacy, access to non-public data, whistle-blowing, travel expenditures by staff, code of conduct, staff credit card usage, data retention, staff duty entertainment, and encouraging pluralism and international diversity.

The user and staff privacy policy (WikimediaFoundation, 2009j) had the following terms. All editors of pages must be publicly identified, by either a username or IP address, and their content automatically licensed under both the Creative Commons Attribution/Share-Alike License 3.0 Unported and the GNU Free Documentation License (GFDL; as a fall-back license). Server logs and users' IP addresses may only be used to combat malicious behavior, to provide site statistics to the general public in aggregate, to solve internal technical or security problems, to aid law enforcement, or with the user's special permission. Server logs may not be used to track or target users, and cookies may only be used for login sessions. Content may only be permanently deleted from a page's history by court order. Email address collection from users is optional, though its absence prohibits the resetting of the user's password, unless the user contacts the headquarters via another means. Users are emailed, if possible, about any subpoenas received requesting the release of their personal information. Users may quash subpoenas by having a lawyer submit a request to do so to the Foundation, though a court (in California) may over-rule such attempts. Last, the user is warned that the Foundation cannot protect users against data mining techniques, which might triangulate their identities from multiple sources. Finally, the site's Terms of Use (WikimediaFoundation, 2009n) offered few additional policies. After reiterating that all contributions are licensed under both CC-BY-SA and GFDL, it added that work imported from other source must comply with CC-BY-SA as well as cite its authors.

## Conclusion

### *Incorporation-divided*

Identification of ownership and organizational structure were divided along a line of incorporation. The non-profit (Wikipedia) and volunteer-run (GateWorld) sites both named their ownership and listed their organizational structures in considerable detail. Both corporate sites (IMDb and Wikia), however, provided only the names of the founders and top executives, leaving their organizational structure inferable only from job postings.

Similarly, regarding funding sources, both non-corporate sites were the most forthcoming, listing donors and major affiliations by name, whereas both corporate sites spoke only of affiliations and third parties in the abstract. Though several affiliated organizations were listed by GateWorld, the site's additional abstract discussion of third parties, as well as its obtaining funding from ads and merchandise sales, indicated a tendency towards corporate policies. Being entirely funded by donations, Wikipedia was the only site not to employ corporate funding practices.

The same spectrum of organizations also reflects protections afforded to users' personal information. Wikipedia, having no advertising or other business partners or subsidiaries, did not profile user types from server logs, required very little personal information from users, and only divulged what little they collected when required by law to do so. GateWorld also did not profile users from logs, only used cookies to track forum login sessions, and only intentionally shared the information necessary to complete a transaction with their business partners, though third party advertisements were not prohibited from collecting user browsing behaviors. Both IMDb and Wikia itemized, in the abstract, the ways in

which they may share user information with partners, subsidiaries, and others.

### *Ideology divided*

Licensing and copyright policies were divided more along ideological, or perhaps editorial, lines. GateWorld and IMDb both claimed copyright and ownership over all content, whether user- or editor-created, on their sites. The most credit given to user contributions was the inclusion of authors' names, or usernames, alongside contributions. Whereas IMDb made clear that this policy was at the core of their licensing-based business model, GateWorld's editors offered no rationale for adopting it. On the other hand, both wikis frequently invoked the phrase "free culture" – a generalization of the free/libre and open source software movement – mandating that all content be released under the Creative Commons Attribution/Share-Alike 3.0 Unported license, and that all content belongs to its creators. Both sites kept a full and public history of all revisions to their content, in order to identify ownership.

### *Common to all*

Common to all sites was a vagueness about physical location, probably due to staff safety and privacy concerns, though the large sites protected their locations more completely than did the small sites. For both IMDb and Wikipedia, on both their websites and WHOIS records, only a P.O. Box could be found in their respective cities. GateWorld and Wikia listed no physical contact information on their websites, though the street addresses of their headquarters could be found in their WHOIS records.

Finally, all of the sites offered considerable narratives about their purposes, intentions,

and histories. Wikipedia was the only site to explain in philosophical detail how its values related to its operations and fundraising, though IMDb's history page described the founders' decision to move the database from being a USENET community project to a corporation as a difficult one, but to which they claim not many in that community objected. Those sites that obtained funding through advertising also did not mention advertising within their statements of passion for their mission. Advertising was also described on separate pages, appearing to be more of a means to an end than a desired activity. Last, the corporate sites took greater pains to obscure their history and past performance pages within their sites, in favor of presenting a simple and unified branding image to all but the most investigative users.

### 4.4.2 Reasons for citing similars

**RC3:** Why did the fansites link to other fansites?

A similar question was answered in §4.3.11, namely to what other types of fansites did external links on these sites go? However, whereas that question is one of characterizing link destinations, this section's question is one of characterizing link origins, namely the contexts in which the sites under study linked to other other sites. For example, GateWorld character pages often linked to IMDb actor pages, but to answer this research question is to also know that such links usually occurred when character pages named the actor who usually portrayed the character.

For this purpose, the pages on which the external links presented in §4.3.11 occurred were manually visited and examined. This section will present, for each website sub-

section, the usual reasons for linking to other fansites. Note that links to large retailers (e.g., Amazon and iTunes) and social networking sites (e.g., Facebook and Youtube) were not considered, because such links always occurred merely for promotional purposes, such as to sell the item featured on the current page or to promote the presence of an account (e.g., GateWorld) on that social networking site.

**GateWorld**

The only fansites to which book pages linked were authors' and actors' official sites. The former occurred on book title pages, providing a link to the first chapter of the book on the author's site. The one actor site link went to Kate Ridder's site devoted to executive producer and actor Richard Dean Anderson, and was located at the end of a page describing a guidebook she wrote about SG-1, as a way of providing more information about that book.

Comic pages contained no links to fansites.

Episode pages' fansite links most often went to IMDb name pages. When the writers, directors, guest stars, etc. were listed near the top of episode guides, that text often linked to those people's IMDb name pages. Links to the producer's weblog were next most common, and usually occurred following a quotation about production details, often appearing towards the bottom of episode pages. Links to interviews with the cast and crew, as well as to press release pages on the network's site, had a similar place, usually following quotations about news or production details for episodes. Three pages contained links to behind-the-scenes photo galleries, from the production section of episode pages. Links to encyclopedic sources, such as the Encyclopedia Mythica and Wikipedia, occurred in notes sections, when discussion was given of cultural or scientific concepts appearing in

the episode. Finally, the pages of highly anticipated episodes often possessed news headlines, usually to trailer, teaser, and behind-the-scenes videos on the studio and network's sites.

Omnipediac pages's links also most often went to IMDb name pages, for the reason noted in this section's introduction. If such links did not go to IMDb, or sometimes in addition to IMDb, they went to the actor's official homepage. Also common were links to the Encyclopedia Mythica, because a common trope in Stargate is to explain mythological figures on Earth by attributing them to powerful aliens. For example, at the end of the character page on Baal, who is the leader of one faction of a powerful enemy race in SG-1, is a link to the Encyclopedia Mythica page on Baal, the Canaanite fertility deity.

Only two external links went from video game pages to other fansites. The first was a "related link" at the end of the Stargate Worlds page to a gaming community site. The second was a link at the end of the SG-1 Mobile game page to its developer's site.

**IMDb**

As explained in §4.3.11, IMDb only included its own affiliates/advertisers' links on the main page for each title and character (and cast/crew), and relegated links that users would probably find more meaningful to sub-pages of those main pages. No external links went to fansites on character sub-pages, only nine links were found on title sub-pages. Although some of those links went to fansites, they were presented on IMDb in simple bulleted lists, without any reasons or contextual evidence given for why those particular links were chosen by users or editors. Hence, this research question cannot be answered for those links.

### Wikia

Actor pages on Wikia most often went to the Wikipedia home and actor pages, as well as to IMDb name pages. Both the Wikipedia homepage and specific actor's page often occurred, because both the author's name and the word "Wikipedia" would be links. This appeared to be a default behavior for creating actor pages on Wikia, and was not often explained in the pages' texts. As on the GateWorld omnipedia, official page links were also commonly included without explanation, as were TV.com profiles. The actor's largest unofficial fansites were also rarely included. Next most common were links to the Battlestar Galactica Wikia wiki, because a number of actors have worked for both franchises. Finally, three links to interviews were used to corroborate facts about a lead actress, as was a GateWorld news page about a lead actor, as well as likely to satisfy personal intrigue about actors' personalities.

On book pages, links to authors' personal, as well as publishers', sites were the default. Two pages linked to short stories created only for publication on a role playing game developer's website. Two pages also linked to an author's site featuring a book by an SG-1 concept artist, which was never published, due to royalty fees. These links were the only way for fans to obtain these publications. The links to Wikipedia were to provided as supplements to Wikia pages on the same topic, and often occurred alongside other fansites that contained pertinent information to the topic. The links to the Star Trek wikia were for an author who had worked for both franchises. The Star Trek Wikia wiki's page for that person was largely identical to the Stargate Wikia's page, and both pages linked to each other.

Of crew pages' three external links, two went to the IMDb name pages of popular crew members, and one to the Battlestar Galactica wiki, because, like the actors, that crew member had worked for both franchises.

On episode pages, the apparent fansites hosted on AOL, which presented a simple front page for obtaining Stargate-related email and instant messaging addresses, were referred to by the pages linking to them as being the studio's official Atlantis and SG-1 sites. Since both of those series had ended by the time of data collection, it seems that the studio replaced their official pages with this generic placeholder page/service. The producer's weblog was linked-to from episode pages in order to direct users to insights into the production process of the series' following episodes or season. Links to GateWorld's news pages served a similar purpose, and links to both GateWorld's and the network's episode summary pages occurred within a list of cross-references to other fansites offering coverage of the same episode. Links to transcripts and screenplays on Dave.tv were also part of this standard block of cross-reference links, as were links to the SG-1 Solutions wiki and MGM site, which provided transcripts for some episodes.

Game pages most often linked to developers' sites and small knowledge base-oriented fansites. Like links on actor pages to IMDb or official sites, developers' sites functioned as a contextless default for most titles. Small fansites served as repositories of trivia, techniques, and discussion around a particular game. Links to developer interviews, as with actor pages, served to corroborate points made in the article, and to allow fans a window into developers' states of mind. Links to purchase games were quite rare, occurring only for less mainstream types of games, such as for mobile phones, and were nearly the only content on pages about those games, suggesting they were possibly placed there by their

developer.

Omnipediac pages most often went to Wikipedia, and almost entirely to pages about science and engineering topics that extend beyond Stargate. On such pages, which were about the same topic as the destination Wikipedia page, a brief introduction to the topic's role in Stargate was given and followed by only the one link to Wikipedia, apparently for those interested in the topic beyond its use in Stargate. Next most common were links to the GateWorld omnipedia, which served as cross-references to others' coverage of the same topic, which was also true for links to character pages on both the SG-1 Solutions wiki and network's site. Kate Ritter's site was consulted for previews of news and episodes as well as for the brief episode guides it offered. Links to IMDb cast and crew pages were used in the same way as on GateWorld's character pages, though much less frequently. The few links to small fansites were to actor interviews, intended to yield insights into a character.

Finally, video pages contained seven external links. The two to the DVD collection's official website came from pages about the collection and its publisher, and were offered only as a supplement to the page's content, not as a reference. Links to GateWorld news pages occurred on pages about the Stargate films, providing a source for rumors that the original film's director, who was not involved in the television adaptation of the franchise and who has publicly voiced disapproval of that adaptation, wished to complete what was originally intended to be a trilogy of films. This was rumored in 2006, and has yet to happen. The link to the network's site was a generic reference without much meaning. And, the link to the partwork publisher was because that publisher produced Stargate's DVD collections.

**Wikipedia**

As on previous sites, actor pages on Wikipedia most often linked to IMDb name pages, actors' official sites, and person-profile pages on indexing sites – including TV.com, All Movie Guide, and TV Guide – as well as to small fansites devoted to individual actors. As on Wikia, these links were part of a standard block of cross-reference links, at the end of most pages, to other sites offering information on an actor. Links to GateWorld news pages, as well as to interviews, were used as references on these pages, usually to information about an actor's future involvement with the franchise. As on Wikia, links to the Wikia wikis of related franchises occurred in the cases where an actor worked for multiple franchises.

Author pages similarly linked to authors' official sites, authors' companies' sites, small devotional fansites, or index sites (i.e., IMDb and ISFDB) by default. Links to interviews were either references for points made in the text, or merely part of the generic default links. The Star Trek Wikia wiki was referenced in the usual way, for authors involved in both franchises.

Crew pages' standard links included IMDb name pages, affiliated studios and production companies, personal profile/index sites (TV.com, ISFDB, and All Movie Guide), crew members' official sites, and small devotional fansites. As on author pages, interview links could either act as references for points made in the text, or as generic links. Links to the Star Trek Wikia wiki were as usual.

Episode pages' standard links were to the studio and network's summary pages, IMDb title pages, screenplays at Dave.tv, and occasionally Wikia and GateWorld season and

episode pages, as well as episode profiles on index sites (TV.com, Rotten Tomatoes, and All Movie Guide). The especial linking to the studio, network, and IMDb might indicate a bias towards other large organizations. Links to Wikiquote pages, which provided a list of key or humorous quotations from an episode, were located in standalone boxes containing the text "Wikiquote has a collection of quotations related to: [episode name]." The format of the list of quotes closely resembled quotation lists on IMDb, possibly indicating stylistic copying. Also like IMDb, these standalone pages existed for most episodes. Links to box office data site The Numbers occurred on pages about the two SG-1 films, as references supporting gross revenue figures given in the text for the films. Finally, one link to an SG-1 Solutions page containing plot spoilers about the Ark of Truth SG-1 film was given as a generic link at the end of the page about the film, having probably been placed there before the film's release.

Finally, the standard links on omnipediac pages, which were primarily about characters and races, went to Stargate Wikia wiki pages on the same topic, network and studio character profiles, as well as race profiles on the GateWorld omnipedia and, rarely, SG-1 Solutions and small fansites. The link to Wikiquote was for translations of humorous quotes by a character who frequently curses his boss's arrogance in the Czech language on Stargate Atlantis. The TV Tropes and TV.com links were generic links to index sites for a popular character and race. Links to IMDb title pages were because they contained interesting awards or trivia pertaining to a character.

The Gospel of Matthew translated into Alteran on Archive.org link was a generic external link on the page about the Alterans (i.e., Ancients), probably merely as a curiosity, as the Gospel of Matthew plays no obvious role in Stargate. The destination page describes

the text in this way: "This newly discovered text is a guide to achieving the required spirituality in order to attain ascension [i.e., enlightenment]." The fact that the translated text belongs to an Abrahamic religion that seeks heaven (i.e., Christianity) rather than enlightenment (e.g., the Dharmic religions of South Asia) is interesting. Perhaps the text's author was a Christian fan of the franchise who wanted to proselytize, or at least defend Christianity, to this community. Stargate's themes often pit science and dogmatic religions against one another, portraying scientists as modern, independent, and rational, whereas followers of dogmatic religions are portrayed as antiquated, subjugated, and fear-mongering. On the other hand, science and enlightenment-oriented religions are often portrayed by the franchise as working towards the same goal, namely overcoming mortal existence via personal effort, in order to live immortally in a suffering-free, powerful, knowledgeable, etc. state.

**Conclusion**

Nearly all pages contained a group of external links, usually isolated and located at either the top or bottom of the page, for the purpose of recognizing people and companies affiliated with the page's topic and/or cross-referencing or directing the user to pages on other sites that covered the same topic. Occasionally, and especially when the current site's coverage of a topic was shallow, such links also provided a supplementary function, directing the user to sites offering greater coverage of the topic. Last, this isolated group of links often contained one or two links to a small, devotional site, usually maintained by one obsessive fan.

Actor, author, and crew pages on the wiki sites often contained links for the purpose of identifying people who had worked in multiple franchises. Such links often went to another

franchise's wiki, which, in turn, linked back to the same Stargate wiki. Also on these page types, as well as on episode and game pages, links to interviews usually served as references supporting arguments or quotes made in the page's text. However, less commonly, interview links could also merely supplement the page somehow, such as by providing intrigue or insight into a crew member or developer's state of mind.

Book and author pages often linked to sample chapters, or to book-like content that had been written by professional authors, but had not been published in print for some reason.

Episode pages often linked to sites providing previews of upcoming episodes, as well as behind-the-scenes footage. Both episode and omnipediac pages also occasionally provided links to the real-world version of a topic appearing in Stargate, such as information about the real ancient history of mythological characters that the franchise has attributed to aliens.

Game pages often linked to fan communities or knowledge bases pertaining to specific games, for the purpose of letting gamers socialize and collaborate around playing the game.

Finally, one instance of Christian proselytizing, or at least defense-taking, was found on a Wikipedia omnipediac page. The franchise's themes often pit science against dogmatic religions, and imply that religions that seek enlightenment (e.g., Dharmic religions) have the same eventual goal as does science. One fan translated a Christian scripture in such a way as to suggest that studying it is a viable means of attaining enlightenment.

### 4.4.3 Length and mass collaboration: thoroughness

**RC4:** To what extent does page length or number of authors correlate with other markers of quality?

Length of page texts was measured in terms of word counts, which were available for the all pages sample (described in §3.3), having been automatically counted by the POSIX `style` utility. Numbers of unique authors were available only for the wiki sites, because only they provided public records of page revisions and those revisions' authors. Also, because it was easily obtainable during the data collection process, the total number of revisions for each page by any author were counted and included as a predictor variable in the analyses, when including that variable did not violate multicollinearity assumptions.

All available revision records for all Stargate-related pages of both wikis were requested in XML format from the wikis' APIs. Only the username field of each revision record was requested, because only that field was required for answering the question, and the maximum number of results allowed by each site for each page (i.e., 500) was requested. The POSIX `wget` utility was used to download the records, was set to infinitely retry each URL until the download succeeded, and was limited to rates of one request every 30 seconds, as well as to download rates of 20Kbps, which are more considerate settings than either the sites' robots.txt files or API documentation required. Other standard POSIX utilities (e.g., `grep`, `sed`, and `sort`) were used parsing each XML results file, and for counting numbers of unique usernames and IP addresses. As with the other multiple regression analyses in this dissertation, iteratively re-weighted least squares (IWLS) was employed (via the `R` `rlm` function) for robustness.

The following sub-sections present both descriptive and modeling results, using either page length or number of authors as the dependent variable, followed by a conclusion.

## Page length

### *Descriptive statistics*

Descriptive results for the page length variable across every site sub-section are presented in table 4.37. The row and column fields have been transposed from the order usually used throughout this dissertation, in order to include all of the site sub-sections in a single table. The abbreviation "N.E.O." is used for cases when Not Enough Observations were available to calculate a statistic. Though the samples were exhaustive, one should always be wary of findings with particularly small sample sizes. Minimum values of zero reflect pages lacking any blocks of descriptive/interpretive text, though such pages could have included lists or tables of links, images, or the like. Textual lists and tables were captured in this analysis.

Table 4.37: Descriptive statistics: the page length variable, for each site sub-section

| Sub-section | St. Dev. | Min. | 1st Q. | Median | Mean | 3rd Q. | Max. | Skew | Kurt. | n |
|---|---|---|---|---|---|---|---|---|---|---|
| **GateWorld** | | | | | | | | | | |
| book | 315.57 | 0 | 111.25 | 142.5 | 226.3 | 166 | 1,479 | 3.07 | 8.66 | 66 |
| comic | 184.85 | 0 | 25.5 | 64 | 168.22 | 323.5 | 546 | 0.86 | -0.89 | 27 |
| episode | 916.79 | 0 | 222.5 | 1,049 | 1,083.24 | 1,604 | 4,660 | 0.92 | 1.07 | 360 |
| omni. | 118.06 | 1 | 31 | 68 | 99.1 | 131 | 1,380 | 4.31 | 33.12 | 2,562 |
| video game | 144.92 | 19 | 34.25 | 87.5 | 145.8 | 205.5 | 414 | 1.06 | -0.19 | 10 |
| **IMDb** | | | | | | | | | | |
| actor | 143.48 | 0 | 0 | 0 | 58.56 | 40.75 | 1,607 | 4.44 | 30.43 | 612 |
| char. | 179.28 | 31 | 80 | 184 | 223.33 | 310 | 713 | 1.31 | 1.87 | 21 |
| crew | 209.76 | 0 | 0 | 0 | 56.26 | 0 | 1,013 | 4.19 | 18.18 | 27 |
| title | 95.21 | 0 | 0 | 82 | 97.15 | 168 | 392 | 0.71 | -0.38 | 369 |

Continued on Next Page…

Table4.37 – Continued

| Sub-section | St. Dev. | Min. | 1st Q. | Median | Mean | 3rd Q. | Max. | Skew | Kurt. | n |
|---|---|---|---|---|---|---|---|---|---|---|
| **Wikia** | | | | | | | | | | |
| actor | 34.32 | 0 | 0 | 7 | 14.67 | 13 | 373 | 6.47 | 56.07 | 373 |
| book | 81.72 | 0 | 0 | 0 | 50.05 | 63 | 502 | 2.08 | 4.53 | 325 |
| crew | 54.2 | 0 | 0 | 0 | 18.1 | 3.5 | 379 | 5.37 | 34.08 | 60 |
| episode | 666.54 | 2 | 363.5 | 624 | 789.5 | 932 | 3,601 | 1.39 | 1.51 | 371 |
| game | 144.37 | 16 | 31 | 96 | 132.13 | 191 | 567 | 2.13 | 5.56 | 15 |
| omni. | 192.15 | 0 | 119.5 | 145 | 199.93 | 202 | 809 | 2.53 | 7.64 | 2,975 |
| video | 42.1 | 0 | 17 | 20 | 30.12 | 22 | 201 | 3.12 | 11.51 | 25 |
| **Wikipedia** | | | | | | | | | | |
| actor | 221.4 | 0 | 84 | 151 | 219.51 | 282 | 1,971 | 2.92 | 14.07 | 371 |
| author | 65.65 | 12 | 42.25 | 62.5 | 82.6 | 98 | 275 | 1.88 | 3.73 | 20 |
| crew | 206.84 | 3 | 70 | 130 | 203.23 | 278 | 1,015 | 1.84 | 3.75 | 53 |
| episode | 602.53 | 0 | 571 | 1,033 | 1,057.11 | 1,477 | 2,443 | 0.39 | -0.47 | 47 |
| game | 1,258.47 | 1,205 | 1,491.5 | 1,778 | 2,199 | 2,696 | 3,614 | 1.34 | N.E.O.[3] | 3 |
| general | 4,696.08 | 1,641 | 4,112.75 | 7,987 | 7,519.38 | 8,999.25 | 16,752 | 0.89 | 1.38 | 8 |
| list | 3,293.07 | 0 | 0 | 0 | 1,415.46 | 62.25 | 14,591 | 2.68 | 6.75 | 56 |
| omni. | 2,867.86 | 74 | 548.25 | 1,047.5 | 2,006.17 | 1,778 | 12,270 | 2.54 | 6.04 | 36 |

For GateWorld, these descriptive statistics say that episode page texts were both the longest on average and had the widest variance, whereas omnipedia pages were the shortest and had the narrowest variance. Video game and omnipedia pages never lacked text, and some episode pages were especially long. All of the distributions were skewed to the left,

[3] As with several places in this dissertation, the abbreviation "N.E.O." is used for cases when Not Enough Observations were available to calculate a statistic. Though the samples were exhaustive, one should always be wary of findings with particularly small sample sizes.

indicating that most page lengths were short, and this was especially true for omnipedia and book pages. Those page types also had high peaks about their most typical values. These findings suggest that episode pages were produced by the site's editors in a much less structured/manufactured manner than were omnipedia pages.

IMDb text lengths – calculated from the free FTP dataset of all name and title pages, and from the matched pages sample of character pages – were longest on average on character pages, though this could be due to the small and non-random matching pages sample. Name and character pages showed the most variance in length, and title pages the least. A few name pages were especially long, though most were relatively short. Character and title pages were mildly left skewed and not very peaked. From these results, name pages appear less routinely structured than do title pages, possibly suggesting more fan contributions to name pages than to title pages. This focus on name over title pages was unique to IMDb. Coming from the matched pages sample, these character pages only represent popular characters, which should be the longest pages in the population of character pages. Considerable variance existed among these popular character pages, suggesting little standardization. This could be due to character pages being a relatively new and peripheral type of content on the site.

On Wikia, episode, omnipedia, and game page texts were the longest on average, and also had the most variance. Actor and crew page texts were the shortest. Like on GateWorld, no game pages were without text, though there were few game pages. Some episode pages were particularly long. All of the page types were left skewed, with actor and crew pages strongly so and highly peaked. These results are most comparable to GateWorld's, though Wikians wrote more about omnipediac and game topics. Unlike IMDb, they gave

only token attention to actors and crew.

Finally, Wikipedia page texts were longest on general, game, and omnipediac pages, and shortest on author pages. This same list reflects the greatest and least variance, if game pages are replaced with list pages. Only actor, episode, and list pages lacked any text. Also, list pages were usually long. All page types were left skewed, with actor, list, and omnipediac pages being the most so. Actor pages were also highly peaked, and list and omnipediac pages moderately so. The focus on Wikipedia pages was clearly towards summary pages, followed by game and omnipediac pages. It is as though Wikia were repackaged for a general public audience. As the purpose of list pages is to accumulate long lists on some topic, those pages were predictably long. Also like Wikia, actor, list, and omnipediac pages seemed to be produced somewhat perfunctorily.

***Modeling results: Iteratively Weighted Least Squares***

As with previous regression results (cf. §4.3.9), because of the large number of models involved, significant coefficients from each website section's model have been grouped across models according to their type (i.e., content-, word usage-, link-, or authorship-related), and sorted according to their coefficients' magnitudes and signs. In addition to the effects discussed in this section, a large number (394) of interaction effects were also found, some as complex as six-way, and usually between the word usage variables. Interactions involving these linguistic variables will not be presented, because: this project's purpose is to identify initial exploratory IQ factors for fansites and not to do an in-depth linguistic study of their texts, interpreting so many interactions would take considerable time and space, and the study of word usage variables is not mandated by the IQ literature. However,

interactions involving the other variable types will be interpreted.

First, regarding **content-related variables**, JPG images were significantly more common on GateWorld episode pages possessing long texts, especially when many conjunctions and links to transcripts were also present. Consistent with §4.2.2, this is probably referring to popular or highly anticipated episode pages, which had been well-documented. However, JPGs were less common on Wikipedia episode and Wikia crew pages possessing long texts, perhaps because copyright-free images of such popular titles and people were not available, or because lossless PNGs, especially of cast and crew members attending ComicCon, were a common format for photographs of popular people on the wikis. Also, PDF images were more frequent on Wikia episode pages, especially when in combination with many prepositions, regular and subordinate conjunctions, total links, and pronouns. This too is consistent with the finding in §4.2.2 that popular wiki episode pages often link to transcript or screenplay PDFs, and with the mass agglomerative writing style discussed in §4.2.5.

Of cast and crew-related variables, crew pages with lengthy texts on Wikipedia often included information about a crew member's occupation, birthplace, titles produced, and children, but tended to have fewer images, birth date details, and links to the person's website. Apparently, the users who write the longest crew articles on Wikipedia were interested in documenting the personal histories of crew members. An actor's height was similarly related to long text on IMDb, as were lists of pictorials, trademarks, interviews, theatrical productions, filmographies as a non-actor (e.g., as a writer or director), and TV commercials about/featuring the actor. The height field was located on main actor pages, whereas the other fields were consolidated on a common sub-page for each actor. Hence,

the height effect probably indicates that more popular/accomplished actors also were more likely to have their physical features documented by users, whereas the other effects occurred together on a page that was intended to be a long list of the actor's accomplishments. On Wikia actor pages, having a "notable roles" (outside of Stargate) field was related to longer texts, though listing many Stargate and theatrical roles were counter-indicated. As in §4.3.11, this suggests that actor pages on Wikia acted as gateways to other similar sites.

Textual sections of pages that were often present on long pages included: author summaries on Wikia book pages, biographies on Wikia crew and actor pages, and career descriptions on Wikipedia crew pages. These are the sort of fields that one would expect on well-documented pages of these types. On the other hand, when summaries of the present state of affairs existed on Wikia omnipediac pages, or when plot summaries that included many short sentences with many to-be verbs existed on Wikia episode pages, those pages tended to have shorter texts. Also, omnipediac pages on Wikia about ships had longer texts when ships' sensors and crew capacities were discussed, and shorter when navigation and propulsion were discussed. Apparently, when Wikia users wanted to provide only a brief summary of an encyclopedic topic, they kept to descriptions of the current state of affairs, only brief and terse plot/history summaries, and the perfunctory features of technologies.

Regarding **word usage variables**, both character and sentence counts were always positively associated with length measured in terms of word counts, affirming that, at least for these sites, text length measured in any of these ways should yield commensurate results. More detail on this will be given in the readability metric results, three paragraphs below. Especially many to-be and auxiliary verbs were present on lengthy GateWorld book and omnipediac pages, though fewer on Wikia actor pages. This is consistent with the reviewer

style from §4.2.5. Conjunctions and pronouns were common to all of these page types, as well as to GateWorld episode and both wikis' crew pages. Prepositions were common all of these page types, except for Wikia crew pages. This suggests the mass agglomeration and general documentation styles.

Sentences were rarely short and often long on lengthy pages on any site. Lengthy crew pages' sentences often began with pronouns, suggesting the reviewing style. Long Wikia book and Wikipedia crew pages also usually contained many interrogative sentences, also suggesting the reviewing style, whereas Wikia omnipediac and GateWorld book pages avoided them, preferring the book/author style. Passive sentences were common on long Wikia actor and crew pages, though not omnipediac pages, implying that the mass agglomeration style dominated the former more than the latter. Wikia actor pages' sentences also often began with articles, and not prepositions, suggesting either the reviewing or possibly general documentation style. Finally, sentences often began with subordinate conjunctions on lengthy Wikia book and Wikipedia crew pages, suggesting the reviewing style.

The following readability measures were associated with lengthy texts on the following pages: Coleman-Liau on Wikia actor pages; Flesch on GateWorld omnipedia, episode, book, and Wikipedia actor pages; Fog on Wikipedia actor pages; Fry on Wikia omnipediac and actor pages; and Kincaid on Wikia crew pages. The following measures were disassociated with lengthy pages: ARI on Wikia episode, crew, actor, and Wikipedia episode pages; Fry on GateWorld episode, and Wikipedia actor pages; Kincaid on Wikia actor pages; Lix on Wikia omnipediac pages; and SMOG on Wikipedia actor pages.

