Scholarly and Research Communication

VOLUME 3 / ISSUE 1 / 2012

Selected Information Management Resources for Implementing New Knowledge Environments: An Annotated Bibliography

Alex Garnett
University of British Columbia
Ray Siemens
University of Victoria
Cara Leitch
Electronic Textual Cultures Lab
Julie Melone
Electronic Textual Cultures Lab
INKE and PKP Research Groups1

Abstract

This annotated bibliography reviews scholarly work in the area of building and analyzing digital document collections with the aim of establishing a baseline of knowledge for work in the field of digital humanities. The bibliography is organized around three main topics: data stores, text corpora, and analytical facilitators. Each of these is then further divided into sub-topics to provide a broad snapshot of modern information management techniques for building and analyzing digital documents collections.

Keywords

Digital humanities; Digital document Collections; Data stores; Text corpora; Analytical facilitators

The INKE Research Group comprises over 35 researchers (and their research assistants and postdoctoral fellows) at more than 20 universities in Canada, England, the United States, and Ireland, and across 20 partners in the public and private sectors. INKE is a large-scale, long-term, interdisciplinary project to study the future of books and reading, supported by the Social Sciences and Humanities Research Council of Canada as well as contributions from participating universities and partners, and bringing together activities associated with book history and textual scholarship; user experience studies; interface design; and prototyping of digital reading environments.

Alex Garnett is a PhD Student in the School of Library, Archival, and Information Studies at the University of British Columbia, Suite 470-1961 East Mall, Vancouver, BC, Canada V6T 1Z1.
Email: axfelix@gmail.com .

Ray Siemens is Canada Research Chair in Humanities Computing and Distinguished Professor in the Faculty of Humanities in English with cross appointment in Computer Science at the University of Victoria, PO Box 3070 STN CSC, Victoria, BC, Canada V8W 3W1.
Email: siemens@uvic.ca .

Cara Leitch is a PhD candidate in English and a Research Assistant at the Electronic Textual Cultures Lab, University of Victoria, PO Box 3070 STN CSC, Victoria, BC, Canada V8W 3W1.
Email: cmleithc@uvic.ca .

Julie Melone was a INKE Postdoctoral Fellow in the Electronic Textual Cultures Lab at University of Victoria, University of Victoria, PO Box 3070 STN CSC, Victoria, BC, Canada V8W 3W1.
Email: jcmeloni@uvic.ca .

CCSP Press
Scholarly and Research Communication
Volume 3, Issue 1, Article ID 010115, 45 pages
Journal URL: www.src-online.ca
Received August 17, 2011, Accepted November 15, 2011, Published March 26, 2012

Garnett, Alex, Siemens, Ray, Leitch, Cara, & Melone, Julie. (2012). Selected Information Management Resources for Implementing New Knowledge Environments: An Annotated Bibliography. *Scholarly and Research Communication, 3*(1): 010115, 45 pp.

This select, annotated bibliography reviews scholarly work in the area of building and analyzing digital document collections. The goal of this bibliography is to establish the baseline knowledge for work in this area, and to provide a set of select, foundational texts upon which to build future research.

In total, this document contains three bibliography topics: 1. Data Stores; 2. Text Corpora; and 3. Analytical Facilitators. Each of these areas is subdivided into a further three topics for a total of nine subsections containing anywhere from six to sixteen documents each: 1A. Digital Libraries; 1B. Technical Architecture and Infrastructure; 1C; Content Management Systems and Open Repositories; 2A. Corpus Building and Administration; 2B. Document Design and Data Classification; 2C. Lessons Learned in Corpus-Building Projects; 3A. Data Visualization and Geographical Information; 3B. Text Encoding and Analytic Tools; and 3C. Reviewing Humanities Computing. Together, these articles provide a broadly accurate snapshot of modern information management techniques pertinent to research in the area of building and analyzing digital document collections.

The bibliography proceeds from fairly late-breaking discussions of digital repository models, through a review of fairly explicit methodologies for analyzing corpora, into a more theoretical broad-level discussion of the digital humanities and related disciplines. We have, without intending to, become journalists with our inverted pyramid of breadth and depth; for this, we present the following narrative review, such that it may serve, with a little effort, as an index of how and why to best leverage premiere scholarship in information management.

Data stores: A review of select bibliographic sources

The articles in this section move gradually over the past decade from general theories of digital library creation, and justifications for the development of grid infrastructure, to more specific case studies of using "big data" in the humanities and social sciences. This retrospective approach to digital humanities scholarship allows us to observe, for example, the evolution of nascent work on the Greenstone digital library project into an active community of open-source and open-access repository developers (and with them, miniaturized two- and three-page academic publications that would have been unthinkable in the humanities not long ago). Similarly, there is a clear turn in the information sciences from predicting new systems and standards for collaboration to carefully observing why and how certain traditional practices have or have not migrated to these new systems, lending a social perspective to our understanding of what it is that makes certain aspects of scholarly practice truly "digital". Whether their authors are talking about "e-Science" (as in the UK) or "cyberinfrastructure" (as in the United States and Canada), it is heartening to see such a broadly receptive dialogue among computer scientists and traditional humanists alike.

In this area, *D-Lib Magazine* should be the primary scholarly publication consulted, although others certainly cover similar topics. Additionally, the fast-paced nature of work in this field, combined with the relatively slow publication schedule of some scholarly journals, makes necessary a frequent review of presentations and reports to affiliated groups of the Association of Computing Machinery (ACM), such as the *Joint Conference on Digital Libraries*, as well as the UK e-Science community, and the Open Repositories user group.

General theories of digital library creation

The articles in this sub-section focus specifically on the formation of digital libraries, including requirements gathering from both a technical and a user-oriented perspective, implementation of solutions that meet those requirements, and other considerations regarding gathering data for inclusion in digital libraries. The creation of data stores for external use is similar to creating a digital library in its technical considerations; thus, in the literature review for data store creation, there will necessarily be many articles on digital libraries.

Levy, David M. & Marshall, Catherine C. (1995). Going digital: A look at assumptions underlying digital libraries. *Communications of the ACM, 38*(4), 77–84.

The authors discuss the foundational elements of digital libraries: how they will be used, how they should be designed, and the relationship they should have to physical libraries. The authors highlight that regardless of the type of library, the purpose of the library is clear: to house and provide access to documents, where "documents" includes a wide range of objects (and in the digital library, importantly, a wide range of digital objects). The access to these objects is provided through technology, and underlying that selection of technology is the work to be done by the library's users. The authors assert that when selecting and implementing technology for document storage, one has to consider three common assumptions: digital library collections contain fixed, permanent documents; digital libraries are used by individuals working alone; and, digital libraries are based on digital technologies. For each digital library plan, one must interrogate these assumptions and adjust priorities and expectations accordingly. Additional elements of the digital library that also have implications for current and future work include the integration of multimedia as stored objects (e.g. not simply PDFs of articles or documents marked up in XML), as well as data versioning.