Given that Kincaid, Flesch, and Fog all prioritize having many sentences and syllables over many words, the fact that some page types both scored highly on these measures

and contained many words suggests that they (i.e., GateWorld omnipedia, episode, book; Wikipedia actor; and Wikia crew pages) would be considered long by most measures of length. Wikia actor pages' low scores on these metrics indicates that they were mostly long in terms of word counts. Fry and SMOG measures prioritize only syllables, so Wikia omnipedia and actor pages were fairly long in terms of syllable counts, and GateWorld episode and Wikipedia actor pages were less so. Because some of these sections scored highly on the previous group of readability metrics (i.e., Kincaid, Flesch, and Fog), this also suggests that sentence length was the primary reason for the previous group's high scores, not syllabic length. ARI and Lix prioritize texts with many sentences and characters. Hence, it makes sense that the sites that are contra-indicated on these measures (i.e., Wikia actor and omnipediac pages) had positive effects on Fry and SMOG, which are purely syllabic. Wikia crew pages' scoring lowly on these measures and highly on Kincaid also suggests that those pages had great syllabic lengths. Finally, Coleman-Liau prioritizes only character length, so Wikia actor pages had both many words and characters.

Regarding **link variables**, lengthy GateWorld episode pages often had many links, especially when many prepositions, nominalizations, and average-length paragraphs, as well as high Flesch and Fry readability scores, were present – that is, when the text was written in the book/author style. Such pages usually did not have many links to outside sites or transcripts, unlike the wikis. By comparison, lengthy Wikia episode pages often included many links, prepositions, pronouns, and transcript PDFs or links thereto, suggesting their writing was closer to the reviewer style.

In terms of PageRanks, lengthy Wikia omnipediac pages often had both high Fry and high PageRank scores, possibly written in either the general documentation or reviewer

styles. Lengthy IMDb title pages also often had high PageRanks, especially when many sentences of average length were present, suggesting the reviewer style. Lengthy GateWorld omnipediac pages often had both many broken links and high PageRanks. Finally, lengthy Wikia episode pages often had both high SMOG and high PageRank scores, suggesting the book/author style.

Finally, regarding **authorship variables**, the relationship was only clear on Wikia crew pages, where longer pages simply had more revisions and fewer authors, possibly indicating obsession with crew members by a few fans. This is also supported by the finding, earlier in this section, that the longest Wikia crew pages were written about crew members' personal lives. Lengthy Wikia omnipediac pages also had many revisions, but only when many pronouns, nominalizations, and interrogative sentence beginnings were also present (i.e., the text was written in the general documentation style). But, if there were more prepositional and pronominal beginnings (i.e., the reviewing style), the opposite was true. Hence, lengthy general documentation omnipediac pages on Wikia were revised by many, but review-oriented pages probably only had a few revisions (i.e., substantive reviews) posted to them. Author counts had a similar relationship with lengthy Wikia omnipediac pages. Whereas long pages possessed many authors as well as sentences beginning with pronouns and interrogative pronouns (i.e., general documentation style), if there were also many prepositions (i.e., reviewing style), the text would probably not be so long. However, on lengthy Wikia episode pages, many revisions, total links, pronouns, conjunctions, and sentences beginning with subordinate conjunctions also often occurred, suggesting the reviewing style. No other reviewing styles were evident on Wikia episode pages.

## Number of authors

### *Descriptive statistics*

Table 4.38 displays descriptive statistics for both the numbers of authors and revision variables, for each site and page type. Though both variables are described, recall that only the numbers of authors variable was used as a dependent variable in the regression modeling presented in the second half of this sub-section. Zero minimum values for these variables represent pages that have been administratively created, but not yet edited, such as pages that automatically redirect to other pages. Also, recall that N.E.O. indicates results having a particularly small sample size, which the reader should be cautious to believe.

Table 4.38: Descriptive statistics: the number-of-authors and revisions variables, for each wiki site sub-section

| Sub-section | St. Dev. | Min. | 1st Q. | Median | Mean | 3rd Q. | Max. | Skew | Kurt. |
|---|---|---|---|---|---|---|---|---|---|
| **Wikia** | | | | | | | | | |
| actor | (n=373) | | | | | | | | |
| authors | 2.9 | 0 | 3 | 4 | 4.65 | 5 | 23 | 2.94 | 11.72 |
| revisions | 6.03 | 0 | 3 | 4 | 5.95 | 6 | 54 | 4.25 | 22.8 |
| book | (n=325) | | | | | | | | |
| authors | 4.97 | 0 | 2 | 3 | 3.94 | 4 | 46 | 5.15 | 34.9 |
| revisions | 19.21 | 0 | 4 | 6 | 9.92 | 8 | 217 | 6.24 | 51.24 |
| crew | (n=60) | | | | | | | | |
| authors | 1.88 | 0 | 2 | 2 | 2.88 | 3 | 9 | 1.22 | 1.4 |

Continued on Next Page...

Table 4.38 – Continued

| Sub-section | St. Dev. | Min. | 1st Q. | Median | Mean | 3rd Q. | Max. | Skew | Kurt. |
|---|---|---|---|---|---|---|---|---|---|
| revisions | 7.79 | 0 | 2 | 3 | 6.17 | 6.25 | 38 | 2.51 | 6.45 |
| episode | (n=371) | | | | | | | | |
| authors | 7.55 | 1 | 13 | 16 | 16.53 | 19 | 47 | 1 | 3.11 |
| revisions | 21.1 | 1 | 20 | 25 | 29.78 | 34 | 196 | 3.15 | 15.36 |
| game | (n=15) | | | | | | | | |
| authors | 5.97 | 1 | 1.5 | 3 | 4.87 | 3 | 20 | 1.82 | 2.14 |
| revisions | 11.03 | 1 | 2.5 | 7 | 9.6 | 10.5 | 43 | 2.29 | 5.87 |
| omnipediac | (n=2,975) | | | | | | | | |
| authors | 8.66 | 0 | 3 | 5 | 7.7 | 9 | 98 | 4.62 | 31.46 |
| revisions | 28.99 | 0 | 5 | 9 | 16.5 | 16 | 396 | 6.34 | 53.78 |
| video | (n=25) | | | | | | | | |
| authors | 3.79 | 1 | 1 | 1 | 2.72 | 2 | 15 | 3.02 | 8.39 |
| revisions | 16.7 | 1 | 2 | 3 | 7.8 | 4 | 69 | 3.29 | 9.97 |
| **Wikipedia** | | | | | | | | | |
| actor | (n=371) | | | | | | | | |
| authors | 83.12 | 1 | 26 | 51 | 85.71 | 119.5 | 367 | 1.4 | 1.24 |
| revisions | 141.27 | 1 | 38.5 | 87 | 144.46 | 192.5 | 500 | 1.33 | 0.73 |
| author | (n=20) | | | | | | | | |
| authors | 31.07 | 4 | 13.5 | 26.5 | 33.05 | 41 | 142 | 2.42 | 7.72 |
| revisions | 54.66 | 4 | 20 | 50.5 | 62.25 | 87.5 | 242 | 1.89 | 5.31 |
| crew | (n=53) | | | | | | | | |
| authors | 68.97 | 2 | 13 | 26 | 54.32 | 68 | 321 | 2.37 | 5.73 |
| revisions | 115.65 | 3 | 19 | 50 | 92.13 | 112 | 500 | 2.16 | 4.31 |

Continued on Next Page...

Table 4.38 – Continued

| Sub-section | St. Dev. | Min. | 1st Q. | Median | Mean | 3rd Q. | Max. | Skew | Kurt. |
|---|---|---|---|---|---|---|---|---|---|
| episode | (n=47) | | | | | | | | |
| authors | 62.45 | 0 | 45 | 65 | 83.04 | 100 | 259 | 1.32 | 1.16 |
| revisions | 139.97 | 0 | 83 | 131 | 175.38 | 211.5 | 500 | 1.3 | 0.82 |
| game | (n=3) | | | | | | | | |
| authors | 160.13 | 0 | 12.5 | 25 | 104.67 | 157 | 289 | 1.68 | N.E.O. |
| revisions | 275.86 | 0 | 24 | 48 | 182.67 | 274 | 500 | 1.67 | N.E.O. |
| general | (n=8) | | | | | | | | |
| authors | 48.11 | 99 | 187.25 | 206.5 | 193.88 | 214.5 | 252 | -1.2 | 1.37 |
| revisions | 110.17 | 262 | 440.5 | 500 | 440.5 | 500 | 500 | -1.44 | 0 |
| list | (n=56) | | | | | | | | |
| authors | 71.49 | 1 | 1.75 | 3.5 | 39.43 | 20.75 | 224 | 1.81 | 1.63 |
| revisions | 176.4 | 1 | 1.75 | 4 | 95.64 | 88 | 500 | 1.76 | 1.37 |
| omni | (n=36) | | | | | | | | |
| authors | 66.86 | 76 | 143.5 | 216.5 | 201.25 | 256.75 | 294 | -0.4 | -1.04 |
| revisions | 131.77 | 151 | 248.5 | 427 | 382.89 | 500 | 500 | -0.65 | -1.19 |

On Wikia, episode and omnipediac pages had the most authors, whereas crew and video pages had the fewest. The same was true for revisions, though adding actors pages to the list of those with the fewest. Omnipedia, episode, and book pages had the most variance, and actor and crew pages the least, on both variables. Episode, game, and video pages always

had both authors and revisions, possibly suggesting the absence of redirect pages. Some omnipediac pages have many authors and revisions, which was also true to lesser degree for book and episode pages. Book and omnipediac pages were the most left skewed on both variables, and episode, crew, and game pages the least. The same was true for kurtosis, though episodes became the third least peaked. Consistent with the text length analysis, many Wikia authors clearly wanted to edit episode pages, as well as omnipediac pages to a lesser degree. Few were interested in editing pages about actors, crew, and videos. These results also show that, like GateWorld, omnipediac, as well as book, pages were produced perhaps more perfunctorily by a smaller set of devoted fans.

On Wikipedia, omnipediac, general, and game pages had the most authors, and author (i.e., pages about book authors) and crew pages the fewest. However, general pages had more revisions on average than did omnipediac pages, perhaps due to general pages' greater popularity/visibility. Game and actor pages had the most variance in number of authors, and author and general pages the least. The same was true of revisions, adding list pages after game pages. Only episode and game pages were without authors and revisions, possibly indicating a redirects pattern opposite that of Wikia. Some popular actor and crew pages had especially many authors, and all except author pages hit the site's maximum 500 result limit for revisions. Author and crew pages were the most left skewed and high-peaked on both variables, whereas general and omnipediac pages were skewed slightly to the right, and omnipediac pages had a slightly flattened/negative kurtosis. Hence, whereas a smaller group of authors focused on preening the public-facing, episode-like general pages, a more diverse group made fewer, possibly more substantive, edits to omnipediac pages. By contrast, a relatively large and homogeneous group of authors revised game pages. Finally,

author and crew pages were usually given little attention, except for pages on the primary cast and crew, which received much.

### *Modeling results: Iteratively Weighted Least Squares*

As with the page length results, only those interactions that did not entirely consist of word usage variables will be presented.

First, regarding **content-related variables**, on most site sections, either a photographic or PDF image meant higher author counts, generally as a sign that more people had invested time and energy in that page, because of its popularity or notability. On Wikia episode and Wikipedia actor pages, PDFs, which most often included supporting materials on these pages (cf. §4.2.2), were associated with more authors, on the latter page type especially when many links were present on the page. PNG images were associated with higher author counts on Wikipedia actor and episode pages. The same was true for JPG images on Wikia book, actor, and omnipediac pages, though only on the latter two when occurring in concert with biographical texts. However, Wikipedia actor pages having either many revisions as well as PDFs, or simply the presence of GIFs, tended to have fewer authors. Due to the rarity of GIFs and PDFs, this was probably because those pages were the obscure pet projects of those few authors.

Within variables typical of actor pages (e.g., marital status, height, and awards), on Wikipedia, texts describing the actor's personal life and background, as well as lists of their notable roles, were most associated with high author counts. These fields were typical of pages about the most popular cast members. However, more nuanced fields – such as lists of actors' filmographies, details of their early life, lists of their theatrical appearances, and

lists of their awards – were associated with fewer, possibly more deeply invested, authors. On Wikia, notable roles, lengthy texts, biographical and birth information, marital status, height, nationality, and hair color were associated with high author counts. These fields are somewhat more role-playing game-like than on Wikipedia. Pages with fewer authors focused on lists of trivia, birth information without textual biographical explication, and nicknames.

Similarly, Wikia book, episode, and crew pages that focused on details – such as when an event occurred, what titles a crew member directed, or episode writers and airdates – tended to have fewer authors. Also, on Wikia omnipediac pages, those pages that covered more substantive issues – such as which actor played a certain character or what was the stargate address of a planet – received more authorial attention than pages that discussed the length, hull characteristics, maximum speed or passenger compliment of a vessel, or behind-the-scenes details.

Fewer **word usage variables** significantly predicted author counts than text length. Wikipedia actor pages with many authors tended to have many to-be verbs, nominalizations, long and passive sentences, sentences beginning with pronouns and with regular and subordinate conjunctions. They also scored highly on the ARI and Fog readability metrics. Wikipedia actor pages with few authors tended to have interrogative sentence beginnings, and scored highly on the other readability metrics. Though somewhat overlapping, these patterns correspond fairly well to the pithy vs. lengthy reviewing styles described in §4.2.5. Whereas more authors contributed to brief and questioning actor pages, fewer contributed to more lengthy and interpretive pages. Wikia actor pages with many authors usually had interrogative and subordinate conjunctive sentence beginnings, whereas such pages with

few authors had many short and passive sentences. This could indicate a scenario reversed from Wikipedia, where, because the site is smaller and the fans presumably more devoted, more authors congregated around the lengthy and interpretive writing than the brief.

**Link variables** also had the most effect on Wikipedia actor pages, with most-authored pages being associated with high PageRanks, overall link counts, and inlink counts. Wikipedia episode pages with many authors also had many overall links. These are clear cumulative advantage trends. However, Wikia omnipediac pages with many authors tended to have lower PageRanks, suggesting that those authors perhaps found those pages through their fandom social networks rather than through search engines.

Finally, regarding **authorship variables**, here referring to only the revisions variable, Wikia actor, crew, episode, and omnipediac pages, as well as Wikipedia actor and crew pages, all had more authors when revisions were higher. This is probably due to simple popularity or mass agglomeration behavior. However, high authorship was also associated with lower revisions on Wikia omnipediac pages having many sentences beginning with prepositions and/or conjunctions, on Wikia crew pages having both many revisions and listing many writing credits, and on Wikia book pages having both many revisions and a high SMOG score. The first of these suggests the reviewing writing style, namely that many authors come to the Wikia omnipediac pages, but few make many revisions, perhaps because those pages are much-watched and difficult to change without arguing with other authors. As for the second, authors show up and participate only for the most accomplished crew members. Finally, for the third, high SMOG scores are common for book pages, especially those manifesting the book/author writing style. That the most quintessential book pages would also be those that drew the most authors and authorial participation

makes sense, especially if they are the most quintessential because they are the most notable books.

## Conclusion

### *Descriptive statistics*

Webpage lengths shared several common patterns across all of the sites, except IMDb. On GateWorld, episode pages were produced by the site's editors in a much more variant, or less consistently structured, manner than were omnipedia pages. Wikia's page lengths were most comparable to GateWorld's, though Wikians wrote more about omnipediac and game topics. The focus on Wikipedia pages was clearly towards series and key episode summary pages, followed by game and omnipediac pages, almost as though Wikia had repackaged for a general public audience. Also like Wikia, actor, list, and omnipediac pages on Wikipedia seemed to be produced somewhat perfunctorily. By comparison, IMDb name (i.e., cast and crew) pages appeared less routinely structured than did title pages, suggesting a pattern opposite from the other sites, that name pages were more the site's focus than title pages. Unlike IMDb, the wiki sites gave only small token pages to actors and crew, and GateWorld gave none.

Regarding numbers of authors, and consistent with the text length analysis, many Wikia authors clearly wanted to edit episode pages, as well as omnipediac pages to a lesser degree. Few were interested in editing pages about actors, crew, and videos. These results also showed that, like GateWorld, Wikia omnipediac, as well as book, pages were perhaps produced by a small set of devoted fans. On Wikipedia, whereas a smaller group of authors

focused on preening the public-facing, episode-like general pages, a more diverse group made fewer, possibly more substantive, edits to omnipediac pages. By contrast, a relatively large and homogeneous group of authors revised game pages. Finally, author and crew pages were usually given little attention, except for pages on the primary cast and crew, which received much.

***Regressing page lengths against IQ criteria***

Regressing page lengths in each website sub-section against all available IQ variables yielded the following cross-section and cross-site patterns. First, content-related variables. JPG images were more common on lengthy GateWorld episode pages, though they were less common on Wikipedia episode and Wikia crew pages, most likely due to the prevalence of lossless PNG photographs (e.g., of actors at conventions) on those sites. Popular wiki episode pages often linked to transcript or screenplay PDFs, especially on pages employing the mass agglomerative writing style discussed in §4.2.5.

Of cast and crew-related variables, the users who wrote the longest crew articles on Wikipedia were most interested in documenting the personal histories of crew members, which the authorship variables revealed to probably be the work of a few obsessive fans. On IMDb, more popular/accomplished actors were also more likely to have had their physical features documented by users. On Wikia actor pages, distinctions between acting roles internal and external to the Stargate franchise confirmed the finding from §4.3.11, that those pages acted as gateways to other similar sites. Finally, when Wikia users wanted to provide only a brief summary of an encyclopedic topic, they kept to descriptions of the current state of affairs, only brief and terse plot/history summaries, and the perfunctory

features of technologies.

Lengthy textual pages on each website sub-section evinced the following writing styles, as defined in §4.2.5. (Note that not all pages of a certain type had the same writing style.) Page types often exhibiting the mass agglomerative style, which involves many people contributing content to a page without giving much thought to the page's organization, included: GateWorld episode, Wikia actor and crew, and Wikipedia crew pages. The somewhat more organized, though prosaic, general documentation style was most in evidence on Wikia actor and crew pages. The book/author style predictably occurred on long GateWorld book pages, but also on long Wikia omnipediac pages. Finally, the reviewing style occurred on lengthy GateWorld book and omnipediac pages, Wikia book and crew pages, and Wikipedia crew pages.

Associating the readability metrics with page lengths allowed each website section to be characterized in terms of several additional length measures, namely the measures to which each heuristic readability metric gives mathematical priority (i.e., sentence, syllable, and character counts). GateWorld book, episode, and omnipedia pages had both many words and many sentences, as did Wikipedia actor pages. Wikia crew pages had both many words, sentences, and syllables. Finally, Wikia actor pages had both many words, syllables, and characters.

Regarding link and authorship variables, whereas lengthy GateWorld episode pages often adopted a book/author writing style with few external links, lengthy wiki episode pages were closer to the reviewer writing style, with many external links. PageRanks were highest on lengthy Wikia omnipediac pages written in the reviewer or general documentation styles, IMDb title pages in the reviewer style, and Wikia episode pages in the book/author

style. Also, lengthy general documentation omnipediac pages on Wikia were revised by many, though more review-oriented omnipediac pages likely had only a small number of substantive revisions posted to them. Author counts followed a similar pattern.

### *Regressing author counts against IQ criteria*

On most site sections, either a photographic or PDF image meant higher author counts, generally as a sign that more people had invested time and energy in that page. However, Wikipedia actor pages with many PDF or GIF images, which were rare, often had fewer authors, probably because those pages were the obscure pet projects of just those few authors.

Texts describing an actor's personal life, background, or nationality, as well as lists of their notable roles, were most associated with high author counts, and were typical of the most popular cast pages on the wiki sites. More detailed and nuanced texts, which were associated with fewer (and probably more deeply invested) authors, on these sites focused on actors' filmographies, early life details, lists of theatrical appearances, lists of awards, and personal trivia. A similar distinction also applied to book, crew, episode, and omnipediac pages on Wikia, where pages that focused on minutiae tended to have fewer authors. Like in §4.3.10, Wikia's fields also appeared to be more role-playing game-like, itemizing the characteristics of people and things, than Wikipedia's.

Wikipedia and Wikia actor pages displayed somewhat opposing writing style patterns, with respect to authorship. Whereas Wikipedia actor pages written in the brief and questioning reviewer style attracted larger numbers of authors, Wikia's authors were drawn to pages written in the longer and more interpretive reviewer style (§4.2.5 for writing style

definitions). This could be because Wikia was a smaller site, with fan authors who were presumably more devoted.

Preferential attachment (i.e., rich-get-richer) processes were in evidence on Wikipedia actor and episode pages, with most-authored pages being associated with high PageRanks, overall link counts, and inlink counts. However, Wikia omnipediac pages with many authors tended to have lower PageRanks, suggesting that those authors found those pages through their fandom social networks or Wikia's own search engine, rather than through large search engine companies or other popular fansites.

Finally, though author and revision counts were in direct correspondence on most pages, on Wikia omnipediac pages employing the reviewer writing style, high authorship was associated with fewer revisions. Though many authors may come to those Wikia omnipediac pages, few make revisions, perhaps because those pages are much-watched and difficult to change without inciting an argument with other authors.

### 4.4.4 Length and mass collaboration: conciseness and organization

**RC5:** In this context, to what extent do wiki-produced articles contain more lists of facts and trivia, longer texts and sentences, and less organization than editorially produced articles?

This question was answered in terms of 10 variables: counts of lists and tables containing facts or trivia (e.g., not external link lists on the wiki sites), counts of page sections, the eight word usage variables measuring aspects of textual length counted by the POSIX `style` utility (i.e., character, word, and sentence counts; counts of relatively short and long

sentences; word counts of each text's shortest and longest sentences; and sentences' average length in terms of words). The list/table and section count variables were measured through manual content analysis of the random pages sample (§3.3), representing a random assortment of pages from each site. Tables were counted as well as lists, because the two were used interchangeably by the sites to present factual and trivia information, according to the nature of the information being presented. The length variables were measured automatically on the all-pages sample (§3.3) by the `style` utility.

For the purposes of this analysis, degree of page organization will be equated with the page section count variable. Although a close, qualitative reading of textual pages could doubtlessly yield a richer account of organizational tactics used by these authors, a quantitative approach also offers certain intellectual advantages. Pages with more sections suggest that authors have put forth more effort towards organizing a page's contents. A quantitative approach can summarize this degree of effort across all of the relevant pages of each, which would be prohibitively difficult to do qualitatively. Such a summary could be used to identify candidate pages and sites for closer examination. Providing such an initial, exploratory view of the nature of information in this new media phenomenon is the purpose of this project, not to critique the specific information organizational techniques of these authors. Also, notice that the research question asks whether there is [more or] less organization, not better or worse.

In this section, the list/table and section variable results are presented first, followed by length variables' results, and a conclusion.

### Lists, tables, and sections

Table 4.39 shows descriptive statistics for the list/table and section variables, with respect to each website.

Table 4.39: Descriptive statistics: the list/table and section variables, with respect to each website

| Section | St. Dev. | Min. | 1st Q. | Median | Mean | 3rd Q. | Max. | Skew | Kurt. | n |
|---|---|---|---|---|---|---|---|---|---|---|
| **Lists/tables** | | | | | | | | | | |
| GateWorld | 1.32 | 3 | 3 | 3 | 3.49 | 3 | 10 | 2.96 | 8.51 | 346 |
| IMDb | 2.21 | 0 | 2 | 3 | 3.47 | 4 | 14 | 2.01 | 4.36 | 320 |
| Wikia | 2.01 | 0 | 1 | 1 | 1.77 | 2 | 15 | 2.96 | 12.8 | 353 |
| Wikipedia | 3.28 | 0 | 4 | 5 | 5.43 | 6 | 23 | 2.28 | 7.4 | 234 |
| **Sections** | | | | | | | | | | |
| GateWorld | 1.81 | 1 | 2 | 4 | 3.48 | 4 | 12 | 1.79 | 4.29 | 346 |
| IMDb | 0.8 | 2 | 4 | 4 | 4.45 | 5 | 9 | 2.3 | 12.08 | 320 |
| Wikia | 1.78 | 1 | 1 | 2 | 2.35 | 3 | 11 | 1.86 | 4.23 | 353 |
| Wikipedia | 2.69 | 0 | 3 | 5 | 5.35 | 6 | 16 | 1.31 | 2.77 | 234 |

Did wiki sites have more lists/tables of facts and trivia than edited sites, in this context, as the IQ literature would suspect? Per page and for a random sample, on average, the IQ literature's suspicion was *true only for Wikipedia, not for Wikia*. GateWorld's pages always contained at least three lists, and rarely had more. IMDb and Wikia had similar

minimums, spreads, and maximums, though Wikia's average values were lower. Wikipedia had the widest spread, the highest averages, and the highest maximums. All of the sites were left skewed a similar amount, indicating that most pages had relatively few lists. Wikia most routinely included its typical amount of lists, probably referring to short lists in Infoboxes, which did indeed occur routinely. Hence, at least for this context, the literature's generalization should be restated to make clear that the assertion only applies to large wikis that provide high-level summary information to the general public (e.g., Wikipedia), not to wikis oriented towards a specific community or topic.

Did pages on wiki sites have less organization (i.e., fewer sections) than on editor-controlled sites? No; in this context, the divide had more to do with site size. The mean values indicate that the large sites usually made more sections, and the smaller sites fewer. Medians and means showed that Wikipedia pages typically had more sections than either of the editor-controlled sites. Most of GateWorld and Wikia's other statistics were comparable. IMDb was the most routinized in its use of sectioning, and Wikipedia the most diverse. Therefore, again, the IQ literature's generalization should be modified. Smaller sites may feature less organized information than larger sites, though editor-controlled sites' sections may be more formulaically applied than wiki sites'.

Going a step beyond the research question, it is reasonable to wonder to what degree a kind of organizing mania existed on these sites. Since these sites exist for the purpose of documenting something, to what degree might a more basic and undiscriminating act of dividing things into lists, tables, and sections have become normative on these sites? The correlations between list/table and section variables within each site are as follows: GateWorld $r = 0.59$, IMDb $r = 0.29$, Wikia $r = 0.57$, and Wikipedia $r = 0.7$. This shows that

a page's having both lists/tables and sections was relatively uncommon on IMDb, fairly and similarly common on both small sites, and quite common on Wikipedia. Hence, it is unlikely that IMDb's users created both lists/tables and sections as part of their documentation project (probably because only IMDb's editors were allowed to create page sections). However, this phenomenon was fairly common for both GateWorld's editors and Wikia's users, and quite common on Wikipedia. As the behavior was most common on the largest and most open site, it might indicate that, when trying to organize information, the general public may try all conceptual tools available to them with similar frequency. This is also suggested/supported by the similar magnitudes of list/table and section counts across both halves of table 4.39.

### Length

Table 4.40 presents descriptive statistics for a variety of text and sentence length measurements, with respect to each site. As in §4.4.3, IMDb's statistics were calculated on name and title pages from the free FTP dataset as well as character pages from the matched pages sample. The variables in this section are defined in detail in appendix A, table A.9.

Table 4.40: Descriptive statistics: a variety of text and sentence length measurements, with respect to each site

| Sub-section | St. Dev. | Min. | 1st Q. | Median | Mean | 3rd Q. | Max. | Skew | Kurt. |
|---|---|---|---|---|---|---|---|---|---|
| **GateWorld** | (n=3,025) | | | | | | | | |
| chars | 2,107.66 | 0 | 157 | 373 | 1,018.11 | 786 | 21,071 | 4.2 | 21.66 |

Continued on Next Page...

Table 4.40 – Continued

| Sub-section | St. Dev. | Min. | 1st Q. | Median | Mean | 3rd Q. | Max. | Skew | Kurt. |
|---|---|---|---|---|---|---|---|---|---|
| words | 463.92 | 0 | 33 | 78 | 219.77 | 164 | 4,660 | 4.22 | 21.77 |
| sents | 30.21 | 0 | 2 | 5 | 14.27 | 10 | 273 | 4.11 | 19.73 |
| s.avlen.w | 7.06 | 0 | 11 | 15.5 | 14.48 | 19 | 44 | -0.35 | -0.1 |
| s.short | 11.89 | 0 | 1 | 2 | 5.58 | 4 | 117 | 4.34 | 23.12 |
| s.long | 4.34 | 0 | 0 | 0 | 1.77 | 1 | 58 | 4.88 | 32.58 |
| s.shortest.w | 6.12 | 0 | 1 | 5 | 6.37 | 10 | 44 | 1.36 | 2.39 |
| s.longest.w | 13.87 | 0 | 18 | 27 | 25.84 | 35 | 102 | 0.07 | 0.57 |
| **IMDb** | (n=1,029) | | | | | | | | |
| chars | 632.36 | 0 | 0 | 0 | 354.1 | 534 | 7,505 | 3.69 | 24.35 |
| words | 133.95 | 0 | 0 | 0 | 75.48 | 116 | 1,607 | 3.7 | 24.88 |
| sents | 8.38 | 0 | 0 | 0 | 4.37 | 7 | 157 | 7.45 | 112.48 |
| s.avlen.w | 9.23 | 0 | 0 | 0 | 7.17 | 15.4 | 38.6 | 0.8 | -0.74 |
| s.short | 3.57 | 0 | 0 | 0 | 1.78 | 3 | 56 | 6.09 | 67.78 |
| s.long | 1.42 | 0 | 0 | 0 | 0.69 | 1 | 20 | 4.66 | 42.36 |
| s.shortest.w | 3.34 | 0 | 0 | 0 | 1.54 | 2 | 25 | 3.41 | 13.38 |
| s.longest.w | 21.26 | 0 | 0 | 0 | 15.57 | 32 | 112 | 1.18 | 0.8 |
| **Wikia** | (n=4,144) | | | | | | | | |
| chars | 1,473.59 | 0 | 129 | 357 | 836.8 | 858 | 15,881 | 4.29 | 24.2 |
| words | 321.83 | 0 | 26 | 74 | 176.62 | 176 | 3,601 | 4.47 | 26.02 |
| sents | 23.13 | 0 | 3 | 6 | 11.83 | 12 | 330 | 6.24 | 50.51 |
| s.avlen.w | 7.29 | 0 | 7.3 | 12 | 11.71 | 16.6 | 111 | 0.64 | 7.99 |
| s.short | 10.02 | 0 | 1 | 3 | 4.9 | 5 | 156 | 6.89 | 64.05 |
| s.long | 5.57 | 0 | 0 | 1 | 1.89 | 2 | 198 | 21.8 | 712.17 |

Continued on Next Page...

Table 4.40 – Continued

| Sub-section | St. Dev. | Min. | 1st Q. | Median | Mean | 3rd Q. | Max. | Skew | Kurt. |
|---|---|---|---|---|---|---|---|---|---|
| s.shortest.w | 342 | 0 | 1 | 1 | 2.27 | 2 | 47 | 4.17 | 25.25 |
| s.longest.w | 17.28 | 0 | 14 | 26 | 25.78 | 36 | 335 | 1.6 | 24.41 |
| **Wikipedia** | (n=522) | | | | | | | | |
| chars | 2,124.96 | 0 | 433.5 | 855 | 1,636.94 | 1,803 | 14,470 | 2.68 | 8.29 |
| words | 430.8 | 0 | 86 | 172 | 333.56 | 373.75 | 2,896 | 2.63 | 7.92 |
| sents | 20.99 | 0 | 6 | 11 | 18.53 | 21 | 146 | 2.55 | 7.92 |
| s.avlen.w | 5.55 | 0 | 13.03 | 15.9 | 16.12 | 19 | 53.5 | 0.75 | 4.84 |
| s.short | 9.79 | 0 | 2 | 5 | 8.09 | 9 | 63 | 2.54 | 7.68 |
| s.long | 4.78 | 0 | 1 | 2 | 3.37 | 4 | 36 | 2.84 | 10.03 |
| s.shortest.w | 3.21 | 0 | 1 | 1 | 2.25 | 2 | 32 | 5.34 | 37.29 |
| s.longest.w | 19.35 | 0 | 27 | 37 | 39.46 | 49 | 203 | 1.77 | 10.05 |

Looking across the means and medians in that table, Wikipedia did indeed possess the longest texts, on average, whether measured in terms of both character, word, and sentence counts as well as sentences' average lengths in words. However, the second longest was not Wikia, but GateWorld, followed by Wikia. IMDb's pages were the shortest. When considering the other sentence variables, the only exception to the previous pattern was that IMDb had shortest of short sentences and GateWorld the longest of short sentences.