Marchionini, Gary. (2000). Evaluating digital libraries: A longitudinal and multifaceted view. *Library Trends, 49*(2), 304–333.

This article summarizes the creation and subsequent decade of development and use of the Perseus Digital Library. As one of the primary digital resources in the humanities since the writing of this article, this report contains many useful lessons for any team looking to create and maintain a large-scale data store. The article details the creation and growth of the library from a HyperCard driven CD-ROM to a fully Web-enabled resource. More importantly, the author discusses how the PDL was not originally conceived as a digital library, but the idea took hold with the increased use of card catalog metadata within each object record, as well as the manner in which users could freely access significant stores of primary source materials. The author concludes by clearly defining three main points for the evaluation of a digital library: (1) efforts must explicate goals ranging from the evaluation of research to product/system testing; (2) efforts must account for and react to the fact that digital libraries are complex systems that must be augmented as both technology and user requirements change; and (3) statistical data, as well as user narratives, must be used to assess impact and performance. The user-facing requirements for evaluation and augmentation will continue to evolve as they have in the decade since this article was published, but the points regarding front-loading planning and evaluation criteria for data stores are well taken.

Thaller, Manfred. (2001). From the digitized to the digital Llibrary. *D-Lib Magazine 7*(2), n.p.

This paper provides another useful set of guidelines for creating and maintaining a digital library, with the focus on ensuring that the digital library is not simply a digitized library. Instead, it is argued, the digital library, while containing digital objects, should be built and evaluated differently than its paper-based counterpart. The author asserts that criteria for digitization (and storage and retrieval) projects should begin with clear criteria based on the use of the resources, and that these criteria should be re-evaluated and change as resource use changes. In addition, other plans for creating a digital library should include planning for large-scale digitization (i.e., of over one million objects) of high quality (i.e., multiple resolutions of digital objects stored as image files). Finally, the author discusses requirements of digital libraries outside the scope of "pure" research, namely, public interfaces, integrated reference systems, and the use of ready-made objects for teaching.

Chowdhury, Gobinda. (2002). Digital divide: How can digital libraries bridge the gap? Lecture notes in computer science. *Digital Libraries: People, Knowledge, and Society, 2555*, 379–391.

The article begins by summarizing the working definition and the state of the world's various "digital divides" in 2002, with an eye toward leveraging digital libraries to help resolve these inequalities. The author notes that physical libraries have themselves been underdeveloped and underutilized for the developing world, particularly with respect to information and communications technology (ICTs) such as public internet terminals, and that the high costs associated with many successful digital library initiatives may not translate well to the developing world. Digital synchronous and asynchronous information delivery mechanisms (e.g., remote reference and subject gateways, respectively) are compared and itemized for their viability in a developing world context, along with then-nascent open access journals, archives, and e-book stores. The author concludes by outlining recommendations for building digital libraries on a limited budget, considering when and where government-backed projects should be outsourced, kept local, or otherwise linked, stressing the importance of improving digital literacy skills.

Coyle, Karen. (2006). Mass digitization of books. *Journal of Academic Librarianship, 32*(6), 641–645.

This article considers several forms of digitization projects that create content that is then stored by others. The author first discusses mass digitization—specifically the Google Books project—and how its goal is not to create collections or maintain anything beyond limited structural mark-up, but instead to digitize everything. This differs from non-mass digitization, which has a specific agenda of preservation. A third form of digitization is "large-scale" projects, which sit somewhere between mass and non-mass digitization projects. The example used of this type of digitization is JSTOR and its goal of creating collections and complete sets of documents (journals, in this case). The author also highlights some of the issues associated with these types of digitization projects, such as adherence to standards (both in file formats and metadata), and the production of preservation-quality digital objects.

Seadle, Michael & Greifeneder, Elke. (2007). Defining a digital library. *Library Hi Tech,* 25(2), 169–173.

Produced several years after the foundational articles presented earlier in this section, this article questions the feasibility of creating a definition of a "digital library" that differentiates data stores from any other electronic resource. The author determines that not only are digital libraries "too young to define in any permanent way" (p. 172), but also that the notions of how users interact with digital content—and the technologies with which they do so—are changing too rapidly to offer a meaningful definition. Rather, it is argued that when creating any large-scale digital resource, be it a data store or a library system, administrators must begin with a set of criteria, a solid plan, and the ability to be flexible in the execution of that plan over time.

Rimmer, Jon, Warwick, Claire, Blandford, Anne, Gow, Jeremy & Buchanan, George. (2008). An examination of the physical and the digital qualities of humanities research. *Information Processing and Management, 44*(3), 1374–1392.

Human-computer interaction (HCI) researchers working on the design of digital reading environments have often questioned how closely these tools should mirror their physical counterparts. The authors report on findings from interviews with humanities scholars on their use of physical and digital information resources. While virtually all respondents are in agreement about the convenience of digital resources, the loss of physical "context" seems to mean different things to different people, ranging from the purely aesthetic (e.g. the excitement of handling ancient texts) to the serendipitous (e.g. having one's interest sparked by physically co-located or otherwise similar resources). The surveyed researchers also demonstrate an awareness of the changing demand for information literacy skills, with mixed opinions on the subject. The tone the authors take is ultimately almost one of sentimentality, with their participants agreeing that digital resources are more reliable, presenting fewer difficulties in resource description and access, but in many cases less pleasurable to actually use. This suggests that the humanities community is aware of the advantages of migrating away from physical resources, but will do so with some regret.

Warwick, Claire, Galina, Isabel, Rimmer, Jon, Terras, Melissa, Blanford, Jeremy, & Buchanan, George. (2009). Documentation and the users of digital resources in the humanities. *Journal of Documentation, 65*(1), 33–57.

This article presents two digital humanities research case studies (User-centered Interactive Search with Digital Libraries, UCIS; and Log Analysis of Information Researchers in the Arts and Humanities, LAIRAH), offering a critical perspective on documentation practices for digital resources. First, the authors distinguish between *technical* (who, what, when, where, why) and *procedural* (how) documentation. They detail recurring issues experienced by the UCIS project in formatting and parsing various mark-up languages for technical documentation. Conversely, the LAIRAH project suffered from an overall *lack* of documentation, and especially procedural documentation, which was undervalued by administrators—except, notably, in the disciplines of archaeology, linguistics, and archival science, which the authors suggest are perhaps better-accustomed to documentation practices—at the expense of novice users. The authors conclude with a discussion of what it means for information resources to be accessible; that is, accessibility requires resources to not only be within logical reach, but also contextually intelligible for novice users, particularly when working with complex, modular documentation.

Marcial, Laura, & Hemminger, Brad. (2010). Scientific data repositories on the Web: An initial survey. *Journal of the American Society for Information Science and Technology, 61*(10), 2029–2048.

Much current research in digital repositories has centred on archiving and reusing, not just of published academic literature, but raw "Big Data." In this article, Marcial and Hemminger (2010) conduct a survey of scientific data repositories (SDRs) on the open Web and develop a framework for their evaluation. They observe, for example, several repository managers' stated intent to capture and index the "dark" (i.e., informal and/or undocumented) data that slips through the cracks of the current scholarly publishing ecosystem, but may still be re-usable or otherwise valuable. Although they identify only four out of 100 surveyed repositories whose scope lies outside the natural sciences, it is shown that a significant majority of SDRs are funded or directly affiliated with individual universities, presenting a clear target for advancement of the digital humanities when working with institutional repositories.

White, Hollie. (2010). Considering personal organization: Metadata practices of scientists. *Journal of Library Metadata, 10*(2), 156–172.

In the interest of making indexed datasets accessible and reusable, this article reports on a small-scale field study of scientists' personal information management practices. Using examples from the *Dryad* data repository with which she is personally affiliated, the author explains that individual datasets originating from different labs often have little more in common than their document format, evidencing the need for descriptive metadata, particularly in disciplines that work primarily with non-standardized, qualitative data. Of the study participants, those who preferred the use of physical information objects were the least inclined to create or use a meta-organizational system, perhaps offering a glimmer of hope for the organization of digital information which may be poorly curated but is nevertheless somehow indexed and searchable without any special effort by the user. For those who preferred to use electronic databases, the most important factor was the ability to view and manipulate the data by some property, which is directly related to the research question itself, highlighting the need for data stores which are tailored to their respective disciplines.

Technical architecture and infrastructure

The articles in this section focus specifically on the technical architecture and infrastructure that supports digital libraries or data stores.

Buyya, Rajkumar and Srikumar Venugopal. (2005). A Gentle Introduction to Grid Computing and Technologies. *CSI Communications, 29*(1), 9–19.

This article is not specific to digital libraries or data stores, but introduces readers to the concept of grid computing, which is often used in digital libraries or data stores. Developers and administrators of large data stores with multiple access methods and user types may be interested in using grid computing, which is an integrated and collaborative technical infrastructure that encompasses machines (processors) and networks (bandwidth) that are managed by multiple organizations, often geographically distributed. Now, readers may be more familiar with the term "cloud computing," which is a form of grid computing.

Balnaves, Edmund. (2005). Systematic Approaches to Long Term Digital Collection Management. *Literary and Linguistic Computing, 20*(4), 399–413.

This article is less about the underlying technical architecture of digital libraries or digital repositories, and more about the continued access to these resources due to licensing of either content or software. The author highlights the issues inherent in the reliance of digital libraries on e-journal subscription contracts and online database vendors, and proposes methods of maintaining rich scholarly archives, while also integrating those maintenance practices with the acquisition practices of the library organization. He outlines some of the risks associated with digital resources, both physical (e.g., fire, media deterioration) and institutional (e.g., agreement expire or voiding), and offers solutions, or at least paths toward mitigating these risks, such as finding alternative suppliers and using open source collections, applying market pressure when the size of the organization warrants it, aligning the organization with large content clearinghouses, interfacing with distributed digital repositories, and implementing content syndication through the use of enterprise-grade content management systems.

Rosenthal, David S. H., Thomas Robertson, Tom Lipkisi, Vicky Reich, and Seth Morabito. (2005). Requirements for Digital Preservation Systems: A Bottom-Up Approach. *D-Lib Magazine, 11*(11), n.p.

Unlike numerous digital preservation models that advocate a top-down approach, the authors propose a bottom-up model for creating systems that remain accessible and stable for the long-term. Of most importance in this article is the authors' "taxonomy of threats," or a set of threats that must be in some way accounted for in system development. Examples of threats include media, hardware, software, and network failures; media, hardware, and software obsolescence; internal or external attacks; operator error; and even economic and organizational failures. Strategies that system architects can and should put into place to survive these threats include data replication, paths for data migration, and system transparency and diversity (e.g., make it clear how systems are put together, and ensure there is sufficient diversity in location, among other factors).

Crane, Gregory, Alison Babeu, and David Bamman. (2007). eScience and the humanities. ***International Journal on Digital Libraries, 7*****(1-2), 117–122.**

In this article, the authors make a call to action for developing a large-scale data architecture for the humanities, noting that any such system must make data "intellectually as well as physically accessible" and citing language barriers as the fundamental challenge for the humanities. Curiously, they point to optical character recognition (OCR, i.e., page-to-digital-text-scanning) and machine translation as a comprehensive solution, despite the fact that the present (and still current) state of the art in machine translation was only sufficient to make the text of a given foreign-language resource intelligible, preserving little of the original text's richness. They go on to extoll the virtues of what is now called "augmented reality" software – that which runs on personal ubiquitous computing (UbiComp) devices such as mobile phones and overlays a layer of geo-locational data on the image captured by the device camera – in the use of historical or anthropological fieldwork. Above all, they stress that humanists must keep abreast of similar infrastructure developments in the natural sciences, both by leveraging new approaches such as crowd-sourcing, and by carefully targeting funding agencies such as the U.S. National Science Foundation (NSF) for collaboration with the sciences.

Gold, Anna. (2007). Cyberinfrastructure, Data, and Libraries, Part 1: A Cyberinfrastructure Primer for Librarians. ***D-Lib Magazine, 13*****(9), n.p.**

As its title suggests, this article serves as a good primer to the cyberinfrastructure needs within a library environment. In this case, the focus is on "e-science" or digitally-enhanced scientific research and communication, but many of the infrastructure issues are similar in humanities work as well: technical architecture, methods for collaboration in a digital environment, computational resources across the grid or in the cloud, data curation, data preservation, and ongoing data management, to name but a few. The author frames her primer as one intended to open up discussion between library practitioners and researchers; to do so, she first provides a brief history of related fields and introduces vocabulary necessary for both groups to communicate successfully with each other. Of particular relevance in 2011 are the sections on data archiving and preservation, curation, access, and interoperability, and the data life cycle.

King, Gary. (2007). An introduction to the Dataverse Network as an infrastructure for data sharing. ***Sociological Methods & Research, 36*****,173–199.**

This article, notably published in a Sociology journal, makes an intriguing claim about data sharing: that its machinations and practices, despite being ostensibly more rigid and quantitative than traditional "analog" scholarship, are not nearly as well understood or officiated as we assume. In order to harmonize digital and "analog" scholarship practices for recognition, distribution, and persistence, the author outlines the infrastructure requirements of the proposed Dataverse Network project. The Dataverse Network is to be a distributed grid, with several independently hosted nodes being indexed by the primary aggregator. Among its other notable features are what the author calls "forward citation" tracking (similar to Google Scholar alerts for tracking the citation of one's own work), and the server-side implementation of the R statistical computing language using the Zelig GUI to promote exploratory data analysis. The author anticipates individual Dataverse nodes being used as ad-hoc syllabi for university courses and other educational opportunities, teaching by data-driven example.