The longest pages divided in terms of site size, rather than by editorial model. GateWorld had the longest pages, followed by Wikia, though Wikipedia's values were similar

to Wikia's. IMDb again had the shortest pages. Wikia's longest pages had unusually long sentences, as well as unusual numbers of short and long sentences. All sites had pages that were empty of textual content (e.g., placeholder pages, redirect pages, multimedia galleries, etc.). This was so common on IMDb as to fill the descriptive statistics table with zeroes up to the median. The other sites' small page sizes followed the same Wikipedia-GateWorld-Wikia pattern as did the averages.

Wikipedia and GateWorld possessed similar character count standard deviations, though diverged on word and sentence counts, where GateWorld showed slightly greater spread than either of the wikis. This indicates more variety in word and sentence counts on GateWorld. Both of the smaller sites (GateWorld and Wikia) had comparable or slightly larger standard deviations than did Wikipedia on all other sentence variable except the longest sentences' word counts. IMDb's standard deviations were comparable or smaller than Wikipedia's, with the exception of sentence average lengths, for which IMDb's value exceed all of the other sites'. This suggests that both GateWorld and Wikia had more variety in their sentence lengths than did Wikipedia, though IMDb's average sentence lengths varied the most. IMDb also evinced the largest spread of especially long sentences, to which both wikis came second.

All of the sites' page lengths, on every measure except sentence average lengths on GateWorld, were left/positively skewed to some degree. This means that most pages on these sites were relatively short. The sentence average length and longest sentence variables consistently displayed the smallest skews. For the average length variable, this could be due to the Central Limit Theorem, which says that the means of sufficiently large numbers of independent random variables will tend towards the normal distribution, which is not

skewed (Rice, 1995). Nevertheless, this variable was slightly left skewed on all sites except GateWorld, where it was slightly right skewed. Of the longest sentences variables, only on GateWorld was the distribution of these sentences not skewed towards shorter long sentences. The long sentence count variable's skew was unusually large and positive on Wikia, indicating that Wikia pages had very few long sentences, though the long sentences that did exist were very long. Examination of the few pages in question showed that their authors used semi-colons, instead of full stops, in several sentences. Skews were usually largest on Wikia and GateWorld, smaller on IMDb, and smallest on Wikipedia.

Regarding kurtosis, Wikia and IMDb's variables were usually the most peaked, GateWorld's moderately, and Wikipedia's the least. GateWorld and IMDb's average sentence length variables were slightly platykurtotic, though leptokurtotic on the wiki sites, indicating greater conformity on the wiki sites to a standard sentence length. Similarly, regarding the longest sentence variable, the editor-controlled sites had wider peaks about the mean, and the wikis had narrower. IMDb displayed a high peak about its sentence counts, as did Wikia about long sentence counts, suggesting either standardization or a few exceptional cases. IMDb and Wikia also had similarly high-peaked (and skewed) short sentence counts, suggesting that their pages both most often only had 1-5 short sentences.

**Conclusion**

Regarding lists and tables of facts and trivia, the IQ literature's expectation that wikis would have more such lists was true only for Wikipedia, not for Wikia. IMDb and Wikia also had similar minimums, standard deviations, and maximums, though Wikia's average values were lower. Most pages on all of the sites had relatively few lists. Hence, the literature's

generalization appears to apply only to large wikis that provide high-level summary information to the general public (e.g., Wikipedia), not to wikis oriented towards a specific community or topic.

Regarding organization of content into sections, the divide had more to do with site size than with editorial models. On average, larger sites usually made more sections per page, and smaller sites fewer. Wikipedia pages typically had more sections than either of the editor-controlled sites. Most of GateWorld and Wikia's other statistics were comparable. IMDb was the most routinized in its frequency of sectioning, and Wikipedia the least. Hence, smaller sites may feature less organized information than larger sites, and editor-controlled sites' sections may be more formulaically applied than wiki sites'.

Evidence was also found to suggest that, when trying to organize information, the general public may try all conceptual tools available to them with similar frequency. Correlations between lists/tables and sections on each site were high on Wikipedia, moderate on GateWorld and Wikia, and small on IMDb, where only professional editors could add page sections. Counts of lists/tables and sections also had similar magnitudes across all sites. Hence, though dedicated/specialized fan authors (i.e., on GateWorld and Wikia) customized to a moderate degree their sections and lists according to the needs of the topic at hand, authors closer to the general public (i.e., on Wikipedia) did so to a high degree.

Regarding page lengths, on average, Wikipedia's pages were indeed the longest and IMDb's the shortest, as the IQ literature would suspect, though GateWorld placed second longest and Wikia third, suggesting that small editor-controlled sites can have longer pages than small wikis. The longest pages occurred on the two smallest sites, and IMDb possessed many pages empty of any textual content (e.g., placeholder pages, having only

titles). GateWorld and Wikia also had more variety in their sentence lengths than did Wikipedia, though IMDb's average sentence lengths varied the most of any site. IMDb also evinced the largest spread of especially long sentences, to which both wikis came second. Most pages on all of these sites were relatively short. Skews were usually largest on Wikia and GateWorld, smaller on IMDb, and smallest on Wikipedia. Finally, Wikia and IMDb's variables were usually the most peaked, GateWorld's moderately, and Wikipedia's the least. Greater conformity to a standard sentence length existed on the wiki sites.

Therefore, the IQ literature's expectations about page organization and length in this context usually only held true for the largest sites.

### 4.4.5 Conclusion

Section 4.4.1 found the sites' agendas to be divided more in terms of either their incorporation statuses or ideologies than their editorial models. The non-profit (Wikipedia) and volunteer-run (GateWorld) sites both named their ownership and listed their organizational structures in considerable detail, whereas both corporate sites (IMDb and Wikia)provided only the names of the founders and top executives. Regarding funding sources, both non-corporate sites were the most forthcoming, listing donors and major affiliations by name, whereas both corporate sites spoke only of affiliations and third parties in the abstract. The same spectrum of organizations also reflected protections afforded to users' personal information. GateWorld and IMDb also both claimed copyright and ownership over all content, whether user- or editor-created, on their sites. On the other hand, both wikis frequently invoked the phrase "free culture," mandating that all content be released under a Creative

Commons license, and that all content belongs to its creators. Finally, common to all sites was a vagueness about physical location, probably due to staff safety and privacy concerns, as well as the offering of considerable narratives on their organizations' purposes, intentions, and histories. However, those sites that obtained funding through advertising did not mention advertising within their statements of passion for their mission, and the corporate sites took greater pains to obscure their history and past performance pages, in favor of presenting a simple and unified branding image.

Regarding why the fansites cited other fansites, nearly all pages contained a group of external links, for the purpose of recognizing people and companies affiliated with the page's topic and/or for directing the user to pages on other sites that covered the same topic (§4.4.2). Actor, author, and crew pages on the wiki sites often contained links for the purpose of identifying people who had worked in multiple franchises. Links to interviews usually served as references supporting arguments or quotes made in the page's text, though they could also provide intrigue or insight into a crew member or game developer's state of mind. Book and author pages often linked to sample chapters, or to book-like content that had been written by professional authors, but had not been published in print for some reason. Episode pages often linked to sites providing previews of upcoming episodes, behind-the-scenes footage, or information about the real-world version of a topic appearing in Stargate. Game pages often linked to fan communities or knowledge bases pertaining to specific games, for the purpose of letting gamers socialize and collaborate around playing the game. Finally, one instance of Christian proselytizing, or at least defense-taking, was found on a Wikipedia omnipediac page.

Quite a few IQ criteria were associated with page lengths (measured in terms of word

counts) and numbers of authors on the sites (§4.4.3). First descriptively, all of the sites, except IMDb, produced the most variety in episode, game, and (on the wiki sites) omnipediac pages, but afforded their cast and crew pages only either perfunctory or little attention. IMDb did the opposite. This was also evident in the author counts, though pages about especially popular cast and crew were given due attention. On GateWorld and Wikia, omnipediac and obscure pages were produced by small groups of devoted fans, whereas, on Wikipedia, a small group of fans preened the most popular pages, and a large number of authors considered, but rarely revised, omnipediac pages.

Now, the results of regressing length against IQ criteria. JPG images were more common on lengthy GateWorld episode pages, as were PNGs on the wikis. Popular wiki episode pages often linked to transcript or screenplay PDFs, especially on pages employing the mass agglomerative writing style (cf. §4.2.5). The users who wrote the longest crew articles on Wikipedia were most interested in documenting the personal histories of crew members, which the authorship variables revealed to probably be the work of a few obsessive fans. When Wikia users wanted to provide only a brief summary of an encyclopedic topic, they kept to descriptions of the current state of affairs, only brief and terse plot/history summaries, and the perfunctory features of technologies. The mass agglomerative writing style was most in evidence on lengthy episode and cast and crew pages; the general documentation style on other cast and crew pages; the book/author style on book and omnipediac pages; and the reviewing styles on lengthy book, crew, and omnipediac styles. GateWorld and Wikipedia actor pages often had many sentences, Wikia crew pages had many sentences and syllables, and Wikia actor pages had many syllables and characters. Whereas lengthy GateWorld episode pages often adopted a book/author writing style

with few external links, lengthy wiki episode pages were closer to the reviewer writing style, with many external links. PageRanks were highest on lengthy Wikia omnipediac pages written in the reviewer or general documentation styles, IMDb title pages in the reviewer style, and Wikia episode pages in the book/author style.

Fewer IQ criteria were associated with pages' author counts. On most site sections, either a photographic or PDF image meant higher author counts, generally as a sign that more people had invested time and energy in that page. Texts describing an actor's personal life, background, or nationality, as well as lists of their notable roles, were most associated with high author counts, and were typical of the most popular cast pages on the wiki sites. More detailed and nuanced texts, which were associated with fewer (and probably more deeply invested) authors, on these sites focused on actors' filmographies, early life details, lists of theatrical appearances, lists of awards, and personal trivia. Whereas Wikipedia actor pages written in the brief and questioning reviewer style attracted larger numbers of authors, Wikia's authors were drawn to pages written in the longer and more interpretive reviewer style, possibly suggesting more that Wikia fans were more devoted. Preferential attachment (i.e., rich-get-richer) processes were in evidence on Wikipedia actor and episode pages, with most-authored pages being associated with high PageRanks, overall link counts, and inlink counts. However, Wikia omnipediac pages with many authors tended to have lower PageRanks, suggesting that those authors found those pages through their fandom social networks or Wikia's own search engine. Finally, though author and revision counts were in direct correspondence on most pages, on Wikia omnipediac pages employing the reviewer writing style, high authorship was associated with fewer revisions, possibly because the social atmosphere around editing those pages was more competitive.

Finally, the IQ literature expected that wiki pages would have more lists of facts and trivia, less-organized pages, and longer texts than pages on editor-controlled sites (§4.4.4). Regarding lists, this expectation was found to be true only for Wikipedia, not for Wikia. IMDb and Wikia also had similar minimums, standard deviations, and maximums, though Wikia's average values were lower. Most pages on all of the sites had relatively few lists. Regarding organization of content into sections, the divide had more to do with site size than with editorial models. On average, larger sites usually made more sections per page, and smaller sites fewer. Smaller sites may feature less organized information than larger sites, and editor-controlled sites' sections may be more formulaically applied than wiki sites'. Evidence was also found to suggest that, when trying to organize information, the general public may try all conceptual tools available to them with similar frequency.

Regarding page lengths, on average, Wikipedia's pages were indeed the longest and IMDb's the shortest, as the IQ literature would suspect, though GateWorld placed second longest and Wikia third, suggesting that small editor-controlled sites can have longer pages than small wikis. The longest pages occurred on the two smallest sites, and IMDb possessed many pages empty of any textual content. Most pages on all of these sites were relatively short, and greater conformity to a standard sentence length existed on the wiki sites.

## 4.5 Conclusion

For a detailed summary of this Results chapter, please see the following Discussion chapter. For a brief summary, please see §6.1 in the Conclusion chapter.

# Chapter 5

# Discussion

## 5.1 Introduction

This dissertation project was designed to expose divisions between sites of different sizes and editorial models, in terms of information quality (IQ) criteria. Indeed, a number of divisions along those lines were found. Additionally, divisions based on sites' business models and missions, characteristics that were common to each site, as well as characteristics that were common to all of the sites were also discovered.

This project was also organized following a Peircean "cable" approach to reasoning (Preucel, 2006, p. 252), whereby a large number of relatively small and independent analyses were conducted, in order to paint a more complete and less fallible picture of this social phenomenon than if only one aspect of IQ had been focused upon in-depth. This approach was most appropriate/important because of the under-studied nature of this phenomenon. The result is that the findings from the various sections of the Results chapter (4) can be easily re-organized and interpreted in terms of types of site divisions.

Additionally, near the end of this chapter, a discussion is offered of how this study's results compare with the Wikipedia case study in Stvilia (2006).

## 5.2 IQ criteria related to site size

### 5.2.1 Small sites

#### Especially devoted fans

The smallest sites in this study (i.e., GateWorld and Wikia) were perhaps most characterized by contributions made to them by especially devoted fans. Their users consistently manifested a positive bias in the ratings they gave to episode and omnipediac pages (§4.3.8). They possessed the longest pages of any site (§4.4.4), and their long pages contained many media objects (§4.4.3). Hence, these sites might be the most likely candidates for library collection development. Leisure studies researchers could also fruitfully study fan devotion on Wikia's actor pages, which were written by many authors in the longer/interpretive reviewer style, as well as on obscure omnipediac pages on both sites, which were made by small groups of devotees (§4.4.3). That Wikia actor pages made by many authors were written in a longer/interpretive reviewer style could also be of interest to linguists.

#### Concerned less with the general public

Unlike the larger sites, GateWorld and Wikia made less of an attempt to accommodate the general public. Wikia's pages were updated less frequently than Wikipedia's and only

periodically, in sync with the academic calendar to which its target users were subject (§4.3.1). This may be of interest to leisure researchers of fan participation, and to collection developers needing to know how to schedule collection updates. Both sites contained more accessibility errors than the larger sites (§4.2.4), and made less of an attempt to organize page content into sections (§4.4.4). These data quality and organization issues could also be important for collection developers, as well as for those who administrate and create the pages, to know, so that appropriate remedies can be applied. That high PageRanked pages on Wikia were written in the same style as highly inlinked Wikipedia pages (§4.3.10) suggests that Wikipedia pages that are popular with the general public may act as conduits for related Wikia pages. This shared writing style and traffic pattern could be of interest to linguists and network researchers, as well as to those creating and administering the pages. Wikipedia may be providing the public face of Wikia. Finally, leisure and mass media researchers, as well as the sites' administrators, may be interested to learn that fans most interested in esoterica do not usually come from the general public or large search engines, but rather probably from fan community social networks and sites' local search engines (§4.4.3).

**Concerned more with providing a rich experience for fans**

Instead of accommodating the general public, the small sites focused on providing a rich experience for their target user group. These sites were the most likely to provide their own interpretive commentary, rather than merely deferring to the studio or mass media, points which may be of interest to leisure and mass media researchers, as well as collection developers (§§4.3.2 and 4.3.5). Their navigation structures were also specific to the franchise,

and contained a similar conceptual structure, which may be of interest to LIS researchers and bibliographers (§4.2.1). Their advertisements were more targetted towards a specific user group, namely younger and possibly more European users than those who visited the larger sites (§4.3.7). Both sites also provided more esoteric, bio-metric, and role-playing game-like information than the larger sites, perhaps in line with the types of games most often played by this younger age group (§§4.3.10 and 4.3.6). This could be of interest to marketing and cultural geography researchers, as well as those who manage advertising on the fansites.

### 5.2.2 Large sites

#### By/for the general public

IMDb and Wikipedia were clearly oriented towards facilitating consumption by the general public. They both relied on either public-made taxonomies/folksonomies or keyword-based search engines for navigation (§4.2.1), and had fewer accessibility errors than the small sites (§4.2.4), both of which may be of interest to LIS researchers. Wikipedia also had more broken links than did Wikia (§4.2.3), which their page administrators/creators may wish to remedy. Large sites' pages were updated continuously and more frequently than the small sites (§4.3.1), possibly of interest to fan participation researchers and collection developers. Finally, several findings could be relevant to marketing and other business research. IMDb displayed a wide variety of advertisements, and Wikipedia a generic donation banner (§4.3.7). Also, both sites linked more often to official franchise and corporate sites than did the small sites, and IMDb had many corporate affiliations (§4.3.6). Last,

their information often merely aggregated data obtained from production and marketing companies (§§4.3.2 and 4.3.5), such as technical, biographical, historical, and reception information.

**Adults wanting thoughtful summaries**

However, only a small sub-set of large sites' users actually maintained the most public facing pages (§4.4.3). Most were older users (compared to the smaller sites) who spent more effort contributing to long and substantive texts and lists on pages about the franchise's most thought-provoking topics (e.g., life elsewhere in the universe, ascension/enlightenment, and cultural commentary; §§4.3.10 and 4.4.4). Judging from Wikipedia's user profiles, these users often identified themselves as American (US and Canadian) males in their late 20s, employed and educated in a variety of fields, less physically active than Wikians, over-weight, either agnostic/atheistic or protestant Christian, heterosexual, and introverted-judging (I–J) on the Myers-Briggs personality matrix (§4.3.7). Additionally, popular cast and crew pages on Wikipedia drew the attention of many general public authors who wrote in the more superficial and questioning reviewer writing style, as well as accumulated PNG images of the people being documented (§4.4.3). These findings should be of interest to any studies of fan types and participation (e.g., in leisure studies, psychology, religious studies, cultural geography, or marketing), to linguists studying the writing styles of the general vs. connoisseuring public in this genre, or to collection developers wanting to know where to find the most thoughtful and celebrity-indexing of articles on large fansites.

## 5.3 IQ criteria related to editorial model

### 5.3.1 Editor-controlled sites

#### User-using, editor-supporting

Though more true for IMDb than GateWorld, both editor-controlled sites contained user interfaces that were essentially superficial mechanisms of taking in content from users, whereas their content management templates and back-end server systems were clearly oriented towards serving editors' and the company's business interests. Most obviously on IMDb, many pages were empty of any content, and those that did have content were on the most popular topics, filled with either technical information provided by the studio or mass agglomerative information from the general public (§4.2.5). It was as if IMDb was merely capitalizing on preferential attachment Web traffic, taking in their share of general fan browsing behavior and excitement surrounding the franchise, and licensing that shallow content as they could (cf. §5.4.1, on business models). This technique could be of interest to marketing and mass media researchers, as well as perhaps to leisure researchers and fan users who could benefit from knowing that people seeking leisure without being part of a fan community can easily wind up in such places.

The following findings explicate the point further. Both sites' interfaces contained many outdated and deprecated media and markup components (§§4.2.2 and 4.2.4). Also, recall that the free dataset and search tools provided by IMDb were antiquated (§3.2.3), and that GateWorld provided no database downloads (§3.2.3). Both sites used formulaic page sectioning and avoided repeatedly answering the same question, probably in order to minimize

editorial effort (§§4.4.4 and 4.2.1). They probably automated the checking of pages for broken links (§4.2.3), either out of editorial pride (for GateWorld) or because licensing content containing broken links to other parties might reflect poorly on the company (for IMDb). LIS researchers may find these interface, information architectural, availability, and knowledge and data management techniques interesting. Marketing notions of brand identity and reputation are also relevant. The content on their pages was at a lower reading level than on the wikis, probably indicative of general public writing (§4.2.5), which may be of interest to linguists and anyone wishing to use this material, such as educators or collection developers. The sites deferred to, and cited, the mass media for critical reception information (§4.3.4), and only used references either to support quotes/claims or to promote one of their affiliates (§4.3.5). This bias for other editor-controlled organizations could be relevant to journalism and business researchers. Finally, users were obligated to sign over the copyright for whatever they contributed to the organization (§4.4.1). This relates to issues of privacy and copyright, common in LIS (cf. §5.4.1).

**Editorial agendas**

However, editorial interests were not consistent across both sites. Whereas IMDb seemed primarily to exist for the purpose of filtering out profitable content from fans, GateWorld's editors were more journalistically motivated (recall that GateWorld's owner has a journalism degree, §4.4.1). GateWorld's editors' ratings of episodes were more balanced and critical than were regular users' (§4.3.8), editors preferred season premier pages having many inlinks and for which they were able to provide critiques (§4.3.9), and GateWorld's users preferred non-premier/-finale episode pages that provided lengthy editorial reviews

(§4.3.9). Pages were usually longer on GateWorld than on IMDb (§4.4.4), and GateWorld's editors wrote the longest episode pages in the chronicling book writing style, with few external links (§4.4.3). They also treated fans' content submissions like newspaper Letters to the Editor, naming their sources, but reserving copyright for themselves (§4.4.1). This difference in motivation allowed GateWorld to extend its content beyond general public mass agglomeration and description, both via editorial commentary and thoughtful fan contributions.

This editorial difference, and its results, should be of interest to mass media and journalism researchers, in that it seems to mirror a common print-based editorial trend. Publications that simply act as generic money filters, having minimal editorial investment, (e.g., classifieds) both produce and attract less substantive of content than do publications with greater editorial investment (e.g., newspapers, magazines, and journals).

### 5.3.2 Wiki sites

#### User freedom for user loyalty

In this context, the free and open-source culture to which both wikis' administrations ascribe, which applies free-use Creative Commons licenses to all user content and allows users to retain copyright (§4.4.1), comes with the price of lost licensing revenue. Therefore, the gamble of these sites is that, by giving users an enjoyable interface and the power to create and collaborate however they can/wish, rather like in a democracy, users will be willing to give back to the organization in support of the freedoms and facilities they enjoy, and not simply take them for granted (§4.2.1). On Wikia, "giving back" means clicking

on the advertisements; on Wikipedia, making a donation. The enjoyable interface on these sites includes: page templates and code generators that produce code free of errors (§4.2.4); ease of embedding a variety of contemporary media formats (§4.2.2); and the freedom to create pages (§4.3.6), sections (§4.4.4), link lists (§4.2.3), and so forth. As wiki sites are very popular, these points are not novel, but relate to considerable research on wikis, open-source, democracy, non-profit business models, interface design, and the like.

**Results of relinquishing editorial control**

What *is* novel are the results of relinquishing editorial control in this context. First, it resulted in the participation and training of more able writers. On both sites, the overall readability level of pages was around one grade higher than on editor-controlled sites (§4.2.5), including the texts written by GateWorld's editors. Evidence also exists for the emergence of socially normative, though undocumented, writing standards on the sites, such as conformance to a standard/conventional sentence length (§4.4.4). The potentially higher quality of writing on wiki fansites deserves more serious linguistic attention.

Second, considerable evidence exists suggesting that wiki users were more invested in the sites' content than those creating content on the editor-controlled sites. Both sites' users most often wrote about themes other than what the targetted advertising and stereotyped user profiling (see §4.3.7) would predict, namely travel and health, suggesting that they may have been writing about personal interests (§4.3.4). Wikipedia users also presented critical reception information in paragraph form, perhaps because they cared enough about the criticism to document it, rather than merely citing it, as did IMDb (§4.3.4). Long episode pages on both sites were written in the reviewer style, and contained many exter-

nal links, compared with GateWorld's episode pages in the book style with few external links (§4.4.3). This could indicate that those wiki pages document what the combined user community values about the topic, rather than what one editor or their organization values. Both sites provided the franchise's production dates, indicating both an interest in the franchise that exceeds the franchise's products, and a willingness to represent that interest on the site (§4.3.6). Wiki pages were also the only to include entire pages on cultural references, which made interpretive connections between aspects of the shows and the real world in which fans live (i.e., they personalized the shows; §4.3.6). Wikia user ratings also suggested the presence of either casual or thrill-seeking syndication viewers who most often sought easily digestible summary details from omnipediac pages (§§4.3.8 and 4.3.9). They were also the only sites where cast, crew, and authors that had worked in multiple franchises known/enjoyed by fans were identified (§4.4.2). Finally, the presence of many links to transcript/screenplay PDFs on pages written in the mass agglomerative style suggest that even pages dominated by general public authors catered to that audience's desire for minutia (§4.4.3). All of these findings suggest the need for more close interpretive or ethnographic work on fans' motivations for participation.

Finally, both wikis' external links and citations regularly encompassed more than only business and mass media affiliates, suggesting that fans included whatever they thought pertinent to the page's topic (§4.3.5). Although collection developers could benefit from knowing that wikis' external links in this genre tend to be less biased in favor of corporations, further work should examine wikis' degree of bias towards non-profit sources.

## 5.4 IQ criteria related to business model & motivation

### 5.4.1 For-profit sites

#### Selfish, possessive, vane, and secretive

Though these words may seem overly harsh, the reasons for choosing them suggest they are not far from the truth. Both IMDb and Wikia displayed behaviors indicative of opportunism and shortcut-taking; IMDb also disregarded many academic and international standards and best practices (§4.2.6). This indicates that they only put forth the effort that they must in order to become successful. Both sites' interfaces were filled with advertisements (§4.3.7), seeking profit from frequent user distraction. They offered only vague protections of user-disclosed information (§4.4.1), covering their own interests over the user's. They obscured their history and performance pages in favor of presenting a pristine-looking brand identity (§4.4.1), knowing, like social networking sites with privacy settings, most users accept the default settings/information and do not investigate (Acquisti, 2009). Regarding secrecy, both sites listed only their founders and executives, provided no organizational charts, and were vague about their disclosures of funding sources and affiliates (§4.4.1), essentially deferring accountability for any wrong-doing by the organizations' employees to its most powerful figures and their lawyers. This practice, which often makes legally confronting large corporations difficult for average citizens, is supported by the law in the United States, which defines a corporation as a separate legal entity, having its own legal personality (Kraakman et al., 2004, p. 2).

Researchers familiar with large corporate business and privacy practices will likely not

be surprised by these findings, though it is interesting that they extend to this genre as well. Namely, a common corporate rationale in the United States and Europe – with which many in academia, government, and non-profits would likely disagree – is that the primary point of establishing a for-profit corporation is to benefit both the company's employees/shareholders and customers, by providing a worthwhile service at a reasonable price. This rationale can be easy to see working in the case of small businesses and tangible goods. However, for such large and intangible corporations and services as these two sites, the degrees of social service and user exploitation manifested are essentially reduced to what their executives are willing to command the company to do (amid cost and performance pressures from shareholders), and whether enough of the international general public is willing to generate revenue for the organization/service to be able to continue in that form. As also shown in §5.3.1, this approach seems to drift more towards the user being used for the company's benefit, rather than vice versa.

### 5.4.2 Non-profit sites

**More generous, less user-disturbing, more protective, etc.**

By comparison, GateWorld (which is an ad- and donation-supported LLC, run by a small group of paid staff and fan volunteers) and Wikipedia pages contained few or no ads (§4.3.7). They both offered detailed and substantial protections of user-disclosed information; their history and performance pages were prominently displayed; GateWorld's branding was somewhat inconsistent throughout the site; their entire staffs were named, and Wikipedia provided detailed organization charts; and they detailed their funding sources

and affiliations by name (all §4.4.1). The antithesis was quite striking.

### 5.4.3 All-fan-made sites

#### Common themes, pages, and texts

Dividing the sites according to organizational mission exposed a triad of sites (GateWorld, Wikia, and Wikipedia) with content that was entirely created or compiled by fans (GateWorld's editors are also dedicated fans). The most characteristic feature of this triad was general agreement on which themes to include on which pages, which should be of interest to narrative and media studies scholars. For example, the sites' episode and character pages contained common themes in their body texts and production notes, agreed on what to include of characters' contexts, and the wikis agreed on which cursory details to include (§§4.3.4 and 4.3.6). Omnipediac pages also contained a core set of topics across the sites (§4.3.6). In addition to common themes, episode, game, and omnipediac pages were also given the most frequent attention, and cast and crew pages the least (§4.4.3). This could be of interest to fan participation researchers, and to bibliographers and collection developers. Finally, they all tended to focus on lengthy textual explications of complex themes (§4.3.4).

#### Less popular pages more substantive

In addition to the trend seen on every site (§§5.2 and 5.6), that highly PageRanked pages tended to be mass-agglomerated and have considerable, though rather shallow, content, pages on the triad of sites with low PageRanks and many inlinks, especially on omnipediac pages about people and technologies, often had more substantive (i.e., thoughtful) content

(§4.3.10). This likely represents participation by devoted fans, which could be interesting to researchers of fan participation, as well as collection developers wanting substantive information.

### 5.4.4 Ad-supported sites

Finally, the ad-supported sites (GateWorld, IMDb, and Wikia) shared in common that none mentioned advertising as part of their missions statements (§4.4.1), suggesting they may have viewed ads as a necessary evil. This may be of interest to marketing researchers.

## 5.5 IQ criteria specific to each site

### 5.5.1 GateWorld

#### Esoteric and devoted

GateWorld was considered by the general public as the place to go for Stargate esoterica (§4.3.10). Its pages included many behind-the-scenes and production details; attention to special effects; and the franchise's depictions of studious academicians, usually archaeologists, linguists, and physicists (§4.3.4). It was also the only site to devote entire pages to episode transcripts, editorial reviews, and making-of information (§4.3.6), though cultural references received less attention than on the wiki sites (§4.3.6). These findings could be of interest to media researchers, members of the represented fields who want to see how they are depicted on television, and bibliographers and collection developers looking for sources and types of esoteric information about the franchise.

#### Standard and dominant ads

Amazon and iTunes ads, for purchasing merchandise or digital downloads of episodes, were standard on many pages. Also, ads for IT products (e.g., computers and cellphones) tended to occupy all of the advertising positions at once on pages where they appeared, possibly because those large and profitable companies could afford to buy all of the site's advertising space (§4.3.7). These advertising behaviors could be of interest to marketing researchers.

### 5.5.2 IMDb

#### Relationships with other companies

IMDb was unique in the way that it cited other companies and identified its affiliates. Technical production details, awards, quotes, and media liner notes all implied, though did not always explicitly mention, a source company (§§4.3.2 and 4.3.4). Also, their including the services of other companies embedded in their own site showed that they seek both opportunistic and stable business partnerships (§4.3.7). This variety of business relationships might be interesting to a marketing or business researcher, as might the variety of sources to a bibliographer.

#### Signs of investment

Two signs of user or editor investment were unique to the site. Sections on cultural references were occasionally interpretive, resembling a smaller or wiki site (§4.3.2). This could be of interest to fan participation researchers. Also, title pages with the highest PageRanks,

which one might expect to contain general public mass-agglomerated content, were often written in the reviewer writing style, suggesting that those pages may have been deemed popular or interesting enough to garner the attention by the site's editorial staff. This ability to distinguish editorial attention from general public fan attention via linguistic and network cues could benefit anyone seeking content on large editor-controlled sites that rises above the quality level of general public mass-agglomeration.

### 5.5.3 Wikia

#### Where to look for substance

Researchers or others seeking the most substantive content on Wikia should direct their attention to several standard locations. On most pages, superficial (e.g., bio-metric) lists were separated from more narrative text sections; biographical and historiographical prose were more common on pages about people; and interpretive sections were more common on pages about plot arcs or events (§4.3.2). Substantive omnipediac pages had higher PageRanks, many authors but few revisions, and were written in the reviewer or general documentation styles (§4.4.3). On the other hand, less substantive omnipediac pages were typically shorter, summarizing the current state of affairs in the plotline, and the perfunctory features of technologies. Finally, episode pages with higher PageRanks that were written in the book style tended to be more substantive. These correspondences between text, page, and popularity types may be of interest to fan participation researchers, linguists, and librarians.

#### Standard targetted ads

Expect most ads on the site to regard either other science fiction franchises or retail merchandise (§4.3.7). This genre- and demographic-based targetting may be of interest to media and marketing researchers.

### 5.5.4 Wikipedia

#### Scarce extreme user investment

As one might expect, especially devoted fan behavior occurred less on the large sites than the small, though it did occur rarely on Wikipedia. Several citations were provided to academic literature (§4.3.5). One instance of Christian proselytizing or defense-taking was observed (§4.4.2). And, only a few fans wrote the longest crew pages (§4.4.3). These occurrences may be of interest to bibliometric researchers, scholars of debates between western science and religion, and fan participation.

#### Scarce ads

Finally, being donation-based, Wikipedia's only ad was the occasional banner at the top of the page containing a fundraising plea from the founder (§§4.3.6 and 4.3.7). The site listed its major donors and endowments, and claimed to have no business affiliations or partnerships (§4.4.1).

## 5.6 IQ criteria common to all sites

### 5.6.1 Content

The following writing styles, documentary approaches, advertisement patterns, and link types occurred on every site.

Six writing styles were present in page texts: mass-agglomeration, general documentation, book/author, general reviewing, pithy/interrogative reviewing, and long/commentarial reviewing (§4.2.5). The mass-agglomerative style occurred most often on lengthy episode, cast, and crew pages; general documentation on cast and crew pages; book/author on book and omnipediac pages; and the reviewing styles on book, crew, and omnipediac pages. This may be of interest to linguistic, narrative, fan participation, and media researchers.

Regarding documentary approach, episode and character pages were standard, all sites gave episode release dates, and most pages had at least one lengthy summary text (§4.3.6). Cast, crew, author, and game developer pages shared a set of common fields, namely: birth dates and places, names, spouses, key episodes, height, trivia, past filmography, titles-by, writer-of, director-of, editor-of, and producer-of (§4.3.6). On the wikis, more authors typically contributed to pages having photos or PDFs, or to actor pages describing the actor's personal life, background, nationality, and notable roles (§4.4.3). On the other hand, few authors contributed to pages with more detailed texts, or to actor pages focusing on actors' filmographies, early lives, theatre work, awards, or personal trivia. All possessed at least an interpretive level of original research (§4.3.6). Finally, all were vague about their physical location, and provided descriptions of organizational purpose, intent, and history (§4.4.1). These are this project's high-level contributions to research on the bibliographic

structures of these fansites' information. Additionally, fan participation researchers, scholars of what qualifies as 'research,' and business researchers of organizations' disclosures and self-representations may find them interesting.

The following set of standard vendor advertisement categories emerged: those dealing with achievement, savings, or status; with spending; with one's personal environment; with good vs. bad investments; with risky vs. predictable behaviors or investments; with fixity vs. transience; and with the mass media vs. the science fiction and gaming industries (§4.3.7). Business and marketing researchers may find these of interest.

Finally, regarding links, a standard group of external links, often present at the end of pages, went to other sites that either provided coverage of the same topic or were affiliates of the current site (§4.4.2). To at least this degree, the fansites appeared to be concerned with users' finding the information they desired, if the originating site was not adequate. This lack of revenue protectionism may be due to the leisure nature of the sites, which a leisure or business researcher may find worthy of further study.

### 5.6.2 Shallow architecture, poor accessibility

On all of the sites, the navigation menus went only to either the franchise or series level, at best, after which the user was required to browse (§4.2.1). Accessibility errors were also more common than HTML errors, the most common being the hiding of page content inside JavaScripts, as well as tables and media objects without alternative textual content (§4.2.4). These findings may be of interest to information architecture and accessibility researchers in LIS, as well as to anyone wishing to use or link-to the content in an unmodified form.

### 5.6.3 Popular means substantial

Perhaps the crux of the mass-agglomerative findings presented variously throughout this chapter is the general trend that popular content – whether popularity is measured in terms of inlinks, PageRanks, or counts of unique authors – was often more substantial/copious, though not always more substantive (§§4.2.2, 4.3.10, and 4.4.3). This is most clearly seen on episode and omnipediac pages, where pages having high PageRanks contained many short words, indicative of the general public, whereas pages having lower PageRanks contained longer words, typical of the more educated/nerdy science fiction user population (§4.3.10; cf. §5.2). Fan participation researchers, and collection developers seeking for different audiences could find this helpful.

### 5.6.4 Commerce-dominated link structures

On the bulk of pages, which were relatively short and with few links or inlinks, higher PageRanks usually equated to higher inlink counts, meaning that Google's reputation-weighted rating of the pages was comparable to unweighted general inlinks, implying that reputation was not a factor. However, among those fewer pages with high PageRanks or inlinks, evidence of preferential attachment, as well as possible manipulation of reputation by the pages' stakeholders and favoritism among large sites, was apparent (§4.3.10). Such biases should be investigated further by mass media researchers.

Among these popular pages, the evidence suggested a network with multiple strongly connected cores both within and across these sites, surrounded by a level of links most often to/from commercial and professional organizations, and less often to/from wikis and

small fansites (§4.3.10). The strong ties between the sites under study, and only weaker ties to other fansites further supports the choice of studying these four sites. Excluding links from the United States, which are difficult to identify based only on domain names (because few sites used the .us country code), inlinking patterns to the sites under study from northern European domains most resembled the inlinks of commercial outlets, whereas those from continental European domains more resembled wiki and fansite inlinks. Commerce links may have dominated the periphery because three of the four sites have a commercial agenda. However, this, and the connection of northern Europeans more to commerce than continental Europeans, deserve further attention by mass media, cultural geography, and social network researchers.

A number of common patterns, consistent with this network model, can be seen in those cases where links from one of the sites under study went to a similar (non-studied) site. Links to sites hosting interviews with the cast and crew most often went to commercial media outlets. Pages about books and authors often linked to online samples of their writing, typically hosted on either a publisher or the author's professional website. Episode pages often linked to previews and behind-the-scenes videos hosted on the network or studio's site. Less common on episode pages were links to sites about the real-world versions of people and things appearing in the franchise, often located on wikis and smaller fansites. Finally, game pages, which were less common than episode or cast/crew pages, most often linked to fan communities and knowledge-bases existing to facilitate gameplay (all §4.4.2).

### 5.6.5 Indiscriminate public organization

One final result could also be significant, and of interest to classification and HCI researchers. Evidence existed suggesting that, when trying to organize information, the general public may try all conceptual tools available to them with similar frequency (§4.4.4).

## 5.7 Comparison with Stvilia

Finally, one of the two case studies in Stvilia (2006), involving both qualitative and quantitative analyses, was of a random sample of page types and collaborative content creation behaviors on Wikipedia. Qualitative analysis of user discussions were used to identify the most common IQ issues affecting Wikipedia, both to give user accounts of the issues and to help choose which IQ variables to use and how to measure them (pp. 158-196). Relevant quantitative findings involved both description and modeling (pp. 175 and 178-185), both of which the remainder of this paragraph will relate to the current dissertation's findings. Stargate-related pages had higher Flesch and lower Kincaid median scores than did Featured or random Wikipedia articles. Considering those metrics' formulae, this suggests that Stargate pages had relatively many sentences with relatively few words and syllables (§4.2.5). Most Stargate pages contained slightly fewer images than Featured Articles' (FA) median of five, though some Stargate pages had upwards of 10 (§4.2.2). Overall, Stargate pages had a lower median character count than did random pages (§4.4.4). Stargate pages' median ages (1,343 days) were older than FAs'. Stargate revision counts had a median of 90, which was more than random articles,' but not than FAs'. Like the current dissertation, Stvilia found evidence of preferential attachment effects in user behaviors (p. 181; cf.

§4.3.10).

Additionally, a non-parametric Kruskal-Wallis ANOVA of the variables across random pages from four sub-genres of Wikipedia could at least confirm that "Most of the measures and metrics showed significant dependence on article genre," (p. 184) a finding consistent with the differences between the current dissertation's results and Stvilia's. If the EFA results are to be believed, they say the following: 'authoritative' pages had many user content reverts, unique editors, total edits, external links, and anonymous user edits; 'complete' pages had many internal links and internal broken links; 'complex' pages had higher Flesch readability scores; 'informative' pages had a high signal-to-noise ratio and low "diversity" (i.e., unique editors / total edits); 'consistent' pages were edited by many administrators and were older; 'current' pages had high "currency" (measure undefined in the master chart on p. 278, though probably the inverse of age in days); and 'volatile' pages took longer to be reverted back to regular content after incidents of vandalism.

These results show that, compared to Featured pages on Wikipedia, the Stargate genre of pages tended to be terser, shorter, and have fewer revisions, but were usually older. Preferential attachment effects in user behavior were confirmed to be a general feature of Wikipedia pages. Finally, if the exploratory factor analysis results are to be trusted, they assert, perhaps controversially, the following: what the current dissertation would call mass-agglomerated pages (§4.2.5) are more 'authoritative,' that 'complete' pages have many links going to other Wikipedia pages (cf. §4.3.10), that 'complex' pages have many syllables but few words and sentences (cf. §§4.4.3 and 4.4.4), that 'informative' pages have few "noise" words and few authors (cf. §4.4.3), and that 'consistent' pages are those that have existed long enough to come under the control of administrators (cf. §§4.3.1,

4.3.6, and 4.3.4). Though all of these findings may be true in some circumstances, as Stvilia notes, they are not very contextually "deep" (p. 185), but have been shown to vary significantly by genre (p. 184). To the degree that the current project studied them, these topics' morphologies in the Stargate genre of Wikipedia pages are presented in the sections numbered throughout this paragraph.

## 5.8 Conclusion

In this chapter, this dissertation's findings were re-organized according to the divisions most evident in the content of the sites under study, namely: by site size, by editorial model, by business model and organizational motivation, by characteristics unique to each site, and by characteristics common to all sites.

Small sites were found to contain lengthy contributions by young and devoted fans, who were concerned less with organizing and making accessible their content for the general public, and more with providing rich content for others like themselves. These sites may be the most likely candidates for library collection development, and their similar information architectures could be interesting to LIS researchers. That Wikia actor pages made by many authors were written in a longer/interpretive reviewer style, as well as Wikia's content connection to Wikipedia, could be of interest to linguists. Both the fact that most users seeking esoteric information probably did not come to the site via large search engines as well as Wikia's periodic updating schedule could be of interest to researchers of fan participation. Finally, their content orientation and marketing towards young Europeans could be of interest to marketing and cultural geography researchers.

Large sites were more made by/for the general public. Whereas IMDb contained little other than such popular public pages, on Wikipedia, only a small group of authors consistently worked on the most public pages, and most authors were older adults who worked on substantive summaries of the franchise's most thought-provoking topics. LIS researchers may find interesting these sites' reliance on user-made navigation structures as well as their higher accessibility quality than the smaller sites. Fan participation researchers and collection developers may wish to note these sites' continuous updating processes. Marketing and business researchers may wish to study their larger variety of advertisements and corporate connections than were present on the smaller sites. Linguists may wish to study the writing styles of the general vs. connoisseuring public in this genre. And a variety of social and psychological science areas would probably be interested in the detailed user profiles available from Wikipedia.

Editor-controlled sites consistently prioritized the editors' or business's interests over users'. Whereas IMDb seemed most oriented towards capitalizing on preferential attachment Web traffic, by taking in their share of general fan browsing behavior surrounding the franchise and licensing that shallow content to other large corporations, GateWorld's editors were journalistically motivated, prioritizing the giving of fair episode ratings, the publishing of their book review-like materials, especially for highly anticipated episodes (though GateWorld's users preferred reviews of more connoisseuring episodes), and being willing to re-publish fan writing as Letters to the Editor. The contrast resembled less vs. more editorially controlled print publications (e.g., classified ads vs. newspaper articles). LIS researchers may find the interface, information architectural, availability, and knowledge and data management techniques of these sites interesting, as may marketing re-

searchers of brand identity and reputation. Linguists and anyone wishing to use this content may wish to know that these sites' reading levels were around one grade level lower than the wiki sites'. Mass media, journalism, and business researchers also may find interesting the resemblance to editor-controlled print publications.

The wiki sites essentially exchanged the giving of freedom over manipulating content to users for users' loyalty and advertising/donation revenue, from which a number of consequences ensued. The first of these consequences was that users produced more advanced writing than was present on the editor-controlled sites. Second, users exhibited greater personal investment in the site's content. Third, users populated the wiki sites' pages with more diverse external links to other sites than were present on the editor-controlled sites, which primarily linked to their affiliates. Linguists may be interested in the higher quality of writing on wiki sites. Fan participation researchers will also find a wealth of signs of personal investment on these sites, deserving of closer interpretive and ethnographic investigation. Finally, business researchers should further examine wikis' linking practices, to determine if they are as biased towards non-profits as editor-controlled sites are towards corporations.

For-profit sites both displayed behaviors indicative of opportunism and shortcut-taking, employing whatever means that their users would tolerate in order to secure revenue and present a pristine branding image. Business researchers may be interested to know that these and several other common large corporate business practices were in evidence on these sites. By contrast, the non-profit sites offered substantial user protections, detailed information about themselves, and a sometimes-inconsistent branding image.

The three sites made entirely by fans (i.e., all except IMDb) demonstrated a common

set of page types, textual themes, and a focus on lengthy textual explications of complex themes. They also demonstrated a pattern contrary to the mass-agglomerative style, where encyclopedic pages about people and technologies, which had low PageRanks but many inlinks, were hubs for devoted fans wanting to create substantive content. Both of these findings could be of interest to fan participation researchers, and to collection developers.

The three ad-supported sites (i.e., all except Wikipedia) interestingly did not mention advertising in their mission statements, suggesting they may have viewed ads as a necessary evil. This may be of interest to marketing researchers.

Each site also possessed IQ characteristics that were unique to just that site. GateWorld was known for a variety of esoteric and especially devoted content, depictions of several types of academicians, along with standard ad links to Amazon and iTunes as well as IT ads that dominated entire pages. The fields of academia in question may be interested to see how they are represented, marketing researchers might find the IT page dominance behavior interesting, and anyone looking for esoteric information on this franchise may want to about this site. IMDb had ways of indicating different degrees of business relationships that it had with other companies, which may be of interest to marketing researchers. It also occasionally contained interpretive cultural reference sections, and title pages that became popular enough to garner editorial attention. The ability to distinguish editorial attention from general public fan attention via linguistic and network cues may be of interest to those researchers. Wikia contained substantive information in several predictable locations, which may be of interest to fan participation researchers, linguists, and librarians. Marketing researchers and the site's users may also benefit from knowing that the site's targetted ads always regarded other science fiction franchises and retail merchandise. Finally,

Wikipedia contained rare extreme fan behaviors, and only rare fundraising pleas from the founder. Fan participation, philosophy of science, and religious studies researchers may find these results interesting.

All sites shared several commonalities in content, information architecture and accessibility, and link structures. Six writing styles were present across all sites, and each often occurred in typical locations, which may be of interest to linguistic, narrative, fan participation, and media researchers. Bibliographers may be interested to know that a number of standard page types and data fields were in evidence. That the sites' organizational descriptions bore commonalities may be relevant to business researchers. Scholars of what constitutes 'research' may wish to know that all of the sites displayed at least an interpretive level of original research. Standard vendor categories emerged, being possibly of interest to marketing researchers. Finally, leisure and business researchers may find interesting that every site routinely directed users to its competitors, if its own content was inadequate.

LIS researchers may benefit from knowing that all of the sites possess only shallow navigation structures, and that accessibility errors were more common that HTML errors.

Also, the mass-agglomeration writing style was indicative of a common page type, where the general public descended upon pages with high PageRanks and amassed a large quantity of content, often with only rudimentary textual quality. Fan participation researchers, and anyone wanting superficial summary information about the franchise, could find this helpful.

Pages both within and between these sites displayed strong interlinking behavior, supports the choice of studying these four sites. Secondary to these strong ties were less frequent links to commercial outlets, perhaps because three of the sites under study had com-

mercial agendas, and weaker ties still to wikis and fansites. Linking patterns from northern European domains were also resemblant of commercial inlinking patterns, as were continental European domains with wiki and fansite inlinking patterns. These findings deserve further attention from mass media, cultural geography, and social network researchers.

Classification and HCI researchers may be interested to learn that evidence existed suggesting that, when trying to organize information, the general public may try all conceptual tools available to them with similar frequency.

Finally, compared with Stvilia's broader analysis of Wikipedia, Stargate pages on Wikipedia were terser, shorter, had fewer revisions, and were older. The preferential attachment effects observed on Stargate pages were also shown to be a common feature of Wikipedia in general.

# Chapter 6

# Conclusion

## 6.1 Summary of results

Like most empirical IQ research, this study was pragmatically motivated, and was empirically conducted in an exploratory rather than confirmatory manner – essentially subjecting these four sites to a barrage of techniques for characterizing IQ. Overall, this process found long lists of results, which were presented in the Discussion chapter, and which are summarized in appendix C, tables C.1-C.5. Differences among the sites primarily occurred along four lines: site size, editorial model, business model, and fan-made vs. non-fan-made sites. This section will summarize those differences, and offer an overall perspective.

In terms of **site size**, the smaller sites (GateWorld and Wikia) were more fan experience-oriented, and the larger sites (IMDb and Wikipedia) more oriented towards high-level overviews of the franchise. Smaller sites had topically focused navigation structures, they facilitated the contributions of devoted fans to obscure pages, they maintained large encyclopedias about peoples and technologies appearing in the franchise, they contained many

original interpretations of the franchise and its production processes, and, when they polled fans' levels of approval of aspects of the franchise (e.g., of individual episodes), those ratings were typically positively biased. Larger sites, by contrast, relied heavily on folksonomies and search engines for helping users find pages about the franchise, and they were often weak on interpretations, instead focusing on collecting production, biographical, and critical reception details. Within these trends, the wiki sites often referred to either each other or GateWorld for content that was more or less specific than each wiki site offered. Also, GateWorld's editors highly anticipated season premier episodes, whereas GateWorld's users more often connoisseuringly preferred normal or obscure episodes, and Wikia's users often appeared to be syndication viewers.

Regarding **editorial model**, the editor controlled sites (GateWorld and IMDb) were generally more self-serving, and the communally controlled sites (Wikia and Wikipedia) more community serving. Editors deferred users with questions to FAQ pages or bulletin boards, insisted that users sign over copyright to them, and wrote in whatever style of text they wished on whatever pages they wished. Though they rarely linked-to or quoted other sites, when they did, it was usually either to support claims they made or to endorse their affiliates. More perfunctorily, they also either manually or automatically monitored and organized content in a uniform manner across all pages on their sites, made little attempt to use the latest forms of media or international coding standards, and produced text at a lower average reading level than did the wikis. By contrast, communally controlled sites directed users with questions to several community discussion procedures/channels, allowed users to keep copyright, wrote most pages in a consistent style, and often linked-to a large variety and quantity of supplementary materials. More perfunctorily, communal

sites' code and media were newer and followed international standards, their texts were at a higher average reading level, and pages were updated either continuously (on Wikipedia) or following breaks in the academic calendar (on Wikia).

**Business models** caused IQ differences in openness and advertising between volunteer/non-profit sites (GateWorld and Wikipedia) and corporate sites (IMDb and Wikia). The more non-corporate was a site, the more likely it was to name its ownership, to detail its organizational structure and business affiliations, and to make its history and performance statements easy to find. Non-profit sites also had either fewer or no advertisements, and any ads they displayed were more on-topic and less about general retail than were commercial sites' ads.

Finally, **sites made primarily by fans** (including the wikis and GateWorld, the editors of which were also devoted fans) made more of an attempt to track every aspect of the franchise than did IMDb. While GateWorld's editors befriended the cast and crew in-person, the wikis focused on accumulating both cursory and cultural reference information from mostly documentary sources, with Wikia amassing esoteric and biometric information and Wikipedia generating more substantive textual descriptions of contexts and topics both within and surrounding the franchise. Interestingly, contrary to the IQ literature's expectations, the wikis' pages did not consistently contain more lists than did edited sites'; rather, larger sites' pages usually had more lists than smaller sites'. The three fan-made sites also often contained consistent themes, tried to accommodate both the general public and more devoted users in their episode pages, and often deferred cast and crew information to IMDb or similar large people-indexing sites. IMDb, by contrast, obtained most of its information from production companies, primarily tried to accommodate the general

public in its episode/title pages, and received cast and crew traffic from many other sites, traffic which it funnelled via links towards its own title pages, probably because that is the core of their content and advertising structure.

Taken together, these findings suggest that each site was probably suited for a different audience. IMDb presented the franchise from a corporate distance, in a manner probably only of interest to people wanting quick facts about an episode they saw when flipping through TV channels. Wikipedia provided a more thoughtful, yet still high-level, account of the franchise's main episodes, characters, and themes. This may be of most interest to casual adult fans of the franchise, who watch the shows often, but lack the time or interest to participate very heavily in a fan community. Wikia appeared to be most associated with college-aged fans who approach their interest in the franchise similarly to the way they might approach an online role-playing game, collaborating with an online community to create a kind of guidebook. Finally, GateWorld was by/for the truly devoted fans, people who wanted to spend large amounts of time and energy describing, dissecting, and extending the franchise.

## 6.2 Contributions of this study

### 6.2.1 Theoretical contributions

*Information quality* researchers perpetually seek both new criteria by which to evaluate information, as well as new understandings of the quality of information in different social contexts. In addition to the unique and contextually specific ways in which this project

operationalized IQ criteria present in the literature, several criteria that are either never or rarely considered by the literature were shown to be related to IQ. *Site size* can influence the types of users who are most likely to participate in creating online content, as well as the nature of their participation. *Editorial and business models* can constrain the types of content that are allowed to appear on a site. For example, *user profiling and targetting*, based on demographics or other user features, can result in typified themes recurring in advertisements or other auto-generated page content. *Writing styles*, measured by word usage, can reveal the type of documentation that an author population intends to create. Finally, facts about the *socio-cultural and geographical situatedness* of the content's authors can open a window into the forms of thought and expression that members of that context are likely to create. The labeled distinctions between fitness vs. representation, perceptual vs. artifactual fitness, and representational accuracy vs. completeness may also be officially unique to this dissertation, though the individual labels occur frequently throughout the IQ literature.

Additionally, all of this dissertation's findings regarding the fansites under study, which are listed in the following paragraphs, could be considered contributions to understanding the IQ of both fansites specifically, and communally created Web 2.0 content on the Internet generally.

*Library and information scientists* (LIS) may find the following of interest. Smaller sites possessed similar conceptual architectures and navigation labeling, whereas larger sites relied more on search engines and folksonomies. Larger sites were more accessible than smaller sites, and accessibility errors (WCAG 1.0) were more common all sites than were HTML validation errors. Editor-controlled sites used a number of interface, informa-

tion architectural, availability, and knowledge and data management techniques to remain in control of user-contributed content. All sites shared a number of standard page types and data fields. Finally, when trying to organize information, it appeared that the general public may try all conceptual tools available to them with similar frequency.

Related to LIS, *collection developers* may also be interested to know that smaller sites contained lengthier and more substantive (i.e., fan immersion-oriented) content, that large sites update their content continuously and smaller sites either periodically or sporadically, and editor-controlled sites' page texts were an average of one reading level grade lower than were wikis'. The fan participation research findings, listed two paragraphs below, may also be of interest.

*Linguists* would perhaps be most interested in there existing evidence for six common writing styles, each of which appeared to some degree on all of the sites. Wikia actor pages made by many authors were written in a longer/interpretive reviewer style, and the style of content on large sites appeared to differ depending on whether the authors were from the general public or were connoisseurs. Wikipedia pages also may have directed traffic to Wikia pages, and wiki authors displayed greater conformance to a standard/conventional sentence length.

*Fan participation* researchers may most want to study the many rich signs of personal investment that occurred on the wikis under study. Also, all of the only-fan-made sites (i.e., all except IMDb) had common sets of page types, textual themes, and a focus on lengthy textual explications of complex themes. They also often followed a pattern contrary to the mass-agglomerative writing style, where encyclopedic pages about people and technologies, which had low PageRanks but many inlinks, were hubs for devoted fans wanting

to create substantive content. On all sites, most mass-agglomerative pages also had high PageRanks, longer than average content, and lower than average quality writing. Furthermore, Wikia notably followed a periodic content updating schedule, which was in sync with breaks in most academic calendars, whereas Wikipedia's pages were updated fairly continuously, with a moderate increase of contributions during the summer. Finally, fans seeking esoteric information about the franchise on/from the smaller sites probably did not come to those sites from large search engines.

*Cultural geography* researchers will find a wealth of socio-cultural and demographic information in the wiki sites' public user profiles. Specifically, this study found that most Wikia users identified themselves as twenty-something, physically active, European students, whereas most Wikipedia users as older, more sedentary, American, and employed. There also existed a resemblance between northern European domain names and corporate inlinks to these sites, and between continental European domains and smaller wiki/fansite links.

*Business and marketing* researchers may find most interesting the distinctions found between editorially vs. communally controlled sites. Editorially controlled sites more tightly controlled their brand identities, evinced a biased for other corporations in their links to other sites, afforded users fewer copyright and privacy protections, and were less forthcoming about their business partnerships than were communally controlled sites. Larger sites also appeared to have more business partnerships, and more diverse advertisements. All sites had commonalities in their organizational descriptions, all had a standard of advertisement categories, and all directed users to their competitors, after the user had been given a chance to view the organization's content. The three ad-supported sites (i.e., all

except Wikipedia) did not mention advertising in their mission/passion statements. Finally, other than the strong hyperlink connections both within and between the four sites under study, links to commerce sites were next most frequent, followed by links to small fansites and wikis.

*Mass media and journalism* researchers may appreciate the resemblance of editorially controlled sites' practices to those of editorially controlled print publications.

*Philosophers* may wish to consider to what degree the forms of interpretive analysis in evidence on these sites constitutes 'research.'

Finally, *social network* researchers may wish to note the strong hyperlink ties both within and between the sites under study, as well as the two levels of external link frequencies mentioned in the business-related results, three paragraphs above.

### 6.2.2 Methodological contributions

As also noted in §3.4.3, and by Stvilia (2006); Stvilia et al. (2009), the IQ literature often uses heuristic and pair-wise methods when statistically principled and multivariate methods would be more appropriate. This dissertation demonstrated the use of a variety of exploratory statistical techniques either never-before or rarely seen in the IQ literature. These include: robust multiple regression, multivariate latent variable models (e.g., principal components analysis and canonical correlation), and generalized linear models (i.e., logistic regression on polytomous ordinal data). For more on these methods and their use in this dissertation, see §3.4.4.

## 6.3 Limitations and generalizability

### 6.3.1 Technical challenges

The following technical and logistical challenges hindered this project.

Two for-profit media companies attempted to either profit from or hinder this research. IMDb's restrictive data gathering policies both limited and made more manually laborious this study's characterizing of cast and crew pages, title pages, character pages, user demographics, and advertisements. Similarly, the Nielsen company's refusal to license its Stargate-related first-run and syndication ratings at a reasonable price necessitated the use of less precise, but publicly available, binary variables in those variables' places.

The availability of more content analytic coders may have made more reliable the analyses of textual themes, advertisements, and link/citation types.

Also, insufficient time and/or subject-specific expert collaborators were available to investigate the following: which fansites had the most up-to-date and authoritative news (i.e., currency as state-of-the-artness), to what extent errors persisted longer on wiki-produced pages than editorially produced pages (i.e., mass collaboration as editorially uneven), matching actor pages in the study of objectivity as impartiality (§4.3.4), and contextually relevant IQ criteria from literatures outside of the general IQ literature (e.g., narrative studies).

### 6.3.2 Limitations to generalizability

Additionally, the generalizability of the current study may be limited by the following factors.

Studying only sites about science fiction-related mass media impedes generalizing to sites about other types of mass media (e.g., sports or headline news). Similarly, sampling only Stargate-related pages means one cannot know for certain how representative these results may be of these four sites overall.

The analysis of only four relatively large sites ignores the "long tail" of very small fan-sites about this franchise. Based on anecdotal evidence, such sites included: the résumé-like sites maintained by cast and crew members, "shrine" (i.e., devotional) sites to individual cast and crew members, and Stargate-related gaming community sites. Though quite a few such sites occurred throughout the results of this dissertation (e.g., when links went between them and the large sites under study), there probably is additional relevant and interesting content on those sites. They were not studied, because there are literally thousands of such sites in existence, and each one would require idiosyncratic data collection and analysis techniques.

Also, this study was not longitudinal, but captured only several months in the histories of these sites. Though there is evidence to suggest that this franchise is fading in popularity (see §4.3.1), meaning that the sites' contents may have accumulated to a peak and may be stagnating, the histories and futures of the sites' developments were neither captured nor predicted. Regarding repeatability, archives of past versions of these sites are only provided by the wiki sites, and those only for a limited time. To repeat this analysis of the editor-controlled sites in the exact states that they were during this study, either a service such as Archive.org's WayBack Machine or this dissertation's author would need to be consulted.

GateWorld, though popular, was an idiosyncratic site run by fan volunteers. The owner's journalism education, personal ties to the studio and advertisers, and knowledge of the

larger sites are probably the most generalizing influences on that site. The site's choice of what types of content and advertising themes to include could have been somewhat unique. Similarly, Wikia's focus on retail ads could have been unique, and the reason for that focus remains unclear. Also, IMDb's unique focus on cast and crew pages may have been due to the "pro," fee-requiring, contact information-hosting service that it provides to/about members of the entertainment industry.

Finally, being donation-based, Wikipedia may not be as concerned with maintaining competitive site features and contents as are the other sites. Or, even if they are – as is perhaps evinced by their recent redesign initiative and responses to media criticism about their having many more male editors than female – they may not have effective means of herding their user population in different directions, unlike editor-controlled sites. Only administrative changes, such as redesigning their site's template, may be feasible.