Voss, Alex, Matthew Mascord, Michael Fraser, Marina Jirotka, Rob Procter, Peter Halfpenny, David Fergusson, Malcolm Atkinson, Stuart Dunn, Tobias Blanke, Lorna Hughes, and Sheila Anderson. (2007). e-Research infrastructure development and community engagement. Proceedings from the *UK e-Science All Hands Meeting 2007.* Nottingham, UK.

This article reviews past work on community development in order to identify barriers to adoption of new technologies by humanities and social sciences researchers. The authors begin by discussing fallacies common to the study of socio-technical systems in this respect, noting that the colloquial "early / late adopter" dichotomy is more often applicable to specific circumstances than to individuals, and the design of these systems is rarely as planned or as discontinuous as we tend to characterize them. They detail nascent work funded by JISC (The UK's Joint Information Systems Committee) in defining "service usage models" to improve our understanding of technology adoption. Finally, echoing a commonly stated principle of community development, the authors highlight that these developments in adoption of new technologies should arise from *within* communities rather than be pushed from outside.

Blanke, Tobias, and Mark Hedges. (2008). Providing linked-up access to Cultural Heritage Data. Proceedings from: *ECDL 2008 Workshop on Information Access to Cultural Heritage.* Aarhus, Denmark.

This short workshop paper presents an example of successfully providing infrastructure for access to cultural heritage data using digital library technologies. The authors state with surprising certainty that the sort of humanities data that begets this infrastructure almost always assumes one of two forms: enormous archival TIFF images, and XML-encoded transcriptions of the content of these images. Therefore, the most important consideration for an archival system is the quick browsing and retrieval of this content: linking corresponding files, allowing for the efficient delivery and storage of thumbnail images, and supporting the creation of personalized workspaces that facilitate the creation and use (for example, of dynamic sorting) of new metadata elements.

Crane, Gregory, Brent Seales, and Melissa Terras. (2009). Cyberinfrastructure for Classical Philology. *Digital Humanities Quarterly, 3*(1), n.p.

The article provides a solid overview of several key concepts in cyberinfrastructure and includes practical examples of these concepts. This article would best serve as a gentle introduction to some of the more technical aspects of the digital humanities, especially to scholars new to the field; however, it is included in this bibliography precisely because it does not provide new information. Later readers of this bibliography may see the term "cyberinfrastructure" in its title and assume that it addresses the same technical concepts and concerns as the Anna Gold article in *D-Lib Magazine* (referenced above) or the Nicolas Gold article in a later issue of *DHQ* (referenced below). This article is dissimilar from those two in its technical depth, but does provide an appropriate discussion of features and functionality for a lay audience. Specifically, the authors remind us all—technically inclined or otherwise—that when "our infrastructure advances incrementally, we may take it for granted" (n.p.), which is problematic as it "does not simply affect the countless costs/benefit decisions we make every day—it defines the universe of what cost/benefit decisions we can imagine" (n.p.). The authors then provide several examples of digital projects that require substantial infrastructure, including digital incunabula, machine-actionable knowledge bases, and digital communities, before providing even more concrete examples of how these projects are used, namely, to produce new knowledge and to extend the intellectual reach of humanity.

Borgman, Christine. (2009). The digital future is now: A call to action for the humanities. *Digital Humanities Quarterly, 3*(4), n.p.

In what could be called "recession-era scholarship," Christine Borgman (2009) issues a supplication to the digital humanities to produce clearly defined goals for advancement in light of limited funding, particularly with regard to data infrastructure and value propositions. She provides a brief history of the development of the digital humanities since 1989, noting that digital scholarship is still segregated from other humanities research in many respects, leaving questions about publishing and tenure still largely unresolved. Given that the agreed-upon best practices for digital libraries have so far produced raw data stores that do not provide any obvious affordance for inexperienced researchers, she concludes that librarians and archivists must remain a valued part of accumulated humanities methods and practice. She voices regrets that the humanities have so far failed to realize some benefits of electronic publishing (such as the agile pre-print sharing practices encouraged by arXiv.org) because digital scholarship in this realm has so far been handicapped by the limits of a preference for print materials. Her principle recommendation for resolving these issues is to focus on the question of what "data" means to humanists, and with it, to encourage documentation practices to facilitate sharing and networked learning. She concedes that her work is premised on a belief that the traditional model of a wizened scholar labouring alone is increasingly dysfunctional, still a difficult proposition for many humanists.

Hicks, Diana, and Jian Wang. (2009). Towards a bibliometric database for the social sciences and humanities. URL: http://works.bepress.com/diana_hicks/18/ [July 15, 2011].

This article, which appears to have been self-published using the SelectedWorks system and may not be peer-reviewed, ironically takes as its subject matter the long-standing issue of unreliable bibliometric authority indicators in the social sciences and humanities. Although the *de facto* bibliometric standard *Web of Science (WoS)* maintains the authoritative *Social Science Citation Index (SSCI),* which is functionally similar to the self-explanatorily dominant *Science Citation Index,* its coverage is much more irregular. The authors identify a number of reasons for discrepancy, including a disagreement over scholarliness across national and international literatures, as well as differences in the evaluation of the impact of journal articles versus other monograph material. All of these serve to perpetuate a systemic overvaluing of SSCI indexing wherein authors and editors alike compete for a prize they may find philosophically objectionable. The authors extoll the virtues of Google Scholar as a powerful resource, which has nevertheless focused on "findability" at the expense of any curated, evaluative bibliometrics, undermining the SSCI in practice (i.e. literature searching) but not in theory (i.e. tenure evaluations). They conclude with detailed statistics on the coverage of various indexing databases, affirming WoS' stature as the most "exclusive" of these databases. In so doing, they provide a rare but expressly negative effect of this: surprisingly poor coverage of non-English humanities materials.

Terras, Melissa M. (2009). The Potential and Problems in using High Performance Computing in the Arts and Humanities: the Researching e-Science Analysis of Census Holdings (ReACH) Project. *Digital Humanities Quarterly, 3*(4), n.p.

This article is a thorough report of a series of workshops intended to bring together an interdisciplinary group to investigate the potential application of grid computing to a large dataset, in this case, historical census records. The results of the workshop, specifically, the description of the benefits that such computing resources would bring about for scholars, are perhaps less authoritative than the questions raised regarding the administration of computing resources and the rights to the data both ingested and produced. It comes as no surprise that when asked for a wishlist of tools and processes for working with such a large dataset as a census, scholars developed a list that ranged from cleaning and managing records to producing algorithms and models for a longitudinal database of individuals throughout the census. Moving from the idea phase to that of technical implementation, the author notes that the "technical implementation [to] perform data manipulation, and output data, is much less of a problem than identifying the research question" (n.p.), then continues on to discuss the specifics of a possible implementation. An important aspect of this discussion is the security requirements needed for working with "commercially sensitive" datasets, and how these factors necessarily limit the use of a distributed, grid-enabled model for computational resources. Finally, the author raises the issue of the fair use and application of data during and after a project, especially when commercial datasets are involved in the production of new knowledge.