### 6.3.3 Aids to generalizability

On the other hand, the generalizability of the current study to other contexts may be aided by the following factors.

These sites are situated within a marketplace, and their prominence requires that, to some degree, they must be concerned with maintaining competitive features and content. In support of this, authors and evidence cited on the pages of the sites studied for this project suggest an awareness, by both users and editors, of a variety of competitor sites. The current study also suggests that larger sites (e.g., IMDb and Wikipedia) may be more obligated by their highly visible social positions, corporate stakeholders, and scale-related technical

challenges to provide only summary or official information and consistent page templates, whereas smaller sites (e.g., GateWorld and Wikia) can afford fans more liberties. This can result in small sites gaining a reputation as being good sources of esoteric information. Closer to marketing, the preferential attachment effects of large sites on the PageRank algorithm, observed in this study, is a very general phenomenon, as are techniques by organizations and individuals to "optimize" (i.e., manipulate) PageRanks. However, the ways in which online communities drive Web traffic toward certain sites or pages without the involvement of search engines is more of a social scientific issue and is perhaps less well-understood. Finally, the marketing techniques by which large corporations place ads on large sites like these are probably quite standardized.

Generalizations can also be made, due to the Stargate pages that were studied being situated within broader organizational contexts. Many of the artifactual fitness-related criteria studied were found to be manifestations of technical, conceptual, or stylistic frameworks that are common to both Stargate and non-Stargate pages alike on Wikia, Wikipedia, and IMDb. Similarly, common information architectural techniques existed both between the two small sites and between the two large sites, and textual reading levels were shown to be tied to editorial model and could be associated with user demographics. Furthermore, the Stargate-related contents of each site were affected by the same organizational agendas, affiliations, funding sources, and office locations as the rest of each site. And, in many ways throughout this study, the ideological and business issues affecting these organizations' use of certain copyright and international standards, as well as whether to impose structures and agendas on users or to let structure emerge from users, stretch far beyond these organizations.

These organizations are also situated within broader social contexts. That these sites' linking strategies are tied to certain companies, demographics, and locations around the world suggests that other similar phenomena might be similarly tied, within social, cultural, geographical, etc. networks. The current study only looked one level out in several of these networks. Content creation patterns on these sites also often followed television broadcast and syndication patterns, suggesting that similar sites might also follow the mass media's timing, probably because of the centralized control that the media exercises over distribution. Similarly, Wikia's content modification patterns' following the schoolyear could be true of many things created by high school and college students.

Several findings suggested issues probably affecting many types of fandom. Perhaps most interesting is that the presence of normative site features, content themes, data fields, page types, and writing styles across all of the sites suggests that common approaches may be employed by sci-fi fans to understand/assimilate a franchise. For example, the social roles of summary documenters, commentators/reviewers, and extenders/extrapolators were observed. The usual *modus operandi* of sci-fi is that it presents people with an imagined environment, which explores, often as a form of social commentary, the possible effects of as-yet-unattained advancements in science and technology. People who encounter many such environments probably develop strategies for processing them – such as finding what in them is generic to many franchises or contemporary social movements, and finding what in them is unique. Also, the ways in which fans interact with each other (e.g., via forums and conventions) and with celebrities and studios (via blogs, chats, studio tours, etc.) are quite standardized. For their own security, privacy, and profit, celebrities and media corporations ensure that there are only so many avenues by which fans can investigate and

interact with the franchise. Websites like those studied here could be re-defining the bounds of celebrity-community engagement. Last, fan polls and Likert scales, like those seen in this study, are very common on many social networking sites and groupware applications used by large organizations, as are the phenomena of fans being positively biased in favor of the thing about which they are a fan, and of editors trying to remain neutral. Interesting differences were observed in this study between fan editors on GateWorld and Wikia vs. professional editors on IMDb. The fan editors struggled more to remain neutral, and the professional editors were more detached and profiteering.

Finally, many of the variables and methods used in this study are quite generalizable and repeatable. All of the studies conducted for this project began by evaluating IQ criteria prescribed by the literature, which often came from other fields. Statistically principled analysis techniques were also preferred over heuristic ones. Regarding repeatability, though obtaining the exact datasets used by this project would be difficult, it would be possible to obtain those datasets from several sources, and the same techniques as employed by this project could be used to obtain future versions of these or most any website. Last, the exhaustiveness and/or randomness of most of the samples used throughout this study minimized observational biases.

## 6.4 Related research directions

Metaphorically, a kind of wave (or funnel) seems to exist surrounding both technological adoption and popular description/interpretation of phenomena. Large corporations often attempt to position themselves at the front crest of the wave, in order to capitalize on what-

ever seems new to the public. This happens whether they are the source of the innovative technology or content, or whether they must acquire/purchase rights to the innovation. As a result, and as seen in this dissertation, the information that they produce is often possessively/restrictively biased towards their own interests, as well as focuses on the obscure, the studio-affiliated, and the most publicly profitable. That is, they ride the wave of the phenomenon's popularity for their own benefit.

As the wave diffuses (or funnel widens), semi-professionals, such as open-source developers or devoted fans, collaboratively replicate and adapt the technology or content into something more reflective of what many different target audiences want (i.e., the long tail). As seen on the wiki sites, their users often pay only token attention to the most superficial public-facing pages, focusing instead on amassing more substantive and detailed/immersive content and system features. Following this, at the most diffuse part of the wave/funnel, the technology or content becomes highly standardized and ubiquitous enough to be considered an infrastructural public service, like e-mail. Skype is perhaps a good example. Though many voice-over-IP (VoIP) technologies exist, that company identified peer-to-peer (P2P) calling as an important innovation, developed their own protocols and codecs for delivering that service, made it easy for novice users to use (and tightly controlled developer use), and locked users into their network, which is incompatible with all other VoIP voice, video, or chat software. They will ride this wave for as long as they are able, collecting subscription fees and ad revenue from locked-in users, until a viable and usable P2P VoIP solution is made by the open-source community, and its protocols begin becoming standardized and distributed throughout the marketplace, at which point Skype will need a new innovation, to remain viable. The movement from ICQ/AIM to

open-source Jabber/GTalk and chat services embedded in social networking sites (e.g., Facebook), along with the death of ICQ in the western world, is another good example.

As with the difference between the 1950s – when several television companies controlled all that Americans could watch on TV – and now, the trends of information and technological ubiquitization are certain to continue. Over time, media companies may find increasing difficulty in establishing an influential innovative niche, as the marketplace becomes ever-more based upon distributed services. The relevance of this to the current research is that the issue of to what degree the free-ified/standardized version of a service or of content captures and extends the value/quality of the earlier corporate version should keep recurring in new forms. Hence, establishing both general and contextually specific IQ criteria, in terms of which to assess the two sides' differences, should long remain valuable.

Broadly speaking, the current study contributes most to understanding the gradual democratization of the news, in this case news about the mass media. Editorially controlled sites are often affiliated with networks and studios, like a print newspaper's having established ties with local police, business, and government offices, such that important and politically correct information is sent to them (i.e., the news comes to them). By comparison, information that has been communally gleaned and accumulated from public sources by semi-professionals and amateurs, may be less filtered by the society's institutions. Being able to characterize such differences in contemporary information from different sources is extremely important, because both researchers and the public should be aware of the comparative benefits and biases/limitations of the information available to them.

# Bibliography

Abbott, V. (2000). Web page quality: Can we measure it and what do we find? A report of exploratory findings. *Journal of Public Health*, 22(2):191–197. Retrieved on 2008-11-25 from http://jpubhealth.oxfordjournals.org/cgi/reprint/22/2/191.pdf.

Abdulla, R., Garrison, B., Salwen, M., Driscoll, P., and Casey, D. (2002). The credibility of newspapers, television news, and online news. In *Association for Education in Journalism and Mass Communication Annual Convention, Miami Beach, FL.*

Acquisti, A. (2009). Nudging privacy: The behavioral economics of personal information. *IEEE Security and Privacy*, 7(6):82–85.

Adler, B. and de Alfaro, L. (2007). A content-driven reputation system for the Wikipedia. In *Proceedings of WWW*, volume 7, pages 261–270. Retrieved on 2009-03-07 from http://www.soe.ucsc.edu/~luca/papers/06/ucsc-crl-06-18.pdf.

Al-Hakim, L. (2007). Information quality factors affecting innovation process. *International Journal of Information Quality*, 1(2):162–176.

AmberlightHCI (2008). Fan psychology: Designing effective fans services online. Retrieved on 2009-05-08 from http://www.amber-light.co.uk/resources/whitepapers/designing_fan_services.pdf.

Anthony, D., Smith, S., and Williamson, T. (2005). Explaining quality in Internet collective goods:

Zealots and good samaritans in the case of Wikipedia. *MIT Innovation and Entrepreneurship Seminar*, 23:2006. Retrieved on 2009-03-07 from http://web.mit.edu/iandeseminar/Papers/Fall2005/anthony.pdf.

Barabási, A. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509.

Barnes, S. and Vidgen, R. (2003). Assessing the quality of a cross-national e-government Web site: A case study of the forum on strategic management knowledge exchange. Retrieved on 2008-11-25 from http://cq-pan.cqu.edu.au/david-jones/Reading/Conferences/HICSS36/DATA/ETEGM04.PDF.

Barnes, S. and Vidgen, R. (2006). Data triangulation and Web quality metrics: A case study in e-government. *Information & Management*, 43(6):767–777. Retrieved on 2008-11-25 from http://linkinghub.elsevier.com/retrieve/pii/S037872060600053X.

Basilevsky, A. (1994). *Statistical factor analysis and related methods: Theory and applications*. Wiley.

BCFilmCommission (2008). Industry profile. Retrieved on 2008-10-07 from http://www.bcfilmcommission.com/about_us/industry_profile.htm.

Beeler, S. and Dickson, L., editors (2006). *Reading Stargate SG-1*. IB Tauris.

Beredjiklian, P., Bozentka, D., Steinberg, D., and Bernstein, J. (2000). Evaluating the source and content of orthopaedic information on the Internet: The case of Carpal Tunnel Syndrome. *The Journal of Bone and Joint Surgery*, 82(11):1540–1540.

Berland, G., Morales, L., Elliott, M., Algazy, J., Kravitz, R., Broder, M., Kanouse, D., Muñoz, J., Hauser, J., Lara, M., et al. (2001). Evaluation of English and Spanish health information on the Internet. *RAND Foundation*, 12:2004.

Bernstam, E., Shelton, D., Walji, M., and Meric-Bernstam, F. (2005). Instruments to assess the

quality of health information on the World Wide Web: What can our patients actually use? *International Journal of Medical Informatics*, 74(1):13–19.

Bernstam, E. V., Walji, M. F., Sagaram, S., Sagaram, D., Johnson, C. W., and Meric-Bernstam, F. (2008). Commonly cited website quality criteria are not effective at identifying inaccurate online information about breast cancer. *Cancer*, 112(6):1206–1213. Retrieved on 2008-12-11 from http://dx.doi.org/10.1002/cncr.23308.

Bharosa, N., Janssen, M., Rao, H., and Lee, J. (2008). Adaptive Information Orchestration: Architectural Principles Improving Information Quality. In *Proceedings of Proceedings of the 5th International ISCRAM Conference*, pages 556–565.

Bizer, C. and Cyganiak, R. (2009). Quality-driven information filtering using the WIQA policy framework. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(1):1–10.

Björnsson, C. H. (1968). *Läsbarhet*. Stockholm: Liber.

Björnsson, C. H. (1983). Readability of newspapers in 11 languages. *Reading Research Quarterly*, 18(4):484.

Blumenstock, J. (2008). Size matters: Word count as a measure of quality on Wikipedia. Retrieved on 2009-03-07 from http://www2008.org/papers/pdf/p1095-jblumenstock.pdf.

Bogenschutz, M. (2000). Drug information libraries on the Internet. *Journal of psychoactive drugs*, 32(3):249–258.

Bradley, A. (2005). Once a sleuth, always a sleuth: A study of fan publications as secondary materials for research in girls series books. Retrieved on 2009-03-07 from http://etd.ils.unc.edu/dspace/bitstream/1901/224/1/alisonbradley.pdf.

Breul, H., Boue, L., and Martin, A. (1999). Overview of Internet information resources of interest to French hospitals. *Semaine des Hopitaux*, (75):257–262.

Buhi, E., Daley, E., Oberne, A., Smith, S., Schneider, T., and Fuhrmann, H. (2010). Quality and Accuracy of Sexual Health Information Web Sites Visited by Young People. *Journal of Adolescent Health*.

Burgess, M., Gray, W., and Fiddian, N. (2007). Using quality criteria to assist in information searching. *International Journal of Information Quality*, 1(1):83–99.

Burkell, J. (2004). Health information seals of approval: What do they signify? *Information, Communication & Society*, 7(4):491–509. Retrieved on 2009-03-07 from `http://www.ingentaconnect.com/content/routledg/rics/2004/00000007/00000004/art00004`.

Butler, B., Joyce, E., and Pike, J. (2008). Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in Wikipedia. Retrieved on 2009-03-26 from `http://www.katzis.org/wiki/images/7/76/Butleretal2008.pdf`.

Calero, C., Ruiz, J., and Piattini, M. (2004). A Web metrics survey using WQM. *Lecture Notes in Computer Science*, pages 147–160.

Chakrabarti, S. (2003). *Mining the Web: Discovering knowledge from hypertext data*. Morgan Kaufmann.

Charnock, D. and Shepperd, S. (2004). Learning to DISCERN online: Applying an appraisal tool to health websites in a workshop setting. Retrieved on 2009-03-07 from `http://her.oxfordjournals.org/cgi/content/full/19/4/440`.

Chen, L., Minkes, R., and Langer, J. (2000). Pediatric surgery on the Internet: Is the truth out there? *Journal of Pediatric Surgery*, 35(8):1179–1182.

Clement, W., Wilson, S., and Bingham, B. (2002). A guide to creating your own patient-oriented website. *JRSM*, 95(2):64.

Cline, R. and Haynes, K. (2001). Consumer health information seeking on the Internet: The state

of the art. *Health Education Research*, 16(6):671–692.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Earlbaum Associates, 2nd edition.

Coleman, B. (2003). Producing an information leaflet to help patients access high quality drug information on the Internet: A local study. *Health Information & Libraries Journal*, 20(3):160.

Coleman, M. and Liau, T. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Collins, J. (2006). An investigation of web-page credibility. *Journal of Computing Sciences in Colleges*, 21(4):16–21.

Comfrey, A. and Lee, H. (1992). *A first course in factor analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates, 2nd edition.

Costello, V. and Moore, B. (2007). Cultural outlaws: An examination of audience activity and online television fandom. *Television & New Media*, 8(2):124. Retrieved on 2009-03-07 from `http://tvn.sagepub.com/cgi/content/abstract/8/2/124`.

Coulter, A., Ellins, J., Swain, D., Clarke, A., Heron, P., Rasul, F., MAGEE, H., and SHELDON, H. (2006). Assessing the quality of information to support people in making decisions about their health and healthcare. *Picker Institute Europe, Oxford*. Retrieved on 2009-03-07 from `http://www.pickereurope.org/Filestore/Downloads/Health-information-quality-web-version-FINAL.pdf`.

Crawley, M. (2002). *Statistical computing: an introduction to data analysis using S-Plus*. John Wiley & Sons Inc.

Cronin, B. (2005). BLOG: See also Bathetically Ludicrous Online Gibberish. Retrieved on 2008-

12-02 from http://www.slis.indiana.edu/news/story.php?story_id=958.

D'Alessandro, D., Kingsley, P., and Johnson-West, J. (2001). The readability of pediatric patient education materials on the World Wide Web. *Archives of Pediatrics and Adolescent Medicine*, 155(7):807–812.

Davison, K. (1997). The quality of dietary information on the World Wide Web. *Clin Perform Qual Health Care*, 5(2):64–6.

Diering, C. and Palmer, M. (2001). Professional information about urinary incontinence on the World Wide Web: Is it timely? Is it accurate? *Journal of Wound, Ostomy and Continence Nursing*, 28(1):55.

DISCERNonline (2009). The DISCERN instrument. Retrieved on 2009-03-12 from http://www.discern.org.uk/discern_instrument.php.

Doupi, P. and Van Der Lei, J. (1999). Rx medication information for the public and the WWW: Quality issues. *Medical Informatics & The Internet in Medicine*, 24(3):171–179.

Dragulanescu, N. (2002). Website quality evaluations: Criteria and tools. *International Information and Library Review*, 34(3):247–254.

Dutta-Bergman, M. (2004). The impact of completeness and Web use motivation on the credibility of e-health information. *The Journal of Communication*, 54(2):253–269.

Ekman, A., Hall, P., and Litton, J. (2005). Can we trust cancer information on the Internet? A comparison of interactive cancer risk sites. *Cancer Causes & Control*, 16(6):765–772.

Elliott, J. (2001). Copyright fair use and private ordering: Are copyright holders and the copyright law fanatical for fansites. *DePaul-LCA Journal of Art and Entertainment Law*, 11:329.

Emerson, J. and Hoaglin, D. (2006). *Resistant multiple regression, one variable at a time*, pages 241–280. Wiley Interscience.

Eppler, M., Algesheimer, R., and Dimpfel, M. (2003). Quality criteria of content-driven websites

and their influence on customer satisfaction and loyalty: An empirical test of an information quality framework. In *Proceeding of the Eighth International Conference on Information Quality*, pages 108–120.

Eppler, M. and Muenzenmayer, P. (2002). Measuring information quality in the Web context: A survey of state-of-the-art instruments and an application methodology. In *Proceedings of the 7th International Conference on Information Quality*, pages 187–196.

Estrada, C., Hryniewicz, M., Higgs, V., Collins, C., and Byrd, J. (2000). Anticoagulant patient information material is written at high readability levels. *Stroke*, 31(12):2966–2970.

Everitt, B. (2005). *An R and S-PLUS companion to multivariate analysis*. Springer Verlag.

Eysenbach, G. and Jadad, A. (2001). Evidence-based patient choice and consumer health informatics in the Internet age. *Journal of Medical Internet Research*, 3(2). Retrieved on 2009-03-07 from `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1761898`.

Eysenbach, G., Powell, J., Kuss, O., and Sa, E. (2002). Empirical studies assessing the quality of health information for consumers on the World Wide Web: A systematic review. *Journal of the American Medical Association*, 287(20):2691–2700. Retrieved on 2009-03-07 from `http://www.paho.org/{English}/DD/IKM/Eysenbach2002c-jama-sysrev.pdf`.

Fallis, D. (2008). Toward an epistemology of Wikipedia. *Journal of the American Society for Information Science and Technology*, 59(10):1662–74. Retrieved on 2008-11-25 from `http://dlist.sir.arizona.edu/2400/01/FallisWikipediaJASIST.pdf`.

Fallis, D. and Frické, M. (2002). Indicators of accuracy of consumer health information on the Internet: A study of indicators relating to information for managing fever in children in the home. *Journal of the American Medical Informatics Association*, 9(1):73–79.

Feng, X. and Liu, Y. (2008). A Study on Evaluation Model of Information Sharing Quality in Virtual Teams. In *Proceedings of the 2008 International Conference on Computer Science*

*and Software Engineering-Volume 05*, pages 117–120. IEEE Computer Society.

Fisher, C. and Kingma, B. (2001). Criticality of data quality as exemplified in two disasters. *Information & Management*, 39(2):109–116.

Fisher, C. W., Lauria, E. J. M., and Matheus, C. C. (2009). An accuracy metric: Percentages, randomness, and probabilities. *J. Data and Information Quality*, 1(3):1–21.

Fiske, J. and Lewis, L. (1992). *The adoring audience: Fan culture and popular media*. Routledge.

Fitzmaurice, D. and Adams, J. (2000). A systematic review of patient information leaflets for hypertension. *Journal of Human Hypertension*, 14(4):259–262.

Fleiss, J. et al. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3):221.

Flesch, R. (1962). *The art of readable writing*. MacMillan Publishing Company.

Frické, M. and Fallis, D. (2004). Indicators of accuracy for answers to ready reference questions on the Internet. *Journal of the American Society for Information Science and Technology*, 55(3):238–245.

Fritch, J. and Cromwell, R. (2001). Evaluating Internet resources: Identity, affiliation, and cognitive authority in a networked world. *Journal of the American Society for Information Science and Technology*, 52(6):499–507.

Fritz, M. and Schiefer, G. (2003). Identification and evaluation of Internet resources for agribusiness information needs. *Atti della*, 4.

Fry, E. (1968). A readability formula that saves time. *Journal of Reading*, pages 513–578.

Fry, E. (1977). Frys readability graph: Clarifications, validity, and extension to level 17. *Journal of Reading*, 21(3):242–252.

Fulda, P. and Kwasik, H. (2004). Consumer health information provided by library and hospital

websites in the South Central Region. *Journal of the Medical Library Association*, 92(3):372.

Gagliardi, A. and Jadad, A. (2002). Examination of instruments used to rate quality of health information on the Internet: Chronicle of a voyage with an unclear destination. *British Medical Journal*, 324(7337):569–573.

Galimberti, A. and Jain, S. (2000). Gynaecology on the Net: Evaluation of the information on hysterectomy contained in health-related Web sites. *Journal of Obstetrics & Gynaecology*, 20(3):297–299.

Garson, D. (2008). Factor analysis: Statnotes. Retrieved on 2008-03-22 from http://faculty.chass.ncsu.edu/garson/PA765/factor.htm.

Gasher, M. (2002). *Hollywood North: The feature film industry in British Columbia*. UBC Press.

GateWorld (2009a). Advertising with gateworld. Retrieved on 2009-03-12 from http://gateworld.net/advertising_with_gateworld.shtml.

GateWorld (2009b). Contribute to gateworld. Retrieved on 2009-03-12 from http://gateworld.net/news/contribute-to-gateworld/.

GateWorld (2009c). Privacy policy. Retrieved on 2009-03-12 from http://gateworld.net/privacy_policy.shtml.

GateWorld (2009d). Site history. Retrieved on 2009-03-12 from http://gateworld.net/site_history.shtml.

GateWorld (2009e). Site staff. Retrieved on 2009-03-21 from http://gateworld.net/site_staff.shtml.

GateWorld (2009f). Write to us. Retrieved on 2009-03-12 from http://gateworld.net/write_to_us.shtml.

GateWorld (2010a). Alpha site. Retrieved on 2010-06-16 from http://www.gateworldalphasite.com/.

GateWorld (2010b). Sci fi & syndication ratings. Retrieved on 2010-04-08 from http://gateworld.net/sci_fi__syndication_ratings.shtml.

GateWorld (2010c). Tv schedule (syndication). Retrieved on 2010-06-02 from http://gateworld.net/tv_schedule_syndication.shtml.

Gaudinat, A., Grabar, N., and Boyer, C. (2007). Machine learning approach for automatic quality criteria detection of health web pages. *Studies in Health Technology and Informatics*, 129(Pt 1):705–709.

Gelder, K. (2004). *Popular fiction: The logics and practices of a literary field*. Routledge.

Gillois, P., Colombet, I., Dreau, H., Degoulet, P., and Chatellier, G. (1999). A critical appraisal of the use of Internet for calculating cardiovascular risk. In *Proceedings of the American Medical Informatics Association Symposium*, volume 775, page 9.

Goodman, L. (1961). Snowball sampling. *The Annals of Mathematical Statistics*, pages 148–170.

Gordon, J., Barot, L., Fahey, A., and Matthews, M. (2001). The Internet as a source of information on breast augmentation. *Plastic and Reconstructive Surgery*, 107(1):171.

Green, C., Kazanjian, A., and Helmer, D. (2004). Informing, advising, or persuading? An assessment of bone mineral density testing information from consumer health websites. *International Journal of Technology Assessment in Health Care*, 20(02):156–166.

Griffiths, K. and Christensen, H. (2000). Quality of Web based information on treatment of depression: Cross sectional survey. *British Medical Journal*, 321(7275):1511–1515.

Griffiths, K. M. and Christensen, H. (2005). Website quality indicators for consumers. *Journal of Medical Internet Research*, 7(5):e55.

Gunning, R. (1969). The Fog index after twenty years. *Journal of Business Communication*, 6(2):3.

Habing, B. (2005). R templates. Retrieved on 2010-01-27 from http://www.stat.sc.edu/~habing/courses/530rF05.html.

Haddow, G. (2003). Focusing on health information: How to assess information quality on the Internet. *Australian Library Journal*, 52(2):169–178.

Halaris, C., Magoutas, B., Papadomichelaki, X., and Mentzas, G. (2007). Classification and synthesis of quality approaches in e-government services. *Internet Research*, 17(4):378–401.

Hanif, F., Abayasekara, K., Willcocks, L., Jolly, E. C., Jamieson, N. V., Praseedom, R. K., Goodacre, J. A., Read, J. C., Chaudhry, A., and Gibbs, P. (2007). The quality of information about kidney transplantation on the world wide web. *Clinical Transplantation*, 21(3):371–376.

Harland, J. and Bath, P. (2007). Assessing the quality of websites providing information on multiple sclerosis: Evaluating tools and comparing sites. *Health Informatics Journal*, 13(3):207–221.

Harrell, F. (2010). R design library. Retrieved on 2010-03-02 from `http://lib.stat.cmu.edu/S/Harrell/library/r/`.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27:83–85. 10.1007/BF02985802.

Häyrinen, K., Saranto, K., and Nykänen, P. (2008). Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *International Journal of Medical Informatics*, 77(5):291–304.

HealthOnTheNet (2008). Code of conduct. Retrieved on 2009-02-27 from `http://www.hon.ch/HONcode/Conduct.html`.

Heinrich, B., Klier, M., and Kaiser, M. (2009). A procedure to develop metrics for currency and its application in crm. *J. Data and Information Quality*, 1(1):1–28.

Herring, S. (2007). Web(log) content analysis: Expanding the paradigm. In *The International Handbook of Internet Research. Springer Verlag*.

Hinman, M. (2006). Venture capitalists invest wiki-millions. *Tampa Bay Business Journal*. Retrieved on 2006-03-10 from `http://tampabay.bizjournals.com/tampabay/stories/2006/03/13/story1.html`.

Hoaglin, D., Mosteller, F., and Tukey, J. (1983). *Understanding robust and exploratory data analysis*. Wiley.

Hoaglin, D., Mosteller, F., and Tukey, J., editors (2006). *Exploring Data Tables, Trends, and Shapes*. Wiley.

Hoffman-Goetz, L. and Clarke, J. N. (2000). Quality of breast cancer sites on the World Wide Web. *Canadian Journal of Public Health*, 91(4):281–284.

HollywoodNorthFilmNet (2008). BC film industry. Retrieved on 2008-10-07 from `http://www.hollywoodnorthpr.com/industry.html`.

Hong, T. (2005). The influence of structural and message features on Web site credibility. *Journal of the American Society for Information Science and Technology*, 57(1):114–127.

Howitt, A., Clement, S., de Lusignan, S., Thiru, K., Goodwin, D., and Wells, S. (2002). An evaluation of general practice websites in the UK. *Family Practice*, 19(5):547–556.

Hsieh, F., Bloch, D., and Larsen, M. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine*, 17(14):1623–1634.

Hu, M., Lim, E., Sun, A., Lauw, H., and Vuong, B. (2007). Measuring article quality in Wikipedia: Models and evaluation. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, pages 243–252.

Huang, J. Y. J., Discepola, F., Al-Fozan, H., and Tulandi, T. (2005). Quality of fertility clinic websites. *Fertility and Sterility*, 83(3):538–544.

Huh, J. and Cude, B. (2004). Is the information "fair and balanced" in direct-to-consumer prescription drug websites? *Journal of Health Communication*, 9(6):529–540.

Ilic, D., Risbridger, G., and Green, S. (2004). Searching the Internet for information on prostate cancer screening: An assessment of quality. *Urology*, 64(1):112–116.

IMDb (2009a). Alternate interfaces. Retrieved on 2009-03-17 from http://www.imdb.com/interfaces.

IMDb (2009b). Authoritative and accurate information about movies & television. Retrieved on 2009-03-17 from http://www.imdb.com/licensing/.

IMDb (2009c). Can i use imdb data in my software? Retrieved on 2009-03-17 from http://www.imdb.com/help/show_leaf?usedatasoftware.

IMDb (2009d). Copyright and conditions of use. Retrieved on 2009-03-12 from http://www.imdb.com/help/show_article?conditions.

IMDb (2009e). How/where [sic] you get your information? How accurate/reliable is it? Retrieved on 2009-03-12 from http://www.imdb.com/help/show_leaf?infosource.

IMDb (2009f). How/where you get your information? how accurate/reliable is it? Retrieved on 2009-03-17 from http://www.imdb.com/help/show_leaf?infosource.

IMDb (2009g). Imdb advertising. Retrieved on 2009-03-17 from http://www.imdb.com/advertising/.

IMDb (2009h). Imdb privacy notice. Retrieved on 2009-03-12 from http://www.imdb.com/privacy.

IMDb (2009i). Submission guides. Retrieved on 2009-03-12 from http://www.imdb.com/Guides/.

IMDb (2009j). Updating information for (title). Retrieved on 2009-03-17 from http://www.imdb.com/updates. The page must be accessed by clicking the ”Update” link at the bottom of any IMDb page about a film or television program. Free registration is required.

IMDb (2009k). Using imdb content for non-commercial purposes. Retrieved 2009-03-17 from

http://www.imdb.com/licensing/noncommercial.

IMDb (2009l). What is the internet movie database? Retrieved 2009-03-17 from http://www.imdb.com/help/show_leaf?about.

IMDb (2010a). Contributor zone: Top contributor messages. Retrieved on 2010-05-01 from http://www.imdb.com/czone/top_msg.

IMDb (2010b). Imdb history. Retrieved on 2010-05-01 from http://www.imdb.com/help/show_leaf?history.

IMDb (2010c). Imdb pro. Retrieved on 2010-06-16 from https://secure.imdb.com/signup/v4/?d=IMDbTab.

InternetArchive (2009). Homepage. Retrieved on 2009-03-21 from http://www.archive.org/.

Jakobsson, A. and Giversen, J. (2009). Guidelines for Implementing the ISO 19100 Geographic Information Quality Standards in National Mapping and Cadastral Agencies. *Guidelines_ISO19100_Quality. pdf Accessed February*.

JDIQ (2010). Journal of data and information quality: About. Retrieved on 2010-06-14 from http://jdiq.acm.org/.

Jenkins, H. (1992). *Textual poachers: Television fans and participatory culture*. Routledge.

Kahn, B., Strong, D., and Wang, R. (2002). Information quality benchmarks: Product and service performance. *Communications of the ACM*, 45(4):184–192.

Karaban, Y. (2009). Www-google-pagerank. Retrieved on 2009-07-22 from http://search.cpan.org/dist/WWW-Google-PageRank/.

Kasal, P., Janda, A., Feberova, J., Adla, T., Hladikova, M., Naidr, J., and Potuckova, R. (2005). Evaluation of health care related web resources based on web citation analysis and other quality criteria. In *Proceedings of the 27th Annual International Conference of the Engineering in Medicine and Biology Society, IEEE-EMBS 2005*, pages 2391–2394.

Katerattanakul, P. and Siau, K. (1999). Measuring information quality of Web sites: Development of an instrument. In *International Conference on Information Systems*, pages 279–285. Association for Information Systems.

Kihlstrom, L. (2001). Evaluating pharmacy benefit management information on the Internet: Purpose, structure, technology, and content. *Managed Care Interface*, 14(5):64–8.

Kim, P., Eng, T., Deering, M., and Maxfield, A. (1999). Published criteria for evaluating health related Web sites: Review. *British Medical Journal*, 318(7184):647–649.

Knight, S. and Burn, J. (2005). Developing a framework for assessing information quality on the World Wide Web. *Informing Science Journal*, 8(3):159–172.

Kraakman, R., Hansmann, H., Davies, P., Hertig, G., Hopt, K., Kanda, H., and Rock, E. (2004). *The anatomy of corporate law: a comparative and functional approach*. Oxford University Press, USA.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Sage Publications, Inc.

Kunst, H., Groot, D., Latthe, P., Latthe, M., and Khan, K. (2002). Accuracy of information on apparently credible websites: Survey of five common health topics. *British Medical Journal*, 324(7337):581–582.

Kunst, H. and Khan, K. (2002). Quality of Web-based medical information on stable COPD: Comparison of non-commercial and commercial websites. *Health Information and Libraries Journal*, 19(1):42–48.

Lacroix, E., Backus, J., and Lyon, B. (1994). Service providers and users discover the Internet. *Bulletin of the Medical Library Association*, 82(4):412.

Langille, M., Rodgers, C., Bernard, A., and van Zanten, S. (2006). A systematic review of the quality of patient information on medical treatment of Crohn's disease and ulcerative colitis

on the World Wide Web. *Gastroenterology*, 130(4, Suppl. 2):A620. 20081210 ILLed.

Latthe, M., Latthe, P., and Charlton, R. (2000a). Quality of information on emergency contraception on the Internet. *British Journal of Family Planning*, 26(1):39–43.

Latthe, P., Latthe, M., and Khan, K. (2000b). Quality of information on female sterilisation on the Internet. *Journal of Obstetrics & Gynaecology*, 20(2):167–170.

Latthe, P., Latthe, M., and Khan, K. (2000c). Quality of medical information about menorrhagia on the WorldWide Web. *BJOG: An International Journal of Obstetrics & Gynaecology*, 107(1):39–43.

Lemon, J. and Fellows, I. (2009). Concordance and reliability: 'concord package for r.

Leung, H. (2001). Quality metrics for intranet applications. *Information & Management*, 38(3):137–152.

Lewiecki, E., Rudolph, L., Kiebzak, G., Chavez, J., and Thorpe, B. (2006). Assessment of osteoporosis-website quality. *Osteoporosis International*, 17(5):741–752.

Lih, A. (2003). Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource. *Nature*, 2004.

Lindemans, M. (2010). Encyclopedia mythica. Retrieved on 2010-04-24 from `http://www.pantheon.org/`.

Lissman, T. and Boehnlein, J. (2001). A critical review of Internet information about depression. *Psychiatric Services*, 52(8):1046.

Llinas, G., Rodriguez-Inesta, D., Mira, J. J., Lorenzo, S., and Aibar, C. (2008). A comparison of websites from Spanish, American and British hospitals. *Methods of Information in Medicine*, 47(2):124–130.

López-Ornelas, M., Cordero, G., and Backhoff, E. (2005). Measuring the quality of electronic journals. *Electronic Journal of Information Systems Evaluation*, 8(2):133–142.

MacCallum, R., Widaman, K., Zhang, S., and Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4:84–99.

Madnick, S. E., Wang, R. Y., Lee, Y. W., and Zhu, H. (2009). Overview and framework for data and information quality research. *J. Data and Information Quality*, 1(1):1–22.

Magnus, P. (2006). Epistemology and the Wikipedia. In *North American Computing and Philosophy Conference*.

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate analysis*. New York: Academic Press.

Marriott, J. V., Stec, P., El-Toukhy, T., Khalaf, Y., Braude, P., and Coomarasamy, A. (2008). Infertility information on the World Wide Web: A cross-sectional survey of quality of infertility information on the Internet in the UK. *Human Reproduction*, 23(7):1520–1525.

Martinez-Lopez, J. and Ruiz-Crespo, E. (1998). Internet. Quality of information available on orthopaedic surgery and traumatology. *Revista de ortopedia y traumatología(Madrid)*, 42(6):469–473.

McClung, H., Murray, R., and Heitlinger, L. (1998). The Internet as a source for current patient information. *Pediatrics*, 101(6).

McInerney, C. and Bird, N. (2005). Assessing website quality in context: Retrieving information about genetically modified food on the Web. *Information Research*, 10(2):10–2.

McKee, A. (2004). How to tell the difference between production and consumption: A case study in Doctor Who fandom. In Gwenllian-Jones, S. and Pearson, R., editors, *Cult television*.

McKemmish, S., Manaszewicz, R., Burstein, F., and Fisher, J. (2009). Consumer empowerment through metadata-based information quality reporting: The Breast Cancer Knowledge Online Portal. *Journal of the American Society for Information Science and Technology*, 60(9):1792 – 1807.

McLaughlin, G. (1969). SMOG grading: A new readability formula. *Journal of Reading*,

12(8):639–646.

McMoneagle, J. (2006). *Memoirs of a Psychic Spy: The Remarkable Life of U.S. Government Remote Viewer 001*. Hampton Roads Publishing.

Meric, F., Bernstam, E., Mirza, N., Hunt, K., Ames, F., Ross, M., Kuerer, H., Pollock, R., Musen, M., and Singletary, S. (2002). Breast cancer on the World Wide Web: Cross sectional survey of quality of information and popularity of websites. *British Medical Journal*, 324(7337):577–581.

Meyer, B. (2006). Defense and illustration of wikipedia.

Michnik, J. and Lo, M. (2009). The assessment of the information quality with the aid of multiple criteria analysis. *European Journal of Operational Research*, 195(3):850–856.

Miettinen, M. and Korhonen, M. (2008). Information Quality in Healthcare: Coherence of Data Compared between Organization's Electronic Patient Records. pages 488–493.

Moore, J. (2001). Copyright protection or fan loyalty: Must entertainment companies choose? Alternate solutions for addressing Internet fan sites. *North Carolina Journal of Law & Technology*, 3:273.

Morahan-Martin, J. and Anderson, C. (2000). Information and misinformation online: Recommendations for facilitating accurate mental health information retrieval and evaluation. *CyberPsychology & Behavior*, 3(5):731–746.

Moran, M. and Oliver, C. W. (2007). Content and design of patient-targeted websites in orthopaedic surgery: The example of total hip replacement. *Annals of the Royal College of Surgeons of England*, 89(8):773–776.

Moskowitz, S. (1990). The origins of science fiction fandom: A reconstruction. In *Foundation: The Review of Science Fiction*, volume 48, pages 5–25.

Moult, B., Franck, L., and Brady, H. (2004). Ensuring quality information for patients: Develop-

ment and preliminary validation of a new instrument to improve the quality of written health care information. *Health Expectations*, 7(2):165–175.

Moustakides, G. V. and Verykios, V. S. (2009). Optimal stopping: A record-linkage approach. *J. Data and Information Quality*, 1(2):1–34.

Murphy, P., Chesson, A., Berman, S., Arnold, C., and Galloway, G. (2001). Neurology patient education materials: Do our educational aids fit our patients' needs? *Journal of Neuroscience Nursing*, 33(2):99–104.

Nandi, K., Zhang, Y., and Migranov, N. (2004). A Semiclassical ANEC Constraint On Classical Traversable Lorentzian Wormholes. *Arxiv preprint gr-qc/0409053*.

Naumann, F. and Rolker, C. (2000). *Assessment methods for information quality criteria*. Professoren des Inst. für Informatik. Retrieved on 2009-03-07 from `http://www.ida.ing.tu-bs.de/academics/seminars/archiv/downloads/ss2004/Yamen_Seminar_SS04.pdf`.

Netcraft (2010a). Uptime summary for en.wikipedia.org. Retrieved on 2010-06-16 from `http://uptime.netcraft.com/up/graph?site=en.wikipedia.org`.

Netcraft (2010b). Uptime summary for gateworld.net. Retrieved on 2010-06-16 from `http://uptime.netcraft.com/up/graph?site=gateworld.net`.

Netcraft (2010c). Uptime summary for imdb.com. Retrieved on 2010-06-16 from `http://uptime.netcraft.com/up/graph?site=imdb.com`.

Netcraft (2010d). Uptime summary for wikia.com. Retrieved on 2010-06-16 from `http://uptime.netcraft.com/up/graph?site=wikia.com`.

Ogushi, Y. and Tatsumi, H. (2000). *Research on the analysis of the current state of the provision and use of health information provided through a new technology medium*. Health & Welfare Ministry Research Group : Tokyo, Japan.

Pandolfini, C. and Bonati, M. (2002). Follow up of quality of public oriented health information on the World Wide Web: Systematic re-evaluation. *British Medical Journal*, 324(7337):582–583.

Paolillo, J. (2002). *Analyzing linguistic variation: Statistical models and methods*. CSLI Publications: Stanford, CA.

Pearson, K. (1901). Mathematical contributions to the theory of evolution. ix: On the principle of homotyposis and its relation to heredity, to variability of the individual, and to that of race. part i: Homotyposis in the vegetable kingdom. *Philosophical Transactions of the Royal Society (London), Series A*, (197):285–379.

Pérez-López, F. and Roncero, G. (2006). Assessing the content and quality of information on the treatment of postmenopausal osteoporosis on the World Wide Web. *Gynecological Endocrinology*, 22(12):669–675.

Pink, D. (2005). The book stops here. *Wired News*. Retrieved on 2007-08-01 from `http://www.wired.com/wired/archive/13.03/wiki.html`.

Pinski, G. and Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing and Management*, 12(5):297–312.

Preucel, R. (2006). *Archaeological Semiotics*. Blackwell.

Primack, D. (2007). Pe week wire. *Private Equity Week*. Retrieved on 2007-01-03 from `http://www.pewnews.com/story.asp?sectioncode=44\&storycode=41174`.

Quinn, K. (2010). R example: ordered logistic regression. Retrieved on 2010-03-02 from `http://www.stat.washington.edu/quinn/classes/536/S/polrexample.html`.

Raggett, D. (2009). Html tidy. Retrieved on 2009-03-21 from `http://tidy.sourceforge.net/`.

Rahnavardi, M., Arabi, M. S. M., Ardalan, G., Zamani, N., Jahanbin, M., Sohani, F., and Dowlat-

shahi, S. (2008). Accuracy and coverage of reproductive health information on the Internet accessed in English and Persian from Iran. *Journal of Family Planning and Reproductive Health Care*, 34(3):153–157.

Rea, L. and Parker, R. (2005). *Designing and conducting survey research: A comprehensive guide*. Jossey-Bass.

Rice, J. (1995). *Mathematical Statistics and Data Analysis*. Duxbury Press, second edition.

Ricoeur, P. (1976). *Interpretation theory: Discourse and the surplus of meaning*. Texas Christian UP.

RNIB (2008). Web accessibility checklists. Retrieved on 2008-03-29 from `http://www.rnib.org.uk/xpedio/groups/public/documents/publicwebsite/public_checklists.hcsp`.

Robbins, S. and Stylianou, A. (2003). Global corporate Web sites: An empirical investigation of content and design. *Information & Management*, 40(3):205–212.

Rolland, Y., Bousquet, C., Pouliquen, B., Beux, P. L., Fresnel, A., and Duvauferrier, R. (2000). Radiology on Internet: Advice in consulting websites and evaluating their quality. *European Radiology*, 10(5):859–866.

Rosenzweig, R. (2006). Can history be open source? Wikipedia and the future of the past. *Journal of American History*, 93(1):117.

Sandvik, H. (1999). Health information and interaction on the Internet: A survey of female urinary incontinence. *British Medical Journal*, 319(7201):29.

Schneider, S. and Foot, K. (2005). Web sphere analysis: An approach to studying online action. *Virtual methods: Issues in social research on the Internet*, pages 157–170.

Scott, W. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321.

Seidelmann, P. (1992). *Explanatory supplement to the astronomical almanac*. Univ Science Books.

Seidman, J., Steinwachs, D., and Rubin, H. (2003). Conceptual framework for a new tool for evaluating the quality of diabetes consumer-information Web sites. *Journal of Medical Internet Research*, 5(4).

Sellitto, C. and Burgess, S. (2005). Towards a weighted average framework for evaluating the quality of Web-located health information. *Journal of Information Science*, 31(4):260.

Selman, T. J., Prakash, T., and Khan, K. S. (2006). Quality of health information for cervical cancer treatment on the Internet. *BMC Womens Health*, 6:9.

Shanks, G. and Corbitt, B. (1999). Understanding data quality: Social and cultural aspects. In *Proceedings of the 10th Australasian Conference on Information Systems*, volume 785.

Shon, J. and Musen, M. (1999). The low availability of metadata elements for evaluating the quality of medical information on the World Wide Web. 945:9.

Siebert, S. (2002). Alternative movie database. Retrieved on 2009-10-10 from `http://www.steffensiebert.de/amdb/index.html`.

Silberg, W., Lundberg, G., and Musacchio, R. (1997). Assessing, controlling, and assuring the quality of medical information on the Internet: Caveant lector et viewor–Let the reader and viewer beware. *Journal of the American Medical Association*, 277(15):1244–1245.

Smith, E. and Senter, R. (1967). Automated readability index. *AMRL-TR: Aerospace Medical Research Laboratories (6570th)*, page 1.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, (15):72–101.

StargateInformationArchive (2009). Homepage. Retrieved on 2009-03-21 from `http://sg1archive.com/`.

Stausberg, J. and Fuchs, J. (2000). Surgical specialty department in the World Wide Web. Tribute

to contemporary life style or information network? *Chirurg*, 71(4):472–477.

Stausberg, J., Fuchs, J., Husing, J., and Hirche, H. (2001). Health care providers on the World Wide Web: Quality of presentations of surgical departments in Germany. *Medical Informatics & The Internet in Medicine*, 26(1):17–24.

Stead, D., Paton, N., Missier, P., Embury, S., Hedeler, C., Jin, B., Brown, A., and Preece, A. (2008). Information quality in proteomics. *Briefings in bioinformatics*, 9(2):174–188.

Stenner, A. (1996). *Measuring reading comprehension with the Lexile framework*. ERIC Clearinghouse.

Stvilia, B. (2006). *Measuring information quality*. PhD thesis, University of Illinois at Urbana-Champaign.

Stvilia, B., Mon, L., and Yi, Y. (2009). A model for online consumer health information quality. *Journal of the American Society for Information Science and Technology*, 60(9).

Su, Y., Peng, J., and Jin, Z. (2008). Assuring information quality in Web service composition. In *Service Systems and Service Management, 2008 International Conference on*, pages 1–6.

Su, Y., Peng, J., and Jin, Z. (2009). Modeling Information Quality Risk for Data Mining in Data Warehouses. *Human and Ecological Risk Assessment*, 15(2):332–350.

Tabachnick, B. and Fidell, L. (2007). *Using multivariate statistics*. Pearson Education, Inc., 5th edition.

Theberge, P. (2005). Everyday fandom: Fan clubs, blogging, and the quotidian rhythms of the Internet. *Canadian Journal of Communication*, 30(4):485.

TheNielsenCompany (2010). Television: How the numbers come to life. Retrieved on 2010-06-01 from `http://en-us.nielsen.com/tab/measurement/tv_research`.

ThePlanet (2010). Homepage. Retrieved on 2010-06-16 from `http://www.theplanet.com/data-centers/`.

Thorne, S. and Bruner, G. (2006). An exploratory investigation of the characteristics of consumer fanaticism. *Qualitative Market Research: An International Journal*, 9(1):51–72.

Tu, F. and Zimmerman, N. (2001). It is not just a matter of ethics: A survey of the provision of health disclaimers, caveats, and other health-related alerts in consumer health information on eating disorders on the Internet. *International Information and Library Review*, 32(3-4):325–339.

Tukey, J. (1979). Nonparametric Statistical Data Modeling: Comment. *Journal of the American Statistical Association*, pages 121–122.

Türp, J., Gerds, T., and Neugebauer, S. (2001). Myoarthropathien des Kausystems: Beurteilung der Qualität von Patienteninformationen im Weltweiten Netz. *Zeitschrift für Arztliche Fortbildung und Qualitatssicherung*, 95(8):539–548.

Tussey, D. (2000). From fan sites to filesharing: Personal use in cyberspace. *Georgia Law Review*, 35:1129.

UCLAAcademicTechnologyServices (2010). R data analysis examples: Ordinal logistic regression. Retrieved on 2010-03-02 from `http://www.ats.ucla.edu/stat/R/dae/ologit.htm`.

U.S.DepartmentOfEducation (2003). National assessment of adult literacy. Retrieved on 2010-04-20 from `http://nces.ed.gov/naal/`.

Vahabi, M. and Ferris, L. (1995). Improving written patient education materials: A review of the evidence. *Health Education Journal*, 54(1):99.

Venables, W. and Ripley, B. (2010). Mass: Main package of venables and ripley's 'modern applied statistics with s'. Retrieved on 2010-03-02 from `http://cran.r-project.org/web/packages/MASS/index.html`.

Veronin, M. and Ramirez, G. (2000). The validity of health claims on the World Wide Web: A systematic survey of the herbal remedy Opuntia. *American Journal of Health Promotion*, 15(1):21–28.

von Danwitz, F., Baehring, T., and Scherbaum, W. (1999). Verbreitung, Gestaltung und Qualität von deutschsprachigen World Wide Web Seiten zum Diabetes mellitus. *Deutsches Diabetes Forschungsinstitut*.

Walker, J. (2007). Google$^{TM}$page rank query. Retrieved on 2009-07-22 from .

Wang, R. and Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–33.

Wann, D. (1997). *Sport psychology*. Prentice Hall: Upper Saddle River, NJ.

Weissenberger, C., Jonassen, S., Beranek-Chiu, J., Neumann, M., Müller, D., Bartelt, S., Schulz, S., Mönting, J., Henne, K., Gitsch, G., et al. (2004). Breast cancer: patient information needs reflected in English and German web sites. *British Journal of Cancer*, 91:1482–1487.

Wikia (2009a). About. Retrieved on 2009-03-12 from http://www.wikia.com/wiki/Wikia:About.

Wikia (2009b). Advertising. Retrieved on 2009-03-18 from http://www.wikia.com/wiki/Wikia:Advertising.

Wikia (2009c). Category: Articles which may contain original research. Retrieved on 2009-03-05 from http://starwars.wikia.com/wiki/Category:Articles_which_may_contain_original_research.

Wikia (2009d). Community central: Licensing. Retrieved on 2009-03-21 from http://community.wikia.com/wiki/Community_Central:Licensing.

Wikia (2009e). Copyrights. Retrieved on 2009-03-18 from http://www.wikia.com/wiki/Wikia:Copyrights.

Wikia (2009f). Database download. Retrieved on 2009-03-21 from http://www.wikia.com/wiki/Wikia:Database_download.

Wikia (2009g). Disclaimers. Retrieved on 2009-03-18 from http://www.wikia.com/wiki/

Wikia:Disclaimers.

Wikia (2009h). Hiring. Retrieved on 2009-03-18 from http://www.wikia.com/wiki/Wikia:Hiring.

Wikia (2009i). Press. Retrieved on 2009-03-21 from http://www.wikia.com/wiki/Wikia:Press.

Wikia (2009j). Privacy. Retrieved on 2009-03-21 from http://www.wikia.com/Privacy_Policy.

Wikia (2009k). Spring 2009 update. Retrieved on 2009-03-21 from http://www.wikia.com/wiki/Wikia:Press:_Spring_2009_Update.

Wikia (2009l). Terms of use. Retrieved on 2009-03-21 from http://www.wikia.com/Terms_of_Use.

WikimediaFoundation (2009a). Advisory board. Retrieved on 2009-03-12 from http://wikimediafoundation.org/wiki/Advisory_board.

WikimediaFoundation (2009b). Board of trustees. Retrieved on 2009-03-12 from http://wikimediafoundation.org/wiki/Board_of_trustees.

WikimediaFoundation (2009c). Bylaws. Retrieved on 2009-03-12 from http://wikimediafoundation.org/wiki/Bylaws.

WikimediaFoundation (2009d). Current events. Retrieved on 2009-03-12 from http://wikimediafoundation.org/wiki/Current_events.

WikimediaFoundation (2009e). Frequently asked questions. Retrieved on 2009-03-12 from http://wikimediafoundation.org/wiki/Frequently_Asked_Questions.

WikimediaFoundation (2009f). General disclaimer. Retrieved on 2009-03-12 from http://wikimediafoundation.org/wiki/Wikimedia:General_disclaimer.

WikimediaFoundation (2009g). Local chapters. Retrieved on 2009-03-12 from http://

wikimediafoundation.org/wiki/Local_chapters.

WikimediaFoundation (2009h). Our projects. Retrieved on 2009-03-12 from http://wikimediafoundation.org/wiki/Our_projects.

WikimediaFoundation (2009i). Press room. Retrieved on 2009-03-12 from http://wikimediafoundation.org/wiki/Press_room.

WikimediaFoundation (2009j). Privacy policy. Retrieved on 2009-03-12 from http://wikimediafoundation.org/wiki/Wikimedia:Privacy_policy.

WikimediaFoundation (2009k). Resolution: Licensing policy. Retrieved on 2009-03-12 from http://wikimediafoundation.org/wiki/Resolution:Licensing_policy.

WikimediaFoundation (2009l). Resolutions. Retrieved on 2009-03-12 from http://wikimediafoundation.org/wiki/Resolutions.

WikimediaFoundation (2009m). Staff. Retrieved on 2009-03-12 from http://wikimediafoundation.org/wiki/Staff.

WikimediaFoundation (2009n). Terms of use. Retrieved on 2009-03-12 from http://wikimediafoundation.org/wiki/Terms_of_Use.

WikimediaFoundation (2009o). Values. Retrieved on 2009-03-12 from http://wikimediafoundation.org/wiki/Values.

WikimediaFoundation (2010). Annual report. Retrieved on 2010-05-01 from http://wikimediafoundation.org/wiki/Annual_report.

Wikipedia (2009a). Administrators. Retrieved on 2009-03-12 from http://en.Wikipedia.org/wiki/Wikipedia:Administrators.

Wikipedia (2009b). Contact us. Retrieved on 2009-03-21 from http://en.wikipedia.org/wiki/Wikipedia:Contact_us.

Wikipedia (2009c). Contributing to wikipedia. Retrieved on 2009-03-07 from http://en.

Wikipedia.org/wiki/About_Wikipedia.

Wikipedia (2009d). Featured article criteria. WWW. Retrieved on 2009-02-25 from http://en.Wikipedia.org/wiki/Wikipedia:Featured_article_criteria.

Wikipedia (2009e). No original research. Retrieved on 2009-03-05 from http://en.Wikipedia.org/wiki/Wikipedia:No_original_research.

Wikipedia (2009f). Peer review. WWW. Retrieved on 2009-02-25 from http://en.Wikipedia.org/wiki/Wikipedia:Peer_review.

Wikipedia (2009g). Reliable sources. WWW. Retrieved on 2009-02-28 from http://en.Wikipedia.org/wiki/Wikipedia:Reliable_sources.

Wikipedia (2009h). Reusing wikipedia content. Retrieved on 2009-03-12 from http://en.Wikipedia.org/wiki/Wikipedia:Reusing_Wikipedia_content.

Wikipedia (2009i). What is a featured article? WWW. Retrieved on 2009-02-28 from http://en.Wikipedia.org/wiki/Wikipedia:What_is_a_featured_article.

Wikipedia (2009j). Wikipedia database. Retrieved on 2009-03-17 from http://en.Wikipedia.org/wiki/Wikipedia_database.

Wikipedia (2010a). Neutral point of view. Retrieved on 2010-06-14 from http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view.

Wikipedia (2010b). Wikignome. Retrieved on 2010-04-07 from http://en.wikipedia.org/wiki/Wikipedia:WikiGnome.

Yadav, S. (2008). Automation of webpage quality determination. *International Journal of Information Quality*, 2(2):152–176.

Yahoo! (2009). Yahoo! site explorer. Retrieved on 2009-07-22 from https://siteexplorer.search.yahoo.com/.

Yamamoto, L. (2000). Copyright protection and Internet fan sites: Entertainment industry finds

solace in traditional copyright law. *Loyola of Los Angeles Entertainment Law Review*, 20:95.

Zeist, R. and Hendriks, P. (1996). Specifying software quality with the extended ISO model. *Software Quality Journal*, 5(4):273–284.

Zhu, X. and Gauch, S. (2000). Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 288–295.

Zhu, Y. (2008). Group Assessment of Web Source/Information Quality Based on WebQM and Fuzzy Logic. *Lecture Notes in Computer Science*, 5009:660.

# Appendix A

# Codebook of variables

## Key

**Name:** the variable's name

**Type(s):** the variable's type. PA means presence-absence/binary, CONT means continuous, COUNT means count, GD means Gregorian date, JD means Julian Day Number, LIST means a list (e.g., of CSS styles or validation errors), ORD means ordinal, and TXT means textual.

**Meaning:** a short description of what the variable signifies

**Example:** the criterion used to determine that the variable had manifested/occurred in an artifact under study

**RQ(s):** which research sub-question(s) the variable was used to answer. AA abbreviates Artifactual-Affective, RA means Reflective Accuracy, and RC means Reflective Completeness. These are the three types of research questions discussed in the literature chapter (2).

Table A.1: Media content types

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| rss | PA | Really Simple Syndication | “rss+xml” | AA2 |
| atom | PA | Atom syndication | “atom+xml” | AA2 |
| css | PA | Cascading Style Sheet | “text/css” | AA2 |
| js | PA | JavaScript | “text/javascript” | AA2 |
| jpg | PA | JPEG image | “.[Jj][Pp]*[Gg]” | AA2 |
| gif | PA | GIF image | “.[Gg][Ii][Ff]” | AA2 |
| png | PA | PNG image | “.[Pp][Nn][Gg]” | AA2 |
| pdf | PA | PDF image | “.[Pp][Dd][Ff]” | AA2 |
| svg | PA | SVG image | “.[Ss][Vv][Gg]” | AA2 |
| mp3 | PA | MP3 audio | “.[Mm][Pp]3” | AA2 |
| aac | PA | AAC audio | “.[Aa][Aa][Cc]” | AA2 |
| m4a | PA | M4A audio | “.[Mm]4[Aa]” | AA2 |
| ogg | PA | OGG audio | “.[Oo][Gg][Gg]” | AA2 |
| flac | PA | FLAC audio | “.[Ff][Ll][Aa][Cc]” | AA2 |
| flash | PA | Flash media | “.[Ss][Ww][Ff]” | AA2 |
| mp4 | PA | MP4 video | “.[Mm][Pp]4” | AA2 |
| m4v | PA | M4V video | “.[Mm]4[Vv]” | AA2 |
| mov | PA | MOV video | “.[Mm][Oo][Vv]” | AA2 |

Continued on Next Page...

Table A.1 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| 3gp | PA | 3GP video | ".3[Gg][Pp]" | AA2 |
| mpg | PA | MPEG video | ".[Mm][Pp]*[Gg]" | AA2 |
| imagesInGallery | COUNT | website specifies number of images in an associated gallery | "Photos (see all # ..." | AA2 |
| videosInGallery | COUNT | ... videos ... | "Videos (see all #)" | AA2 |

Table A.2: Links

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| total | COUNT, LIST | links of any type | (any URL) | RA12 |
| external | COUNT, LIST | links leaving the site | otherdomain.com/... | RA12 |
| broken | COUNT, LIST | link without working destination | (untraversable URL) | AA3 |
| inlinks | COUNT, LIST | links to the site from elsewhere | (from Yahoo!) | RA11 |
| pagerank | CONT | Google PageRank (integers, 0-10) | (from Google) | RA11 |

Continued on Next Page...

Table A.2 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| firstchp | PA | links to first chapter of a book | "Read the first chapter" | RA12 |
| review | PA | links to editorial review | "Read Review!" | RA12 |
| transcript | PA | links to transcript | "Transcript" | RA12 |
| forum | PA | links to forum | "Discuss" link | RA12 |
| itunes | PA | links to iTunes | "Download on iTunes" | RA8,12 |
| amazon | PA | links to Amazon | "Download full episode" | RA8,12 |
| officialSite | PA, LIST | lists links to official site(s) | "official sites" | RA3,12 |
| miscSites | PA, LIST | lists links to sites dubbed "misc" | "miscellaneous" | RA3,12 |
| photoSites | PA, LIST | lists links to photo gallery sites | "photographs" | RA3,12 |
| soundSites | PA, LIST | lists links to sound gallery sites | "sound clips" | RA3,12 |
| videoSites | PA, LIST | lists links to video gallery sites | "video clips" | RA3,12 |
| showtimes | PA, LIST | lists links to showtime sites | "showtimes" | RA3,12 |

Table A.3: Validation

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| html | COUNT, LIST | validation errors | “Warning: missing ” | AA4 |
| wcag | COUNT, LIST | accessibility errors | “Access: [3.2.1.1]: <doctype> missing” | AA4 |
| total | COUNT, LIST | both error types | (see last two) | AA4 |

Table A.4: Ratings

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| editor | ORD | GateWorld editor ratings | “***” | RA9-10 |
| fan | CONT, ORD | GateWorld & Wikia fan ratings | integers, 1-10 or 1-5 | RA9-10 |
| network | CONT | Nielsen rating when first broadcast | “1.7” | RA9-10 |
| synd | CONT | Nielsen rating when in syndication | “0.8” | RA9-10 |

Table A.5: Short text fields

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| title | TXT, PA | something's title | "=Stargate=" | RA7 |
| author | TXT, PA | book's author's names | "AUTHOR:" | RA7 |
| writers | TXT, PA, COUNT | writers | "WRITTEN BY:" | RA7 |
| directors | TXT, PA, COUNT | directors | "DIRECTED BY:" | RA7 |
| starring | PA, COUNT | main cast | "Cast" | RA7 |
| gueststars | TXT, PA | guest stars | "GUEST STARRING:" | RA7 |
| composer | COUNT, PA | composers or narrators | "composer" | RA7 |
| editors | COUNT, PA | (script) editors | "editor" | RA7 |
| illustrator | COUNT | illustrators | "illustrator" | RA7 |
| producers | COUNT, PA | (executive) producers | "producer" | RA7 |
| effects | COUNT | effects/fx managers | "fx" | RA7 |
| cinematography | PA | cinematographers/photographers | "Cinematography" | RA7 |

Continued on Next Page...