Hedges, Mark. (2009). Grid-enabling Humanities Datasets. *Digital Humanities Quarterly, 3*(4), n.p.

This article provides a solid foundational definition of data and infrastructures that are "grid-enabled" and provides example applications of use in and for humanities research. The author describes the grid using an "analogy of public utilities, for example an electricity grid, where a consumer can connect a diversity of electrical appliances, making use of open and standard interfaces (e.g., a plug), and consume electricity, without knowing or caring about its origin." Thus, to the consumer, the electricity that results is their only reality, not all that comes before it (or is behind it). Cloud computing is also mentioned as a similar type of technology: to the end user, storage "in the cloud" means that their data is housed and maintained elsewhere—possibly within a distributed network, possibly not—and they can access this data from various clients, interfaces, and locations; the technology that powers all of these actions is not of concern to them. The author highlights that although humanists have done a good job of producing large datasets that are accessible from beyond their home institution, the tools for carrying out new modes of research—and thus creating new knowledge—have "lagged behind." Two projects—LaQuAT (Linking and Querying Ancient Texts) and gMan—are then discussed in terms of the technologies used and the relationship of the projects to grid-enabled computing (where it could or does enable research and where it breaks down). The conclusions made with regard to grid-enabled computing in the humanities are not a surprise: technology can enable access and discovery, but humanities research is inherently interpretive and not scientific. While there are significant advances that can be made by creating and extending networks of datasets, repositories, and tools, the "best" we can do in the humanities is to aggregate questions and answers, not provide definitive ones.

Blanke, Tobias, and Mark Hedges. (2010). A Data Research Infrastructure for the Arts and Humanities. In Simon C. Lin and Eric Yen (Eds.), *Managed Grids and Cloud Systems in the Asia-Pacific Research Community* (pp. 179–191). Boston, MA: Springer.

This article presents an in-depth look at an existing research infrastructure used by a community of classics scholars in order to understand best practices for data interoperability in the humanities. The authors begin by defining three essential virtues of "virtualized" resource access: location-free technology, autonomy of data management regimes, and heterogeneity of both the storage mechanisms and the data. They claim, intriguingly, that virtualization can hide "irrelevant" differences between data resources (in other words, making different formats functionally equivalent whenever it is convenient to do so), offering detailed system specifications from the Linking and Querying Ancient Texts (LaQuAT) project as positive evidence.

Groth, Paul, Andrew Gibson, and Johannes Velterop. (2010). The anatomy of a nanopublication. *Information Services and Use, 30*(1-2), 51–56.

In this article, the authors posit a concept model for what they call "nanopublications"; that is, semantically-enabled, one-off data snippets that they believe will help to drive down the lowest common denominator of scholarly publishing. Refereed journal articles can easily take at least a year to bring to publication, and it is only after this that they can be authoritatively referenced by potential collaborators. While the humanities have historically been less incentivized than other disciplines to advance the speed at which the gears of the academic knowledge economy are turned, they are certainly no less dependent on certain norms for attribution and, particularly in the design of reading environments for the digital humanities, annotation. Both are made more dynamic by these proposed advancements in sentence-level document structure.

Content management systems and open repositories

Articles in this section focus on implementations of the primary digital repository solutions in use over the last decade, namely Greenstone and Fedora, as well as middleware used to bridge systems. Other content management systems are discussed in articles herein as well, both as reference to the state of the field in past years, as well as for some indication of the types of systems under consideration for data storage in the future.

Witten, Ian H., David Bainbridge, and Stefan J. Boddie. (2001). Greenstone: Open-Source Digital Library Software. *D-Lib Magazine,* 7(7), n.p.

This early article, relative to the creation of digital libraries and to the Greenstone software, describes the basic functionality of the Greenstone digital library software as a content management system. Although the software has undergone a great deal of development in the ensuing decade, this overview describes the basic features of the software which are still in place: the ability to construct and present collections of information, the ability to search both full text and metadata, and the ability to browse by metadata elements. Additionally, even this early iteration of Greenstone had the ability for developers to create and install plugins—in this case to accommodate different document and metadata types. The bulk of the article is designed to introduce library professionals to the primary interface for the Greenstone system, the "Collector," so as to demonstrate the ease of use for creating and managing collections, including adding material to collections and distributing these structures both as self-contained installable libraries or Web-accessible libraries.

Witten, Ian H. (2003). Examples of Practical Digital Libraries: Collections Built Internationally Using Greenstone. *D-Lib Magazine,* 9(3), n.p.

This follow-up article to the aforementioned introduction to Greenstone highlights some of the myriad ways organizations used the software to build digital libraries in the first few years of the package's general release. The examples shown highlight the use of Greenstone in many countries (and thus housing and displaying content in many languages), in several different contexts (historical, educational, cultural, and research), and to store different types of source material (text, images, and audio). Greenstone is geared toward maintaining and publishing collections—which necessarily includes an interface—and not for the pure data storage, disassociated from an interface, that may be used by large, distributed research groups.

Witten, Ian H. and David Bainbridge. (2007). A Retrospective Look at Greenstone: Lessons From the First Decade. Proceedings from*: 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 147-156). New York: ACM.

This retrospective of ten years of Greenstone development and production helped researchers to better understand the original (and continuing) purpose of Greenstone; specifically, that its goal is to enable a relatively easy method for constructing and publishing a digital library. As such, the program meets its goal—as evidenced by the hundreds of organizations using it both at the time of this retrospective and today—but it does not necessarily meet current researcher needs. For instance, that Greenstone is bundled with an interface (two, actually, one for the Reader and one for the Librarian), and that those interfaces are the only methods through which data can be accessed or added, necessarily limits its usefulness for a research endeavour in which the data and the interface must remain separate. Interestingly, by the time of this retrospective, Greenstone developers were already ensuring data inter-operability between other digital repository software, such as DSpace and Fedora.

Smith, MacKenzie, Mary Barton, Mick Bass, Margret Branschofsky, Greg McClellan, Dave Stuve, Robert Tansley, and Julie Harford Walker. (2003). DSpace: An Open Source Dynamic Digital Repository. *D-Lib Magazine, 9*(1).

A few years after Greenstone was released and gained traction within institutions, MIT Libraries and Hewlett-Packard Labs began collaborating on the development of an open source digital repository called DSpace. This article provides an overview of DSpace for library professionals; it first describes the impetus behind development (to manage institutional research materials and publications in a stable repository, specifically for MIT but with the hopes of wider adoption), and then describes the unique information model. The authors then describe elements of DSpace with regard to its metadata standard, user interface, workflow, system architecture, inter-operability, and persistent identifiers. Each of these elements is explained briefly, with enough information provided to give the reader a clear sense of the usefulness and maturity of the software at this point in time, without overwhelming them. The remainder of the article is a description of the MIT Libraries' DSpace implementation, such that potential users could gain an understanding of the policies and procedures in place for such a process.