Table A.5 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| miscCos | COUNT | companies outsourced to | "Production Company:" | RA7 |
| name | TXT, PA | actor or character's names | "Name:" | RA7 |
| altName | PA | alternate names, aliases | "Sometimes credited as:" | RA7 |
| birthName | PA | birth or real names | "Real Name:" | RA7 |
| nickName | PA | nicknames or callsigns | "Nick Name:" | RA7 |
| birth | PA | birth dates or places | "Birth Date:" | RA7 |
| occupation | PA | person's specialties, employers | "Occupation:" | RA7 |
| education | PA | person's education, alma maters | "Alma mater:" | RA7 |
| died | PA | death date or place | "Died:" | RA7 |
| height | PA | person's height or weight | "Height:" | RA7 |
| booksAbout | COUNT | books written about | "Books about" | RA7 |
| moviesPortIn | COUNT | movies about | "Was portrayed in" | RA7 |

Continued on Next Page...

Table A.5 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| spouse | PA | marital and family status | "Married to:" | RA7 |
| gender | PA | person's gender | "Gender:" | RA7 |
| nationality | PA | person's nationality, ethnicity | "Nationality:" | RA7 |
| haircolor | PA | person's hair color | "Hair:" | RA7 |
| eyecolor | PA | person's eye color | "Eyes:" | RA7 |
| triviaItems | COUNT | trivia or goofs | "Trivia:" | RA7, RC5 |
| commercials | COUNT | commercials featuring | "Commercials:" | RA7 |
| theatre | COUNT, PA | theatre shows featuring | "Theatre:" | RA7 |
| otherWorks | PA | other works by | "Other Works:" | RA7 |
| awards | COUNT, PA | awards received by | "Awards:" | RA7 |
| quotes | COUNT | quotes by | "Quotes:" | RA7 |
| trademarksC | COUNT | person's trademark role | "Trademark:" | RA7 |
| salary | COUNT, PA | person's earnings or net worth | "Salary:" | RA7 |

Continued on Next Page...

Table A.5 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| wherenow | PA | person's current job | "Where are they now:" | RA7 |
| interviews | COUNT | interviews with | "Interviews:" | RA7 |
| articles | COUNT | articles about | "Articles:" | RA7 |
| pictorials | COUNT | pictorials about | "Pictorials:" | RA7 |
| covers | COUNT | magazine covers featuring | "Covers:" | RA7 |
| filmAct | COUNT, PA | films actor | "Filmography as Actor:" | RA7 |
| filmNonAct | COUNT | films as non-actor | "Filmography as Writer:" | RA7 |
| adaptations | PA | other works that adapt this | "Adaptations:" | RA7 |
| admissions | COUNT | numbers of tickets sold | "Admissions:" | RA7 |
| akaTitles | COUNT | titles in other languages | "Aka Titles:" | RA7 |
| boxOffice | PA | money earned in ticket sales | "Box Office Gross:" | RA7 |

Continued on Next Page...

Table A.5 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| biz | PA | business/earnings details | “Business:” | RA7 |
| cast | COUNT | full cast | “Cast (in credits order):” | RA7 |
| certificates | COUNT | ratings in each country | “Certificates:” | RA7 |
| copyright | PA | copyright or license holders | “Copyright Holder:” | RA7 |
| costumes | COUNT | costume designers | “Costume Designer:” | RA7 |
| countries | COUNT | production countries | “Country of Production:” | RA7 |
| critics | COUNT | critical review publications | “Critics:” | RA7 |
| locations | PA | filming locations | “Locations:” | RA7 |
| authStory | TXT, PA | story authors | “STORY BY:” | RA7 |
| authArt | TXT, PA | artistic directors | “ART BY:” | RA7 |
| authColor | TXT, PA | color directors | “COLOR BY: ” | RA7 |

Continued on Next Page...

Table A.5 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| authCover | TXT, PA | main cover artists | "MAIN COVER BY: " | RA7 |
| authTeleplay | TXT, PA | teleplay authors | "TELEPLAY BY:" | RA7 |
| authExcerpts | TXT, PA | excerpts authors | "EXCERPTS WRITTEN BY:" | RA7 |
| charPlayedBy | TXT, PA | actor playing this character | "PLAYED BY:" | RA7 |
| planetHometo | TXT, PA | planet's inhabitants | "HOME TO:" | RA7 |
| raceHome | TXT, PA | race's home planet | "HOMEWORLD: " | RA7 |
| shipUsedBy | TXT, PA | ship's usual occupants | "USED BY:" | RA7 |
| epFirstApp | TXT, PA | where first appeared | "FIRST APPEARED:" | RA7 |
| epKey | TXT, PA | character's pivotal episodes | "KEY EPISODE:" | RA7 |
| sgroles | COUNT | actor's characters in Stargate | "Stargate roles:" | RA7 |

Continued on Next Page...

Table A.5 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| publisher | TXT, PA | title's publishing company | "PUBLISHER:" | RA7 |
| developer | TXT, PA | game's developing company | "DEVELOPER:" | RA7 |
| distributor | COUNT | distributing companies | "Distributor:" | RA7 |
| platform | TXT, PA | game's compatible systems | "PLATFORM:" | RA7 |
| issue | TXT, PA | magazine's issue number | "ISSUE NUMBER: " | RA7 |
| epnumber | TXT, PA | episode's production code | "EPISODE NUMBER:" | RA7 |
| dvdnumber | TXT, PA | DVD's order in box set | "DVD DISC:" | RA7 |
| genres | TXT, PA, COUNT | title's genres | "Genres:" | RA7 |
| keywords | TXT, PA | title's key subject matter | "Keywords:" | RA7 |
| mediatype | PA | primary media format (e.g., DVD) | "Media type:" | RA7 |

Continued on Next Page...

Table A.5 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| othermedia | COUNT | other available formats | "Other media:" | RA7 |
| pages | PA | book's page count | "Pages:" | RA7 |
| isbn10 | PA | book's ISBN-10 number | "ISBN10:" | RA7 |
| isbn13 | PA | book's ISBN-13 number | "ISBN13:" | RA7 |
| issn | PA | book's ISSN number | "ISSN:" | RA7 |
| datetable | PA | table describing events | "DtBroadcasted:" | RA7 |
| content | COUNT | DVD's contents | "==Contents==" | RA7 |
| feature | COUNT | magazine's feature sections | "Features" | RA7 |
| contributors | COUNT | magazine's regular authors | "==Regulars==" | RA7 |
| contributed | COUNT | magazine author's bib | "==Works Contributed==" | RA7 |

Continued on Next Page...

Table A.5 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| size | PA | a technology or place's size | "Size:" | RA7 |
| era | PA | time period when existed | "Era:" | RA7 |
| species | PA | character's species' name | "Species:" | RA7 |
| address | PA | a stargate address | —address= | RA7 |
| affiliation | PA | affiliated organization | —affiliation= | RA7 |
| allegiances | PA | to team, country, etc. | —allegiances= | RA7 |
| alliances | PA | with others | —alliances= | RA7 |
| anthem | PA | national anthem | —anthem= | RA7 |
| appearance | PA | first appearance | —appearance= | RA7 |
| armament | PA | weaponry | —armament= | RA7 |
| avionics | PA | flight control interface | —avionics= | RA7 |
| capacity | PA | to carry something | —capacity= | RA7 |
| capital | PA | of a state or region | —capital= | RA7 |
| class | PA | of a ship | —ship= | RA7 |
| commander | PA | of a ship, facility, or team | —commander= | RA7 |

Continued on Next Page...

Table A.5 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| conc | PA | concurrent events | —conc= | RA7 |
| conflict | PA | in what war the conflict occurred | —conflict= | RA7 |
| control | PA | interface of a weapon | —control= | RA7 |
| cost | PA | financial cost | —cost= | RA7 |
| countermeasures | PA | of a ship in battle | —countermeasures= | RA7 |
| crew | PA | of a ship | —crew= | RA7 |
| currency | PA | of a peoples | —currency= | RA7 |
| designer | PA | race who created | —designer= | RA7 |
| discharge | PA | power output | —discharge= | RA7 |
| dist | PA | distinguishing features | —dist= | RA7 |
| domination | PA | under the control of | —domination= | RA7 |
| engine | PA | propulsion mechanism | —engine= | RA7 |
| established | PA | date begun | —established= | RA7 |
| executive | PA | executive branch of a council | —executive= | RA7 |
| firstuse | PA | first episode used in | —firstuse= | RA7 |

Continued on Next Page...

Table A.5 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| formed | PA | when/where a group was formed | —formed= | RA7 |
| founding | PA | founding document | —founding= | RA7 |
| fragmented | PA | date a group became fragmented | —fragmented= | RA7 |
| fuel | PA | a vessel's fuel type | —fuel= | RA7 |
| function | PA | a technology's primary purpose | —function= | RA7 |
| galaxy | PA | galactic location | —galaxy= | RA7 |
| govt | PA | a people's form of government | —govt= | RA7 |
| hdsystem | PA | type of hyperdrive engine | —hdsystem= | RA7 |
| headofstate | PA | chief of government | —headofstate= | RA7 |
| homeplanet | PA | a people's home planet | —homeplanet= | RA7 |
| hull | PA | a ship's hull's composition | —hull= | RA7 |

Continued on Next Page...

Table A.5 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| interest | PA | reason for Earth's interest | —interest= | RA7 |
| judicial | PA | judicial branch of a council | —judicial= | RA7 |
| legislative | PA | legislative branch of a council | —legislative= | RA7 |
| manufacturer | PA | a technology's manufacturer | —manufacturer= | RA7 |
| material | PA | a raw material | —material= | RA7 |
| maxspeed | PA | a ship's top speed | —maxspeed= | RA7 |
| model | PA | a weapon's intended purpose/scope | —model= | RA7 |
| navigation | PA | a ship's navigation system type | —navigation= | RA7 |
| next | PA | the next in a series | —next= | RA7 |
| origin | PA | location of creation | —origin= | RA7 |
| othersystems | PA | a ship's misc interesting systems | —othersystems= | RA7 |
| passengers | PA | a ship's occupancy | —passengers= | RA7 |

Continued on Next Page...

Table A.5 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| place | PA | location of an event | —place= | RA7 |
| pointoforigin | PA | seventh stargate symbol | —pointoforigin= | RA7 |
| population | PA | of a race or place | —population | RA7 |
| power | PA | something's power source | —power= | RA7 |
| poweroutput | PA | something's power output | —poweroutput= | RA7 |
| prev | PA | the previous in a series | —previous= | RA7 |
| range | PA | a weapon's range | —range= | RA7 |
| rank | PA | a military person's rank | —rank= | RA7 |
| religious | PA | a people's state religion | —religious= | RA7 |
| reorganized | PA | date of a group's reorganization | —reorganized= | RA7 |
| restored | PA | date of a group's restoration | —restored= | RA7 |

Continued on Next Page...

Table A.5 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| result | PA | outcome of an action/battle | —result= | RA7 |
| sensor | PA | on a ship or facility | —sensor= | RA7 |
| shieldgen | PA | shield generation capability | —shieldgen= | RA7 |
| skeleton | PA | a ship's minimum crew | —skeleton= | RA7 |
| status | PA | of an action or people | —status= | RA7 |
| target | PA | of an attack | —target= | RA7 |
| tech | PA | a people's technological sophistication | —tech= | RA7 |
| type | PA | a weapon's underlying principle | —type= | RA7 |
| width | PA | physical width | —width= | RA7 |
| region | PA | a DVD's playable region | —region= | RA7 |
| blu-ray | PA | a title's Blu-Ray availability | —blu-ray= | RA7 |

Continued on Next Page...

Table A.5 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| audiolang | PA | a title's audio languages | —audiolang= | RA7 |
| subtitlesLang | PA | a title subtitle languages | —subtitlesLang= | RA7 |
| numdiscs | PA | number of discs in a DVD set | —numdiscs= | RA7 |
| numeps | PA | number of episodes | —numeps= | RA7 |
| influences | PA | an author's influences | —influences= | RA7 |
| signature | PA | an author's signature | —signature= | RA7 |
| first | PA | a character's first appearance | —first= | RA7 |
| last | PA | a character's last appearance | —last= | RA7 |
| creator | PA | crew who conceived of something | —creator= | RA7 |
| caption | PA | picture or table caption | —caption= | RA7 |

Table A.6: Long text fields

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| shortdesc | TXT, PA | 1-2 sentence introduction | (below title) | RA7 |
| bio | TXT, PA | character biography | “==Biography==” | RA7 |
| career | TXT, PA | actor’s career summary | “==Career==” | RA7 |
| personallife | TXT, PA | actor’s personal life summary | “==Personal life==” | RA7 |
| previously | TXT, PA | previous plot developments | “==Previously on SG-1==” | RA7 |
| plot | TXT, PA | plot summary | “==Plot==” | RA7 |
| tech | TXT, PA | technology, person, battle summary | “==Overview==” | RA7 |
| society | TXT, PA | summary of alien society | “==Society==” | RA7 |
| analysis | TXT, PA | plot discussion | “ANALYSIS” | RA7 |
| chardev | TXT, PA | character development discussion | “CHARACTER DEVELOPMENT” | RA7 |
| notes | TXT, PA, COUNT | in-universe fact list | “NOTES” | RA7 |
| questions | TXT, PA | unanswered questions list | “UNANSWERED QUESTIONS” | RA7 |

Continued on Next Page...

Table A.6 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| production | TXT, COUNT | production crew quotes | “PRODUCTION” | RA7 |
| cheats | TXT, PA | video game cheats/tricks | “CHEATS” | RA7 |
| authsum | TXT, PA | author’s summary of book | “==Author’s summary==” | RA7 |
| pubsum | TXT, PA | publisher’s summary of book | “==Publisher’s summary==” | RA7 |
| english | TXT, PA | English summary of book | “==English==” | RA7 |
| german | TXT, PA | German summary of book | “==German==” | RA7 |
| gameplay | TXT, PA | video game experience summary | “GAMEPLAY” | RA7 |
| altRealNotes | TXT, COUNT | alternate reality plot summary | “==In alternate realities==” | RA7 |
| reception | TXT, PA | critical reception summary | “==Reception==” | RA7 |

Table A.7: Dates

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| release | GD, JD, PA | generic release date | “Released:” | RA7 |
| airdate.us | GD, JD, PA | episode US airdate | “ORIGINAL U.S. AIR DATE” | RA7 |
| airdate.syn | GD, JD, PA | episode US syndication airdate | “SYNDICATION AIR DATE” | RA7 |
| production | PA | production dates | “Production:” | RA7 |
| created | GD, JD | webpage creation date | (from MediaWiki API) | RA1 |
| lastmod | GD, JD | webpage last modification date | (from MediaWiki API) | RA1 |

Table A.8: Readability

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| kincaid | CONT | Kincaid Formula | (found with `style`) | AA5 |
| ari | CONT | Automated Readability Index | (found with `style`) | AA5 |

Continued on Next Page...

Table A.8 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| cl | CONT | Coleman-Liau Formula | (found with `style`) | AA5 |
| flesch | CONT | Flesch Reading Ease formula | (found with `style`) | AA5 |
| fog | CONT | Gunning-Fog Index | (found with `style`) | AA5 |
| lix | CONT | Lix formula | (found with `style`) | AA5 |
| smog | CONT | Simple Measure of Gobbledygook | (found with `style`) | AA5 |
| fry | CONT | Fry readability graph | (found by re-expressing graph) | AA5 |

Table A.9: Word usage

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| chars | COUNT | characters in the text | “a” | RC4-5 |
| words | COUNT | words in the text | “word” | RC4-5 |
| w.avlen.ch | COUNT | words’ average length, in characters | “word” = 4 | RC4-5 |

Continued on Next Page...

Table A.9 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| w.avsyl | CONT | words' average syllables | "ex-am-ple" = 3 | RC4-5 |
| sents | COUNT | sentences in the text | "This is one." | RC4-5 |
| s.avlen.w | CONT | sentences' average length, in words | "This is one." = 3 | RC4-5 |
| s.short | COUNT | sentences under 10-12 words long | | RC4-5 |
| s.long | COUNT | sentences over 25-27 words long | | RC4-5 |
| para | COUNT | paragraphs in the text | (≥ 2 newlines) | RC4-5 |
| p.avlen.s | COUNT | paragraphs' average length, in sentences | | RC4-5 |
| s.quest | COUNT | sentences ending in a question mark | "Is this one?" | RC4-5 |
| s.pass | COUNT | sentences in passive voice | "This was made passive." | RC4-5 |
| s.longest.w | COUNT | the text's longest sentence, in words | | RC4-5 |
| s.shortest.w | COUNT | ... shortest ... | | RC4-5 |
| v.tobe | COUNT | "to be" verbs | "will be different" | RC4-5 |

Continued on Next Page...

Table A.9 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| v.aux | COUNT | modal auxiliary verbs | "can, could, should" | RC4-5 |
| conj | COUNT | coordinating and subordinating conjunctions | "and, but, because" | RC4-5 |
| pron | COUNT | pronouns | "I, you, who" | RC4-5 |
| prep | COUNT | prepositions | "of, to, for" | RC4-5 |
| noms | COUNT | nominalizatons, verbs changed to nouns | "-ment, -ance, -ion" | RC4-5 |
| s.b.pron | COUNT | sentences beginning with pronouns | "I went...." | RC4-5 |
| s.b.int | COUNT | ... with interrogative pronouns | "Who are you?" | RC4-5 |
| s.b.art | COUNT | ... with articles | "The kite flew." | RC4-5 |
| s.b.sub | COUNT | ... with subordinating conjunctions | "Even if...." | RC4-5 |
| s.b.conj | COUNT | ... with coordinating conjunctions | "And still...." | RC4-5 |
| s.b.prep | COUNT | ... with prepositions | "To go with them...." | RC4-5 |

Table A.10: Revisions & authors

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| totalRevisions | COUNT | a wiki page's total revisions | (from MediaWiki API) | RC4 |
| uniqueAuthors | COUNT | a wiki page's unique authors | (from MediaWiki API) | RC4 |

Table A.11: Lists, tables, & sections

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| listTabInfo | COUNT | lists, tables, and infoboxes on page | "{{Infobox" | RC5 |
| sections | COUNT | content sections on page | "==section==" | RC5 |

Table A.12: Vendor types

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| amazon | PA | Amazon.com | “Buy on Amazon” | RA8 |
| itunes | PA | Apple iTunes | “Download on iTunes” | RA8 |
| IT | PA | information technology | “Dell computers” | RA8 |
| games | PA | games and sports | “World of Warcraft” | RA8 |
| scifi | PA | science fiction, fantasy | “new Fringe episodes” | RA8 |
| auto | PA | automotive | “Toyota Prius” | RA8 |
| pet | PA | house pets | “dogs and cats” | RA8 |
| travel | PA | travel | “picture of a beach” | RA8 |
| finance | PA | banking, business, casting | “credit card offers” | RA8 |
| health | PA | health and food | “Assurant health insurance” | RA8 |
| fam | PA | family and romance | “find a date” | RA8 |
| sweepstakes | PA | contests and sweepstakes | “enter to win” | RA8 |
| acad | PA | academia, educational | “enroll in distance courses” | RA8 |
| security | PA | military and security | “join the national guard” | RA8 |
| fashion | PA | retail and high fashion | “designer sunglasses” | RA8 |
| massmedia | PA | non-scifi mass media | “new Survivor episodes” | RA8 |
| self | PA | site advertises itself | “Narnia wiki” (on Wikia) | RA8 |
| rental | PA | media rental services | “Rent it at...” | RA8 |

Continued on Next Page...

Table A.12 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| furnishings | PA | home furnishings, general retail | “couches at Target” | RA8 |
| charity | PA | charity causes | “help end Darfur genocide” | RA8 |

Table A.13: Evidence source types

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| local | PA | website cites itself | (local URL) | RA3,6; RC3 |
| gw | PA | GateWorld | “Article on GateWorld” | RA3,6; RC3 |
| imdb | PA | IMDb | “PLAYED BY: (IMDb URL)” | RA3,6; RC3 |
| wikia | PA | Wikia | “Article on Stargate Wikia” | RA3,6; RC3 |
| wp | PA | Wikipedia | “Article on Wikipedia ” | RA3,6; RC3 |
| publisher | PA | publisher’s description | “From Fandemonium Ltd.” | RA3,6; RC3 |

Continued on Next Page. . .

Table A.13 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| user | PA | user forums, comments, messageboards | "IMDb Mini Biography By:" | RA3,6; RC3 |
| studio | PA | film studio, game developer, content owner | "Official Stargate SG-1 website" | RA3,6; RC3 |
| tvnetwork | PA | TV network airing content | "Official Syfy website" | RA3,6; RC3 |
| adsite | PA | site that hosts ads | "Gatenoise from Moon-catchin"' | RA3,6; RC3 |
| sgwiki | PA | Stargate SG-1 Solutions | "stargate-sg1-solutions.com" | RA3,6; RC3 |
| rda.com | PA | Kathleen Ritter's Lexicon | "rdanderson.com" | RA3,6; RC3 |
| sg1archive | PA | Stargate Information Archive | "sg1archive.com" | RA3,6; RC3 |
| savesg1 | PA | Save Stargate SG-1 | "savestargatesg1.com" | RA3,6; RC3 |
| fansite | PA | non-Stargate, but related fansite | "christopherheyerdahl.net" | RA3,6; RC3 |

Continued on Next Page...

Table A.13 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| indexsite | PA | site hosting celebrity information | "MovieTome" | RA3,6; RC3 |
| books | PA | book, magazine, CD | "Stepping Through the Stargate" | RA3,6; RC3 |
| producer | PA | Stargate producers' weblogs | "Joseph Mallozzi's weblog" | RA3,6; RC3 |
| DVDextras | PA | bonus content on DVDs | "As noted on the DVD audio commentary..." | RA3,6; RC3 |
| massmedia | PA | the mass and news media | "Tapping joins 'Stargate Atlantis'. Chicago Tribune." | RA3,6; RC3 |
| personal | PA | cast and crew's personal websites | "amandatapping.com" | RA3,6; RC3 |
| acad | PA | an academic work | "Reading Stargate SG-1" | RA3,6; RC3 |

Table A.14: Original research types

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| interp | PA | gives interpretations and opinions | "ANALYSIS, NOTES" sections | RA4 |
| unansQuest | PA | identifies unanswered questions | "UNANSWERED QUESTIONS" section | RA4 |
| prodDet | PA | gives production details | "PRODUCTION" section | RA4 |
| cultRef | PA | identifies cultural references | "Jonathan Glassner had written The Wizard of Oz references into his own scripts...." | RA4 |
| bio | PA | discusses biographical or historical context | "Brad Wright and Jonathan Glassner had worked together on the MGM television series The Outer Limits since 1995." | RA4 |
| reception | PA | discusses critical reception | "Critical reception" section | RA4 |

Table A.15: Page types

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| ep.guide | PA | episode guide | "Stargate SG-1 Season Ten: Unending" | RA7 |
| ep.trans | PA | episode transcript | "Transcript by" | RA7 |
| ep.review | PA | episode review | "Read the GateWorld review" | RA7 |
| ep.makingof | PA | behind-the-scenes | "In The Making: Origin" | RA7 |
| awards | PA | list of awards | "List of awards and nominations received by Stargate Atlantis" | RA7 |
| omni.char | PA | a character | "Jack O'Neill " | RA7 |
| omni.peoples | PA | a group, race, or species | "The Asgard" | RA7 |
| omni.place | PA | a place | "Cheyenne Mountain" | RA7 |
| omni.tech | PA | a technology | "the stargate" | RA7 |
| omni.scinature | PA | science or nature | "anti-matter" | RA7 |
| omni.lang | PA | a language | "Rillaanian language" | RA7 |
| omni.cultref.out | PA | a non-Stargate cultural reference | "The Simpsons" | RA7 |

Continued on Next Page...

Table A.15 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| omni.cultref.in | PA | a Stargate-specific cultural reference | “Tuttleroot soup” | RA7 |
| omni.battles,wars | PA | a battle or war | “Battle of Abydos” | RA7 |
| omni.sg-overall | PA | a Stargate series overall | “SG-1” | RA7 |
| book | PA | a guide to a book or magazine | “Stargate Atlantis: The DVD Collection 86” | RA7 |
| comic | PA | a guide to a comic | “Doomsday World 3” | RA7 |
| vg | PA | a guide to a video game | “Stargate SG-1: The Alliance” | RA7 |
| dvd | PA | a guide to a DVD | “Stargate SG-1: The Complete Tenth Season” | RA7 |
| template | PA | page template | “Template:Infobox Battle” | RA7 |
| admin | PA | administrative page | “SGCommand:WikiNode” | RA7 |
| list | PA | page containing only lists | “List of Atlantis personnel” | RA7 |

Continued on Next Page...

Table A.15 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| actor | PA | an actor | “Christopher Heyerdahl” | RA7 |
| crew | PA | a crew member | “Damian Kindler” | RA7 |
| author | PA | a book or comic author | “Jennifer Fallon” | RA7 |
| date | PA | summary of events by date or time | “2005” | RA7 |
| disambig | PA | a disambiguation page | “Perseus” | RA7 |
| demographics | PA | demographics about real-life | “a guide to Vancouver, BC” | RA7 |

Table A.16: Characters' personal details

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| altname | PA | an alternate name or alias | "Alternate Names:" | RA5 |
| rank | PA | ranks or degrees | "Rank:" | RA5 |
| decorations | PA | military decorations | "==Military Decorations==" | RA5 |
| superiors | PA | boss or superior | "...a member of Sheppard's exploratory team..." | RA5 |
| colleagues | PA | colleagues, group | "Allegiances:", "Race:" | RA5 |
| station | PA | station, post, assignment | "...stationed at Earth's Atlantis base..." | RA5 |
| age | PA | age, in years | "Age:" | RA5 |
| gender | PA | gender | "Gender: " | RA5 |
| bio | PA | biography, family details | "==Biography==" | RA5 |
| characteristics | PA | characteristics, abilities | "==Abilities and skills==" | RA5 |

Continued on Next Page...

Table A.16 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| alienContact | PA | the person's non-human contacts | "She also made friends with the Tok'ra Martouf." | RA5 |
| ep.appearedIn | PA | key appearances | "KEY EPISODES:" | RA5 |
| playedBy | PA | actor(s) portraying the character | "Actor:" | RA5 |

Table A.17: Textual themes

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| archae | PA | archaeology, egyptology | "Daniel is an archaeologist and linguist..." | RA5 |
| ling | PA | linguistics | "...who speaks more than twenty-three languages...." | RA5 |
| mil | PA | military honor, hero | "...in honor of the fallen physician who had so heroically preserved his life." | RA5 |

Continued on Next Page...

Table A.17 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| travel | PA | time/space travel | "On a standard reconnaissance mission to planet..." | RA5 |
| tech | PA | technology, weapons | "...expert on alien technology." | RA5 |
| slave | PA | slavery, pretense | "...aliens responsible for taking humans from Earth and seeding them all across the galaxy." | RA5 |
| medicine | PA | medicine, healing | "...Goa'uld technology such as hand and Healing Devices." | RA5 |
| possess | PA | demon possession, parasites, brainwashing, free will | "...she was taken over by a Tok'ra symbiote..." | RA5 |
| romance | PA | romance, sex, marriage | "==Romance==" | RA5 |
| injoke | PA | in-jokes, self-deprecation | "Teal'c: Daniel Jackson's preliminary electroencephalogram proved anomalous. O'Neill: I dare you to say that again." | RA5 |

Continued on Next Page...

Table A.17 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| races | PA | races, species | "the Ori" | RA5 |
| science | PA | science, nature | "the multiverse and entropic cascade failure" | RA5 |
| invade | PA | overtaking by force | "...we will spread Origin to all the unbelievers." | RA5 |
| boom | PA | explosions, bombs | "...they blew up Vorash's sun." | RA5 |
| homage | PA | homages, cameos, cultural references | "zombies, Wizard of Oz, Simpsons, Star Trek, Farscape, 1969..." | RA5 |
| gov | PA | large federal governments | "Kinsey was the Senator of Indiana and the chairman of the Senate Appropriations Committee, which controlled the Stargate Program's budget." | RA5 |
| rsch | PA | research and development | "Area 51 contains labs for medical research, geology, space metallurgy, and artifacts study." | RA5 |

Continued on Next Page...

Table A.17 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| treasHunt | PA | treasure hunts, puzzles, riddles | "Vala Mal Doran stole a tablet written in Ancient code. The tablet told of an incredible treasure hidden on Earth...." | RA5 |
| con | PA | betrayal, deception, lying | "Vala Mal Doran is a thief and con artist, former Goa'uld host, and current member of SG-1." | RA5 |

Table A.18: Production information

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| budget | PA | the production's budget | "I can do it for $55 million." | RA5 |
| sequel | PA | plans for sequels | "There has been plenty of talk about doing a sequel...." | RA5 |
| why | PA | actor/crew motivations | "I have a certain manual-labourist view of acting." | RA5 |

Continued on Next Page...

Table A.18 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| prefs | PA | actor/crew preferences | "Jaye Davidson despised the costumes he wore..." | RA5 |
| deleted | PA | deleted scenes | "In a deleted scene in season 5..." | RA5 |
| webpres | PA | the franchise's Web presence | "Stargate has the distinction of being the first film to have an official website." | RA5 |
| effects | PA | special effects, music | "==Visual effects==" | RA5 |
| locations | PA | filming locations | "...filmed in three days in Arizona." | RA5 |
| mistakes | PA | mistakes made, discrepancies | "...Kawalsky is referred to as Major, but he is wearing the rank insignia of an Air Force Captain." | RA5 |
| reception | PA | critical reception | "==Release and reception==" | RA5 |
| fans | PA | fan following | "Fans of the character set up campaigns..." | RA5 |
| merch | PA | merchandising | "A wide area of merchandise is available for the Stargate franchise." | RA5 |

Continued on Next Page...

Table A.18 – Continued

| Name | Type(s) | Meaning | Example | RQ(s) |
|---|---|---|---|---|
| quotes | PA | crew/cast/char quotes | “Quotes:” | RA5 |

Table A.19: Qualitative information

| Name | Type(s) | Meaning | RQ(s) |
|---|---|---|---|
| disclaimers | TXT | copyright, licensing, about us, history, etc. pages | RC1-2,7 |
| homepages | TXT | home/index pages for each site’s sub-sections | RC1-2,7 |
| templates | TXT | HTML/CSS templates of each site’s sub-sections | AA1, AA6, RA7 |

# Appendix B

# Re-expression and rotation of the Fry Graph

Fry (1968) introduced a simple chart for educators and medical professionals to use in determining the readable grading level of texts. He took samples of many textbooks used at different public school grade levels, counted their average numbers of sentences ($x$) and syllables ($y$) in 100-word samples, plotted – by hand, and not using a consistent scale on the y-axis – those variables against each other, and drew by hand a curved line between the clusters of data points. People inside academia and out have been using that graph and more attractive versions of it for the past 40 years. The original Fry Readability Graph (FRG) can be found in figure B.1.