Witten, Ian H., David Bainbridge, Robert Tansley, Chi-Yu Huang, and Katherine J. Don. (2005). StoneD: A Bridge Between Greenstone and DSpace. *D-Lib Magazine, 11*(9).

After Greenstone and DSpace each gained a strong user-base, developers for both projects collaborated on a software bridge used to migrate between Greenstone and DSpace. This article delineates the similarities and differences in these two digital library systems. Today, the article may be of greater importance, as it articulates the different goals and strengths of each system, and identifies situations in which each system would be better utilized. For example, DSpace "is explicitly oriented towards long-term preservation, while Greenstone is not"; DSpace is designed for institutions, while Greenstone is designed for anyone with basic computer literacy to run inside or outside an institutional environment, and so on.

Staples, Thornton, Ross Wayland, and Sandra Payette. (2003). The Fedora Project: An Open-source Digital Object Repository Management System. *D-Lib Magazine, 9*(4).

This early article provides an overview of the Fedora (Flexible Extensible Digital Object and Repository Architecture) project. The Fedora architecture is based on object models, on which data objects are in turn based. The software internals are configured to deliver the content in these objects based on the models the objects follow, via Web services. This article outlines this architecture in a basic way, providing an understanding of the fundamental differences between Fedora and systems like Greenstone and DSpace, namely, that the former is based on multiple layers (Web services, core subsystem, and storage) and public APIs (application programming interfaces, in this case for management and access). After a description of these layers, the article notes the features already present in this early version of Fedora, such as XML submission and storage, parameterized disseminators, methods for access control and authentication, and OAI metadata harvesting. The remainder of the article describes four cases for the use of Fedora: "out of the box" management and access of simple content objects; as a digital asset management system; as a digital library for a research university; and for distributed content objects.

Lagoze, Carl, Sandra Payette, Edwin Shin, and Chris Wilper. (2005). Fedora: An Architecture for Complex Objects and their Relationships. *International Journal on Digital Libraries, 6*, 124-138.

This article describes in rich detail the Fedora architecture (based on version 2), namely, the structures and relationships that provide the framework for the storage, management, and dissemination of digital objects within the repository. Additionally, the authors describe their "motivation for integrating content management and the semantic web" (p. 125) as a need driven by the Fedora user community at the time; thus, semantic relationships between objects, and the need to represent, manipulate, and query these relationships and objects, became the developmental focus. The bulk of this article focuses on detailed descriptions of the Fedora digital object model as well as the Fedora relationship model. Both sets of descriptions are important to understand the basic principles of this software. The authors also take a moment to discuss the ways in which Fedora (as of version 2) had been implemented for real-world collections, and also the ways in which Fedora differs from institutional repositories such as DSpace, arXiv, and ePrints: Fedora was designed from the beginning for extensibility, modularity, and as a pluggable service framework.

Han, Yan. (2004). Digital Content Management: The Search for a Content Management System. *Library Hi Tech, 22*(4), 355-365.

This article outlines the systems analysis process undertaken by the University of Arizona Library for the selection of a digital content management system. The key elements of this article include the predetermined criteria against which candidates were judged, as well as the eventual performance of each system in the evaluation process. The bulk of the article is devoted to detailed analyses of Greenstone, Fedora, and DSpace with respect to the predetermined criteria for digital content management: preservation, metadata, access, and system features based on the needs of the University of Arizona Library. While the core of the article describes the criteria and considerations that should have been determined by the group as a whole during Year One of the project, the appendices provide documentation of both the University of Arizona criteria and results of the analysis. Although the University of Arizona selected DSpace for their content management system, the reasons DSpace won out over Fedora do not point to any inherent failings of Fedora as a content management system.

Allinson, Julie, Sebastien François, and Stuart Lewis. (2008). SWORD: Simple Web-service offering repository deposit. *Ariadne, 54*.

This article documents the work of the JISC-funded SWORD (Simple Web service Offering Repository Deposit) project from 2007. The impetus behind SWORD was to create a lightweight deposit API that would be inter-operable with open repositories such as DSpace, Fedora, and ePrints. The authors list a now commonly stated goal of data repositories that informed the development of SWORD: to support a wide range of large-scale, heterogeneous data formats with linked metadata. They justify their choice of the lightweight ATOM protocol for publishing Web resources, particularly with respect to designing the repository to programmatically "explain" its policies and procedures as part of the deposit process, and briefly detail the functionality of the available SWORD repository clients. Note that SWORD functionality is also built into the Microsoft Word Article Authoring Add-In, available at http://research.microsoft.com/en-us/downloads/3ebd6c86-95b0-4dc3-950e-4268508f492e/default.aspx.

Aschenbrenner, Andreas, Tobias Blanke, David Flanders, Mark Hedges, and Ben O'Steen. (2008). The Future of Repositories? Patterns for (Cross-)Repository Architectures. ***D-Lib Magazine, 14*****(11/12).**

In this relatively recent article, the authors examine the growth of repositories in the previous decade, and especially the evolution of what are, by the time of publication, the major players in the field: DSpace and Fedora. However, the goal of the article is to investigate how repository architectures could or should change as the needs of libraries and end users change. The authors examine whether every institution even needs a local repository; this speaks to collaborative and administrative work more than technical requirements, although technological know-how (and potential lack thereof both internally and externally) also underlies this question. Of particular relevance is the authors' discussion of the future, and of the desire for an open repository environment in which "repository components can be mixed, and external services can be employed to fit an institution's capabilities and needs."

Brase, Jan. (2009). DataCite–A global registration agency for research data. Proceedings from: ***Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology*****. Beijing, China.**

This article highlights the little-known fact that since 2005, the German National Library of Science and Technology (TIB) has offered a Digital Object Identifier (DOI) registration service for persistent identification of research *data*, by virtually identical means to the assignment of DOIs to published articles elsewhere in the world. The DataCite initiative thus seeks to enable researchers across the globe to assign permanent, unique, and citable identifiers to their datasets. The authors note a key issue with current linking of data sets: while Web search engines, including Google Scholar, do a reasonably good job of encouraging "findability", they face the common (and, in this instance, magnified) problem of poor metadata. Likewise, linking data sets on the Web from the articles in which they are mentioned solves the problem of organizing and locating resources according to current norms, but only works for datasets that correspond directly to published articles. This presents an interesting philosophical question, which has been troubling in practice if not in theory for the sciences, and has so far gone mostly unaddressed in the humanities and social sciences: do the benefits of standalone data publication outweigh the difficulties of making such a change to the academic knowledge economy?