Though this graph is convenient for manually determining the score of a small number of texts, for this dissertation, FRG scores for over 5,000 text samples were required. Re-expressing curved graphs to make them more linear, so that linear modeling assumptions are not violated, is a common task in exploratory statistics, one that was applied in some form to nearly every variable in this project's analysis (discussed in section 3.4.4). Additionally, if the line can also be rotated using

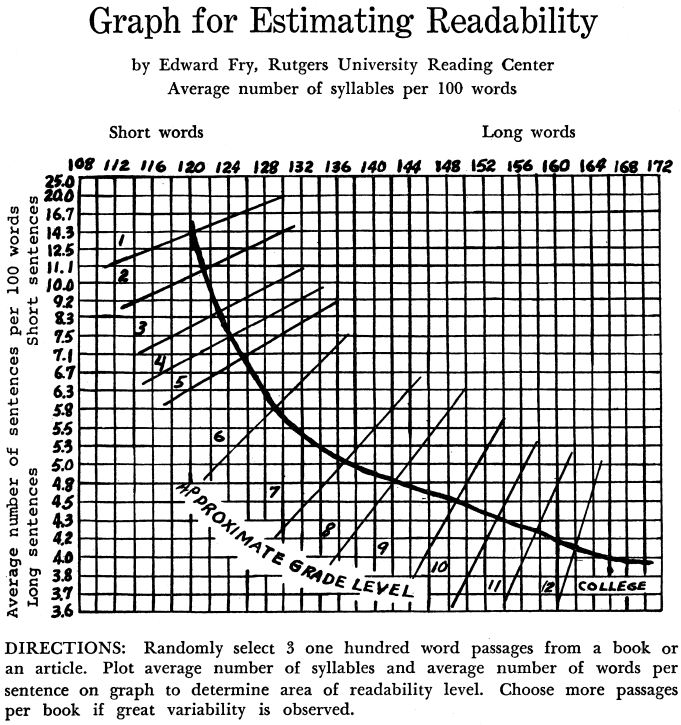



Figure B.1: Fry's Readability Graph

trigonometric functions, it should be possible to divide the line into discrete grade categories (as indicated on the FRG by lines perpendicular to the curve), such that the value of only a single transformed $x$ variable is necessary to automatically categorize a text.

To capture the chart's data, the points where each grade level begins on the line – e.g., grade one begins at about (120, 14.3) – were read and brought into the R statistical environment. Although the original graph appears to have two distinct slopes, this is only due to its inconsistent scale on the y-axis. The FRG with a consistent y-axis is shown in figure B.2.

An exponential transformation of -2.35, or a negative cube root for a similar result and simpler calculation, was found to be optimal for linearizing this line, following the method of Hoaglin et al. (1983, pp. 97-127). Figure B.3 shows the transformed line bisected by an iteratively re-weighted least squares regression line, using the `rlm` function from the MASS library in R. Each point on the line represents a grade level on the original Fry Graph.

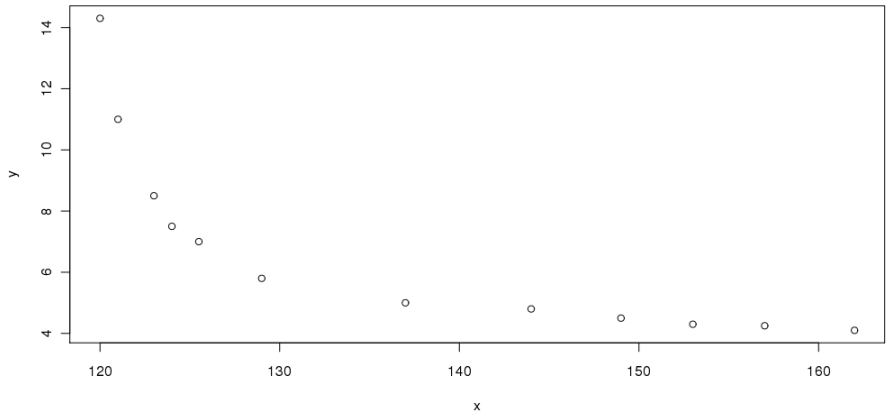



Figure B.2: Fry Graph with consistent axes

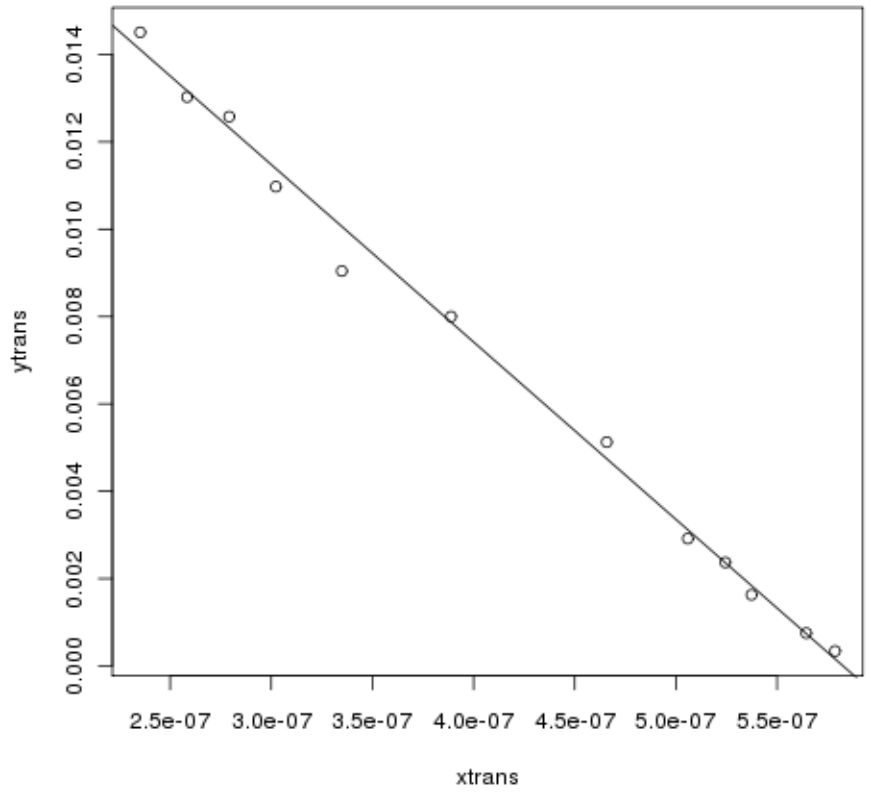



Figure B.3: Fry Graph re-expressed as a line

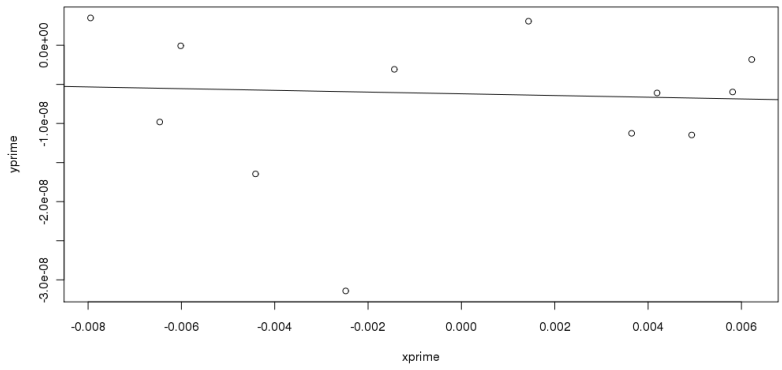



Figure B.4: Final re-expressed and rotated Fry Graph

To rotate the line counter-clockwise to the horizontal, since trigonometric functions are typically centered about the origin, the line was centered about its median, and the trigonometric equality for "angle of inclination," $\theta = arctan(slope)$, was used to find the angle of inclination in radians. The absolute value of the slope was used, because the direction of rotation in two dimensions is controlled by the following matrix operation: $xprime = xcos(\theta) - ysin(\theta)$, and $yprime = xsin(\theta) + ycos(\theta)$, where *xprime* and *yprime* are the rotated coordinates. The result is the essentially straight and horizontal line – i.e., the y-axis is only on the order of $10^{-8}$, and the slope is $-1.096524e-07$ – seen in figure B.4, again with an `rlm` regression line.

The existence of this line, and the $x$ positions of its points, means that, after transforming and rotating the original sentence ($x$) and syllable ($y$) values into this space, one can determine the Fry readability level of the text using only the *xprime* values. For example, if *xprime* is less than or equal to the left point ($x \leq -0.007946735$), then the text is in grade/category one. The following Java-like pseudo-code demonstrates how to use this transformation and rotation process on any appropriate $x$ and $y$ values:

```
double x; //syllables per 100 words
```

```
double y; //sentences per 100 words


double xtranscent = ((1.0/Math.pow(x,3.0)) - 4.27367e-07); //centered, transformed x

double ytranscent = ((1.0/Math.pow(y,3.0)) - 0.006562631); //centered, transformed y

double xytranstheta = 1.570772; //degree of rotation in radians

double xprime = (xtranscent*Math.cos(xytranstheta)) -

                (ytranscent*Math.sin(xytranstheta)); //xprime


//xprimes converted into Fry grade levels

if (xprime <= -0.007946735) System.out.println("1");

else if (xprime > -0.007946735 && xprime <= -0.006464033) System.out.println("2");

else if (xprime > -0.006464033 && xprime <= -0.006014878) System.out.println("3");

else if (xprime > -0.006014878 && xprime <= -0.004411306) System.out.println("4");

else if (xprime > -0.004411306 && xprime <= -0.002479615) System.out.println("5");

else if (xprime > -0.002479615 && xprime <= -0.001437369) System.out.println("6");

else if (xprime > -0.001437369 && xprime <= 0.001437369) System.out.println("7");

else if (xprime > 0.001437369 && xprime <= 0.003647179) System.out.println("8");

else if (xprime > 0.003647179 && xprime <= 0.004192260) System.out.println("9");

else if (xprime > 0.004192260 && xprime <= 0.004934298) System.out.println("10");

else if (xprime > 0.004934298 && xprime <= 0.005811316) System.out.println("11");

else if (xprime > 0.005811316 && xprime <= 0.006220658) System.out.println("12");

else if (xprime > 0.006220658) System.out.println("13");

else System.out.println("0");
```

# Appendix C

# Summary of conclusions

The following tables summarize the conclusions reached throughout this dissertation, both for each website, across all of the websites, and for each research sub-question. The codes in the RsQ column correspond to the research sub-question codes given throughout the Literature and Results chapters. Also, for consistency, the results have been grouped into several tables, following the site-size, editorial models, etc. categories presented in the Discussion and Conclusion chapters. Findings common to all of the sites are presented in the "Common" column, and in the same row as the most similar site-specific findings.

Table C.1: Summary of conclusions: Site size-related

| RsQ | GateWorld | IMDb | Wikia | Wikipedia | Common |
|---|---|---|---|---|---|
| AF1 | topically focused navigation | folksonomies and search engines | topically focused navigation | folksonomies and search engines | navigation only went to franchise/series-level, after which browsing was necessary |

Continued on Next Page...

Table C.1 – Continued

| RsQ | GateWorld | IMDb | Wikia | Wikipedia | Common |
|---|---|---|---|---|---|
| AF4 | many WCAG errors, esp. on obscure pages | fewer WCAG errors | many WCAG errors, esp. on obscure pages | fewer WCAG errors | most WCAG errors were for inaccessible JavaScripts, or media files and tables without textual alternatives |
| RA2 | more critical interpretations | more production, biography, and reception info | more critical interpretations | more production, biography, and reception info | |
| RA3 | strongest on interpretion | weakest on interpretion, though cultural reference sections contained some user interpretations | interpretive sections common on plot/event pages | weak on interpretation | at least an interpretive level of investigation |
| RA5 | prioritized rich fan experience | prioritized high-level overview | prioritized rich fan experience | prioritized high-level overview | |
| RA5 | | most info came from production and marketing companies | | most info came from production and marketing companies | |
| RA6 | provided release dates of non-episodes | | provided release dates of non-episodes | | provided episode release dates |
| RA6 | | links to official sites common | | links to official sites common | book, cast, crew, and game pages shared title and people-related fields across all sites |

Continued on Next Page...

Table C.1 – Continued

| RsQ | GateWorld | IMDb | Wikia | Wikipedia | Common |
|---|---|---|---|---|---|
| RA6 | had unique transcript, review, and making-of pages | lists of cultural references | pages on cultural references | pages on cultural references | cast, character, crew, and author pages had many similar fields everywhere except GateWorld |
| RA7 | ads on IT | diverse ad portfolio | ads on sci-fi and retail | | standard set of vendor categories |
| RA8 | positively biased fan ratings | | positively biased fan ratings, though some cast/crew not rated highly | | |
| RA10 | perhaps known as the place-to-go for esoteric info | title pages in the reviewer style, and pages listing quotes by characters, had highest PageRanks | more esoteric info, lists of biometric info | more textual and substantive, even in lists | high PageRanks or inlinks usually meant more substantive content, low meant more obscure content |
| RC3 | often supported production details with a link to the producer's weblog | | often supported production details with a link to the producer's weblog | | interviews either supported arguments/quotes or gave crew member or game developer views |
| RC4 | a few devoted fans made encyclopedic and obscure pages | | a few devoted fans made encyclopedic and obscure pages | small group of fans made most popular pages, many authors made encyclopedic pages | |
| RC4 | JPGs common on episode pages | | PNGs common on episode pages | | |

Continued on Next Page...

Table C.1 – Continued

| RsQ | GateWorld | IMDb | Wikia | Wikipedia | Common |
|---|---|---|---|---|---|
| RC5 | fewer sections per page | more sections per page | fewer sections per page | more sections per page | |
| RC5 | | more lists of trivia on average | fewer lists of trivia than average | more lists of trivia on average | no info-organizational technique was more common than another |
| RC5 | | many pages empty of content | some pages empty of textual content | some pages empty of textual content | |

Table C.2: Summary of conclusions: Editorial model-related

| RsQ | GateWorld | IMDb | Wikia | Wikipedia | Common |
|---|---|---|---|---|---|
| AF1 | directed users with questions to FAQs or forums | directed users with questions to FAQs | directed users with questions to community | directed users with questions to community | |
| AF2 | older image formats | older image formats | newer image formats | newer image formats | episode pages had more photos and common media formats |
| AF3 | no broken links | no broken links | moderate broken links | many broken links | |
| AF4 | much invalid HTML | much invalid HTML | little invalid HTML | little invalid HTML | WCAG errors more common than HTML errors |
| AF5 | text at 7-9th grade level | text at 7-9th grade level | text at 8-9th grade level | text at 8-9th grade level | six common writing styles |

Continued on Next Page...

Table C.2 – Continued

| RsQ | GateWorld | IMDb | Wikia | Wikipedia | Common |
|---|---|---|---|---|---|
| AF6 | more opportunistic, less use of international standards | more opportunistic, less use of international standards | less opportunistic, more use of international standards | less opportunistic, more use of international standards | |
| RA1 | pages updated at editorial whim | pages updated at editorial whim | pages updated every 2/3 year on average | pages updated quarterly on average | |
| RA1 | | | pages updated in spurts, in sync with academic calendar: small edits over semester breaks, new page creation in late summer | pages updated continuously | |
| RA4 | presented critical reception info in lists | presented critical reception info in lists | | discussed critical reception info in paragraphs | |
| RA5 | citations supported claims or referenced an affiliate | citations supported claims or referenced an affiliate | more variety & quantity of resources | more variety & quantity of resources | |
| RA6 | | | provided dates about editorial and production processes | provided dates about editorial and production processes | at least one long summary text per page |
| RA7 | Amazon and iTunes downloads | evidence for both stable and opportunistic business partnerships | users may be college students, European | users may be young professionals, American | most targeted users were probably tech-savvy males in their 20s |

Continued on Next Page...

Table C.2 – Continued

| RsQ | GateWorld | IMDb | Wikia | Wikipedia | Common |
|---|---|---|---|---|---|
| RA11 | | | crew pages linked to other page types | cast, crew, and episode pages linked to many other page types | each site more often linked to commercial external sites than to small fansites |
| RC1, RC2, RC6 | editors owned all content | editors owned all content | contributors owned their own content | contributors owned their own content | |
| RC3 | | | identified people who work in multiple sci-fi franchises | identified people who work in multiple sci-fi franchises | episode pages linked to preview and behind-the-scenes footage, and info about the real-world version of a topic in the episode |
| RC3 | encyclopedic pages often linked to the Encyclopedia Mythica | | encyclopedic pages either cross-referenced GateWorld's encyclopedia or linked to Wikipedia for beyond-Stargate info | encyclopedic pages usually linked to GateWorld, Wikia, or small fansites | most pages had a group of external links to affiliate/similars |
| RC4 | | | popular episode pages linked to transcript or screenplay PDFs | popular episode pages linked to transcript or screenplay PDFs | the presence of a photo or PDF on a page usually indicated higher author counts/investment |

Continued on Next Page...

Table C.2 – Continued

| RsQ | GateWorld | IMDb | Wikia | Wikipedia | Common |
|---|---|---|---|---|---|
| RC4 | long episode pages in book/author writing style with few external links | | long episode pages in reviewer writing style with many external links | long episode pages in reviewer writing style with many external links | |
| RC5 | page sections more formulaically applied | page sections more formulaically applied | page sections less formulaically applied | page sections less formulaically applied | |
| RC5 | | | standard sentence length | standard sentence length | |

Table C.3: Summary of conclusions: Business model-related

| RsQ | GateWorld | IMDb | Wikia | Wikipedia | Common |
|---|---|---|---|---|---|
| RA4 | described substantive aspects of characters' contexts | | | described substantive aspects of characters' contexts | |
| RA7 | relatively few ads | many ads | many ads | only site not to advertise | |
| RA11 | encyclopedia pages received the most inlinks (followed by episodes and books), and were a clique | character pages were a clique, and often linked to title pages | actor and encyclopedic pages formed a clique | Episode and general (followed by encyclopedic) pages received the most inlinks from other Wikipedia pages | each site had a multi-core network of strong connections between pages of different types |

Continued on Next Page...

Table C.3 – Continued

| RsQ | GateWorld | IMDb | Wikia | Wikipedia | Common |
|---|---|---|---|---|---|
| RA11 | the four sites most often linked to the GateWorld encyclopedic pages | | game pages received many inlinks | the four sites often linked to Wikipedia actor, episode, science, and nature pages | commercial links occurred alongside links to/from northern Europe, wiki/fansite links to/from continental/eastern Europe |
| RC1, RC2, RC6 | named ownership | only founders and top executives were named | only founders and top executives were named | named ownership | |
| RC1, RC2, RC6 | described org structure, funding sources, and affiliations in detail | affiliations only described in the abstract | affiliations only described in the abstract | described org structure, funding sources, and affiliations in detail | lengthy descriptions of org purposes, intentions, and histories |
| RC1, RC2, RC6 | | took greater pains to obscure their history and past performance | took greater pains to obscure their history and past performance | only site without business affiliations | vagueness about physical location |

Table C.4: Summary of conclusions: Fan-made site-related

| RsQ | GateWorld | IMDb | Wikia | Wikipedia | Common |
|---|---|---|---|---|---|
| RA3 | decent production details | | decent production details | strong production details and biographies | |

Continued on Next Page...

Table C.4 – Continued

| RsQ | GateWorld | IMDb | Wikia | Wikipedia | Common |
|---|---|---|---|---|---|
| RA3 | only site to list unanswered questions | | biographical and historical accounts rare, only on pages about people | discussed public reception at length; pages that focused on reception info provided few original interpretations | |
| RA4 | themes on title pages largely agreed with all except IMDb | | themes on title pages largely agreed with all except IMDb | themes on title pages largely agreed with all except IMDb | |
| RA4 | character pages thematically closer to wikis than to IMDb | | included cursory/Infobox details | included cursory/Infobox details | |
| RA4 | focused on depictions of academia, explosions, film production | | texts often about themes of travel and medicine | texts often about themes of travel and medicine | |
| RA6 | episode and character pages had production notes | rarely had more than one long summary text per page | episode and character pages had production notes | episode and character pages had production notes | covered episode and character pages |
| RA10 | low PageRanked episode pages had more trivia | character pages may be the target of other highly PageRanked sites | low PageRanked episode pages had more trivia | low PageRanked episode pages had more trivia | |
| RA10 | high PageRanked episode pages had info for the general public | | high PageRanked episode pages had info for the general public | high PageRanked episode pages had info for the general public | high PageRanked pages had simpler language |

Continued on Next Page...

Table C.4 – Continued

| RsQ | GateWorld | IMDb | Wikia | Wikipedia | Common |
|---|---|---|---|---|---|
| RA10 | encyclopedic pages about people and tech had highest PageRanks | | encyclopedic pages about people and tech had highest PageRanks | encyclopedic pages about people and tech had highest PageRanks | |
| RC3 | GateWorld linked to IMDb, instead of providing cast/crew pages | (no data available) | cast and crew pages often deferred to Wikipedia and IMDb actor pages, or other index sites | cast and crew pages often linked to IMDb or other index sites | book and author pages linked to either sample or unpublished content |
| RC4 | cast and crew given little attention | cast and crew given much attention | cast and crew given little attention | cast and crew given little attention, except for a few fans' obsessing over a few cast members | many authors focused on general details about cast and crew, only a few focused on more obscure trivia |
| RC5 | longest pages; could have longer pages than Wikia | shortest pages on average | long pages; could have shorter pages than GateWorld | longest pages on average | most pages had few links and were relatively short |

Table C.5: Summary of conclusions: Unique to each site

| RsQ | GateWorld | IMDb | Wikia | Wikipedia | Common |
|---|---|---|---|---|---|
| RA1 | | | | most pages were created towards the end of SG-1 | |
| RA5 | | | | only site to cite academic literature | |

Continued on Next Page...

Table C.5 – Continued

| RsQ | GateWorld | IMDb | Wikia | Wikipedia | Common |
|---|---|---|---|---|---|
| RA8 | balanced editor ratings | (no data available) | | (no data available) | |
| RA8 | editor and fan ratings usually agreed | | most wikia fans were probably "syndication" viewers | | |
| RA8 | editors were "high-anticipation" viewers | | | | |
| RA8 | most fans "connoisseuringly" preferred normal or obscure episodes | | | | |
| RA9 | "high-anticipation" viewers preferred pages about premiers, with many inlinks, links to supplementary material, and interrogative texts | (no data available) | "syndication" viewers preferred mid-season premiers and finales | (no data available) | |
| RA9 | "connoisseuring" viewers preferred normal episode pages with links to editorial reviews and lengthy editorial content | | highly rated encyclopedic pages about characters, places, and tech contained quick summaries of those topics; highly rated encyclopedic pages about people had lengthy texts and many images | | |

Continued on Next Page...

Table C.5 – Continued

| RsQ | GateWorld | IMDb | Wikia | Wikipedia | Common |
|---|---|---|---|---|---|
| RA11 | | | | | pages about actors, series, lists, cultural references, etc. acted as network boundary pages |
| RC3 | | | | episode pages usually linked to other large orgs | game pages often linked to fan communities and knowledge bases |
| RC4 | | | more authors on pages in interpretive reviewer style | actor pages in interrogative reviewer style had many authors | |
| RC4 | | | encyclopedic pages with many authors had low PageRanks, PageRanks highest on lengthy pages in the reviewer or general documentation styles and episode pages in book/author style | preferential attachment effects on actor and episode pages | more authors usually meant more revisions |

## Jonathan D. Warren

February 25, 2011

**Objective** **Research data analysis for social or environmental service organizations**


**Personal information**

**Office:** 1320 E 10th St, LI 011
Bloomington, IN 47405

**E-mail:** jonathan@warren.info
**Web:** warren.info


**Education**

**Indiana University**, Bloomington, Indiana, 47405

**Ph.D., Information Science**, May 2011

- **Minor:** Applied Statistics
- **Core subjects:** Exploratory data analysis (i.e., statistical data mining), network analysis, longitudinal and multi-level analyses, macro-scale sociology
- **GPA:** 3.975

**Masters, Information Science**, Aug 2008

**Reed College**, Portland, Oregon, 97202

**B.A., Religious Studies**, May 2002

- **Core subjects:** South Asian and Chinese philosophy and history, hermeneutics, ethnography
- **GPA:** 3.5

**School for the Creative and Performing Arts** at Lafayette High School, Lexington, Kentucky, 40503

**Diploma with honors**, May 1998

- **GPA:** 4.0

**Professional experience**

**Adjunct Lecturer** **Sep 2005 - present**

**School of Library and Information Science, Indiana University, Bloomington, IN**

- **Information Architecture for the Web**: every fall and spring, taught 25-50 masters students about Web programming, accessibility, information architecture, requirements analysis, project management, and graphic design.
- **XML workshop**: every summer, taught around 10 masters students how to create markup languages using XML-related technologies.
- Occasionally taught around 150 masters students about information retrieval, network computing, Java, and enterprise website project management.
- **Address:** 1320 E 10th St, LI 011, Bloomington, IN 47405
- **Research supervisor:** John Paolillo, Associate Professor (812-322-4847)
- **Teaching supervisor:** Howard Rosenbaum, Associate Dean (812-855-3250)

**Logistics Coordinator** **Oct 2003 - Apr 2005**

**Mercer Delta Consulting (now Oliver Wyman), Portland, OR**

- Managed daily operations of a $20 million business unit's global supply chain, including: planning, execution, and efficiency analysis of warehousing, fulfillment, and shipping. Until April 2005, I consulted remotely as needed.

- Supported the organization through 600% growth in demand, by creating and revising logistics-related information systems, standard operating procedures, and deliverables both in the office and in collaboration with global vendors and suppliers.
- Helped research and plan the business unit's migration from desktop publishing software to enterprise-scale software.
- Both hired and trained my successor, and I believe they have been satisfied with her.
- **Address:** 1631 NW Thurman Street, Suite 100, Portland, OR 97209
- **Supervisor:** Mark Allen, Director of Resource Management (503-419-5300). Mr. Allen is no longer with the company; this is the receptionist's number.

**Trainer, Customer Service Representative** **Sep 2002 - Oct 2003**

**Blockbuster District Training, Portland, OR**

- Trained Customer Service Representatives and entry level Store Managers for 160 retail stores in Oregon.
- Learned warehousing and retail industry standard operating procedures, customer service, conflict resolution, risk management, just-in-time (JIT) inventory management, and barcoding.
- **Address:** 7535 SW Barnes Rd #109, Portland, OR 97225
- **Supervisor:** Eve Cason, Store Manager (503-296-9900, main store number)

**Media Archivist, Clerk** **May - Aug 2001**

**Reed College Library, Portland, OR**

- Troubleshot Recordbuilder database errors.
- Entered electronic journal catalog data using FileMaker Pro.
- Organized the college's media storage by the Library of Congress' system.
- Maintained 10 language lab and video editing stations.
- **Address:** 3203 SE Woodstock Blvd, Portland, OR 97202
- **Supervisor:** Betty Woerner, Media Librarian & Heather Whipple, Electronic Resources Librarian (503-777-7352, current Media Center number)

**Hardware and Network Technician** **May - Aug 2000**

**Micro Computer Analysts, Lexington, KY**

- Designed, built, and serviced Windows 2000, NT, and 98-based servers, workstations, and networks for clients including: AT&T, @Home, and the University of Kentucky.
- Attended MCSE and A+ training courses.
- **Address:** 128 Southland Dr, Lexington, KY 40503
- **Supervisor:** Fred Wachs, Business Manager (859-275-1228, receptionist's number)

## Statistical expertise

- **General and generalized linear models** (multiple regression, logistic regression, robust techniques, etc.)
- **Latent variable models** (principal components analysis, factor analysis, canonical correlation, etc.)
- **Longitudinal multi-level / mixed-effects models**
- **Social and semantic network analysis**
- **Clustering and smoothing techniques**

## Computer skills

**Operating systems:** most Microsoft and Apple OSs since 1987, Debian/Ubuntu (expert); FreeBSD, RedHat, Solaris, and SuSE (proficient)

**Programming languages:** Java (expert); Bash, C, Javascript, Perl, PHP, Prolog, SQL, XSLT (proficient); ASP, Python (basic)

**Statistics packages:** R / S-Plus (expert); Excel, SPSS (proficient); Matlab, Mathematica, M-plus, Mx, SAS, StatView, VARBRUL (basic)

**Markup languages:** XML-related technologies (DTDs/Schema, RDF, ontologies, XLink, XPath, XQuery, Xerces, oXygen, etc.) and (X)HTML/CSS (expert); LaTeX and BibTex (proficient); SGML, Postscript (basic)

**Project management:** custom Web 2.0 software created by Mercer Delta, eGroupWare (expert); MS Project, MS SharePoint, MS Visio (proficient)

**Publishing, graphic design:** MS Office, OpenOffice, WordPerfect (expert); Photoshop, Illustrator, Gimp, Inkscape (proficient); InDesign, QuarkExpress, AutoCAD, Flash, Fireworks (basic)

**Servers & databases:** Apache, Tomcat, MySQL, PostgresSQL (proficient); MS IIS, MS SQL, MS Access, MS Exchange, MS Navision/Dynamics, Oracle (basic)

**Content management systems:** Drupal, Joomla, Wordpress (proficient); IBM Content Manager, Sakai (basic).

## Peer-reviewed publications

Warren, J; Stoerger, S; and Kelley, K. (undergoing revisions). Longitudinal gender and age bias in a prominent amateur new media community. *New Media & Society.*

Paolillo, J; Warren, J; and Kunz, B. (2010). Genre emergence in amateur Flash. In Mehler, A., Sharoff, S., & Santini, M. (Eds.), *Genres on the Web: Computational Models and Empirical Studies.* New York: Springer.

Paolillo, J; Warren, J; and Kunz, B. (2007). Social network and genre emergence in amateur Flash multimedia. *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, IEEE Computer Society Press, 70b.

Warren, J. (2002). Vipassana in the Pacific Northwest. Thesis. Reed College: Portland, OR.

## Peer-reviewing activity

(2010, January). Reviewed an article on online hate networks for JASIST.

(2009, December). Reviewed an article on Twitter use by the US Congress for JASIST.

## Academic service

**Indiana University**, Bloomington, Indiana, USA

**Doctoral Steering Committee** **2008-2009**

- Reviewed issues of importance to the SLIS doctoral program. The committee was comprised of the Director of the Doctoral Program, three faculty members assigned by the Dean, and two voluntary doctoral students.

**Participated in faculty searches** **2008**

**Curriculum Steering Committee** **2005-2008**

- Reviewed proposals for new SLIS courses. The committee was comprised of the Associate Dean, three faculty members assigned by the Dean, three elected masters students, and one voluntary doctoral student.

**SLIS_PHD listserv moderator** **2007-2010**

- Moderated the school's listserv for doctoral students

## Fellowships, grants, & awards

Hawaii International Conference on System Science (HICSS, is peer-reviewed)

- Nominated for best paper, 2007

Indiana University

- Margaret Griffin Coffin fellowship, 2008-2009
- Federal GAANN fellowship, 2007-2008
- Chancellor's Scholar award, 2005
- Chancellor's fellowship, 2004-2008
- SLIS grant for travel to Association for Library and Information Science Education (ALISE) conference, 2009
- SLIS grants for travel to Sunbelt Social Networking and American Society for Information Science & Technology (ASIS&T) conferences, 2008
- SLIS grant for travel to HICSS, 2007
- Doctoral Student Research Forum: finalist for best presentation, 2006