Green, Richard and Chris Awre. (2009). Towards a Repository-enabled Scholar's Workbench: RepoMMan, REMAP and Hydra. *D-Lib Magazine,* 15(5/6).

This article describes the genesis of the Hydra project, which seeks to develop a repository-enabled "Scholars' Workbench", or, a flexible search and discovery interface for a Fedora repository. After outlining four years of research at the University of Hull, during which time the RepoMMan tool (a browser-based interface for end-user interactions within a repository) and the REMAP project (process-oriented management and preservation workflows) were developed, the authors discuss the beginning of a three-year development commitment between the University of Hull, the University of Virginia, Stanford University, and the Fedora Commons to develop a set of Web services and display templates that can be configured within a reusable application framework to meet the myriad needs of an institution.

Sefton, Peter. (2009). The Fascinator: a lightweight, modular contribution to the Fedora-commons world. Proceedings from: *Fourth International Conference on Open Repositories.* Atlanta, Georgia.

The Fascinator is, in the words of its creator, "a useful, fast, flexible web front end for a repository [Fedora] using a single fast indexing system to handle browsing via facets, full-text search, multiple 'portal' views of subsets of a large corpus, and most importantly, easy-to administer security." It also includes a client application for packaging and indexing research objects that is designed to monitor a user's desktop and automatically create local or remote backups for reuse by colleagues or other researchers. The system is designed to be extensible, so that plugins may eventually be developed to programmatically interpret different research objects (there are currently plans for a generic interpreter to read column headers out of CSV files as ad-hoc metadata), potentially diminishing the need for annotation of shared research objects.

Reilly, Sean and Robert Tupelo-Schneck. (2010). Digital Object Repository Server: A Component of the Digital Object Architecture. *D-Lib Magazine, 16*(1/2).

This article introduces the CNRI Digital Object Repository Server (DORS) and raises an interesting question for consideration; namely, to what extent do research initiatives need to disassociate themselves from design problems and simply provide the most streamlined access to digital content? The DORS is described here as a flexible, scalable, streamlined package for depositing, accessing, and long-term storage and management of digital assets. It uses persistent identifiers, object identifiers as keys, has a uniform interface to structured data, maintains metadata associated with objects, includes authentication features, and provides automatic replication; however, documentation on the project is limited, with the only available written content being this article and the source code itself. Despite this lack of information, DORS should be watched for future developments and possibilities.

Kucsma, Jason, Kevin Reiss, and Angela Sidman. (2010). Using Omeka to Build Digital Collections: The METRO Case Study. *D-Lib Magazine, 16*(3/4).

This case study is included here as an example of the ways in which managers of digital content are leveraging more lightweight repository solutions for long-term preservation and access to collections. This article outlines the ways in which the Metropolitan New York Library Council (METRO) used Omeka, a software platform for creating and managing digital collections on the Web, to build a directory of digital collections created and maintained by libraries in the metropolitan New York City area. The case-study approach addresses Omeka's strengths and weaknesses, with an emphasis on original record creation and the pluggable system architecture. Although a recent article (written mid-2010), the case study is based on Omeka version 1.0; however, the software has matured considerably since then, and continues to do so. The key factors of a pluggable system architecture and the maintenance of content within a flexible environment point toward the type of repositories and user expectations we will likely see in the future.

Viterbo, Paolo Battino, and Donald Gourley. (2010). Digital humanities and digital repositories. Proceedings from: *28th ACM International Conference on Design of Communication*. Sao Paolo, Brazil.

This article is a case study of the Digital Humanities Observatory (DHO) in implementing a digital repository. In addition to such commonly-stated requirements as the ability to deal with heterogeneous data resources, the authors note that the repository must have the ability to support projects at any individual stage of development, given that adoption may vary considerably among its user community, and that the need to support browsing of resources will necessarily supersede a desire to use a lightweight access protocol. These and other specifications – such as the decision to use the content management system Drupal rather than the more powerfully object-oriented Django largely because of the former's large open-source developer community, and an avoidance of Flash in favour of HTML5 in light of the recently released iPad – are unusually articulate and particularly helpful, as they befit true digital humanists.

Corpora: A review of select bibliographic sources

The articles in this section review scholarly work in the area of corpus building, establishment, facilitation, and (semi-)automatic generation of information resources. It is important to note that "corpus linguistics" is a bit of a misnomer that need not entail any research into linguistics per se. Large textual corpora are at least as valuable for literary study as they are linguistic analysis, and beyond that, they inspire pragmatic meta-analyses and case studies surrounding their role in digital libraries and archives. This bibliography reviews the past decade of work in automatic and manual corpus-building and text classification, and would serve as an excellent resource to any who are beginning work with large corpora. Among the articles reviewed are several instances of using large corpora in related fields such as natural language processing, network analysis, and information retrieval, as well as epistemological meditations on the creation and use of document encoding standards.

Three publications in particular should produce related content: *D-Lib Magazine*, *Digital Humanities Quarterly*, and the *International Journal of Corpus Linguistics*.

Additionally, the fast-paced nature of work in this field combined with the relatively slow publication schedule of some scholarly journals makes necessary a frequent review of presentations and reports to associated groups of the Association of Computing Machinery (ACM) such as SIGDOC (the Special Interest Group on Design Of Communication), as well as the Canadian Symposium on Text Analysis (CaSTA).

Corpus building and administration

The articles and reports in this section focus specifically on the theoretical framework underlying the eventual technical architecture of a corpus. Although not discussed in detail in this bibliographic essay, Susan Armstrong's 1994 edition, *Using Large Corpora*, brings together numerous essays concerning corpus-building and corpus linguistics, many of which are reference points to the later research outlined below.

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics, 19*(2), 313-330.

Developed initially between 1989 and 1992, the Penn Treebank was the first large-scale treebank, or parsed corpus. Although a parsed corpus necessarily has different conceptual, theoretical, and technological underpinnings than lexical corpora, the lessons learned with regard to resource use and technical architecture remain valuable. Written after initial development, this article details the decisions and actions (and reactions) made while constructing a corpus of more than 4.5 million words of American English and annotating the part-of-speech (POS) information for each word. The problems and solutions regarding collaborative work of this type are not unrelated to the problems and solutions encountered during lexical corpus development; in both situations, research methodology must be agreed upon and documented for all parties involved, and tests (spot checks) of any manual work should be planned into the project. The article also indicates the research efforts initially reliant on the output of the Penn Treebank, rightly details the limitations of the initial design, and looks ahead to future iterations of the corpus. These developmental stages and the manner of project documentation and description serve as a good model for future corpus development projects.

Davis, Boyd. (2000). Taking Advantage of Technology. Language And Digital Technology: Corpora, Contact, and Change. *American Speech, 75*(3), 301-303.

Situated at the beginning of the new millennium, this brief essay reminds researchers that the inclusion of technology in their studies of language can move the work forward in multiple ways. Beginning with the notion that the reproduction of large corpora and databases is nothing new (the scribe has created databases on paper for centuries), Davis reminds us what *is* new is the digitization of corpora and thus different entry points of contact for different researchers. In short, the author argues that while digital technology can support the study of language contact and change, it can also be "the vehicle and perhaps an impetus of change" (p. 703) as well.

Crane, Gregory and Jeffrey A. Rydberg-Cox. (2000). New Technology and New Roles: The Need for 'Corpus Editors.' Proceedings from: *The fifth ACM Conference on Digital Libraries*. New York: ACM.

In this report, the authors explain the need for a clearly defined professional role devoted to corpus maintenance, that of a "corpus editor," or one who manages a large collection of materials both thematically (as a traditional editor) and with technical expertise (at the computational level). At the time of writing, this sort of position was unheard of as "no established graduate training provides" a learning path toward gaining this expertise. Crane and Rydberg-Cox (2000) outline some of the possible tasks for the "corpus editor," which are important to note for anyone determining personnel to include in such a project, but the underlying argument is perhaps more important: the need for technical and academic expertise to be brought together in the formal instruction of humanists, if corpora (and digital libraries in general) are to fulfil their promise.

Crane, Gregory, Clifford E. Wulfman, and David A. Smith. (2001). Building a Hypertextual Digital Library in the Humanities: A Case Study on London. Proceedings from*: The first ACM/IEEE-CS Joint Conference on Digital Libraries.* New York: ACM Press.

This detailed article describes the digitization of the initial London Collection (11 million words and 10,000 images) and its inclusion in the Perseus Digital Library. As the authors clarify, a collection of this size, with information more precise than collected before, allowed the developers to "explore new problems of data structure, manipulation, and visualization" (p. 1) and in greater detail than before. The authors remind us that digital libraries should be designed for users to systematically expand their knowledge; that good design of collections is crucial for broad acceptance; that the size of collections matters (the bigger the collection, the more useful it is); that users should be able to work with objects at a fine level of granularity; and finally that study objects should contain persistent links to each other.

Crane, Gregory and Clifford Wulfman. (2003). Towards a Cultural Heritage Digital Library. Proceedings from*: The 3rd ACM/IEEE-CS Joint Conference on Digital Libraries.* Washington, DC: IEEE Computer Society.

As the Perseus Project continues to grow and establish itself as a model cultural heritage collection, papers such as this continue to appear, documenting research and technological factors leading to its success. In this paper, the authors articulate issues encountered during the creation and maintenance of the collection, specifically in the realm of audiences, collections, and services. The authors begin by reminding readers of the premise of the Perseus Project and its corpora: that "digital libraries promise new methods by means of which new audiences can ask new questions about new ideas they would never otherwise have been able to explore" (p. 1). The authors then delineate trade-offs that were considered in creating and maintaining the collection, including the perceived neglect of the core collection versus the need to generalize, and the issue of exploring new domains versus the rigors of disciplinarity. Additionally, the authors describe what they consider to be the basic services to include in such a collection (or technical framework for a collection): document chunking and navigation services, an XML server, visualization tools, citation linking, quotation identification and source tracking, named-entity identification, semantic services, authority-list editors, runtime automatic linking, and automatic evaluation services.

Sinclair, John. (2005). Corpus and Text – Basic Principles. In Martin Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 1-16). Oxford: Oxbow.

Burnard, Lou. (2005). Metadata for Corpus Work. In Martin Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 30-46). Oxford: Oxbow.

Wynne, Martin. (2005). Archiving, Distribution and Preservation. In Martin Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 71-78). Oxford: Oxbow.

These three chapters from the 2005 text *Developing Linguistic Corpora: A Guide to Good Practice* are especially useful for newcomers to corpus linguistics. Although these texts are geared toward linguistic rather than lexical corpora, the underlying quality of the corpus, as a collection of texts, remains the same; as such, in this volume the editor has brought together numerous experts in the field to describe aspects of corpus building in relatively non-technical terms. One of the stated goals of the edited collection is specifically to be a starting point for scholars and researchers new to the field, and even five years later this text fulfils this mission. Sinclair's (2005) essay specifically addresses the basic principles of corpus development: who builds it, what is it for, and how can it be approached. Of particular interest are Sinclair's admonitions of what corpora are not; he contends the following are not corpora: the Web, an archive, a collection of citations, a collection of quotations, a concordance, and a single text. This assertion will be directly challenged in later essays as work with corpora spreads into fields other than those purely linguistic in nature. Although the delimitations of corpora may be contentious, Burnard's (2005) entry in this collection discusses the scope of metadata used in linguistic corpora, and acknowledges these types of metadata are useful for other purposes as well. Specifically, Burnard (2005) identifies the following types of metadata: editorial, analytic, descriptive, and administrative. All of these working together, he contends, are part of the interpretative framework within which the corpus operates. Wynne's (2005) essay, on the other hand, addresses the question of what happens after the creation of the corpus. Wynne (2005) offers observations of the scope of ongoing development (if any), the rights and responsibility of the developer, how and where the corpus is stored and who has access to it, how users will find it, and what archival format the text(s) should assume for reasons of sustainability.

Thelwall, Mike. (2005). Creating and Using Web Corpora. *International Journal of Corpus Linguistics, 10*(4), 517-541.

In Sinclair's (2005) *Corpus and Text – Basic Principles*, he contends that the Web should not be considered a corpus. On the other hand, in this essay (published in the same year), Mike Thelwall (2005) discusses both the opportunities and problems of using commercial search engines and the Web itself as a corpus, as well as ways to mitigate those problems through alternative data collection strategies. First, Thelwall (2005) describes how the Web does indeed meet the requirements for classification as a corpus insofar as it is a body of text that can be used for research. In his estimation, (as of 2003, which should be noted is before the boom of Web 2.0/user-generated content) "Web English is not representative of written or spoken English" (p. 519). He also notes that it is difficult to determine the importance of Web content, which is "normally implicit" in corpus construction; that much Web content is in fact replicated or nearly-replicated; and that authorship of Web content is often indeterminate. Thelwall's (2005) solution for working through these issues with the Web as corpus is to create a personal Web crawler "for more direct control over data collection" (p. 526), and he dedicates the remainder of the essay to a discussion of the use of a Web crawler for building Web corpora.

Sharoff, Serge. (2006). Open-Source Corpora: Using the Net to Fish for Linguistic Data. *International Journal of Corpus Linguistics, 11*(4), 435-462.

Like the Thelwall (2005) essay published a year prior, this paper outlines a methodology specifically for collecting what the author terms an "open-source" corpora, or corpora that are collected from the Web and distributed in such a way as to reproduce "a random snapshot of Internet pages which contain stretches of connected text in a given language" (p. 435). Although this, again, differs from Sinclair's (2005) view that the Web itself is not a corpus, Sharoff (2006) carefully details the series of actions necessary to create Internet corpora, including pre-planning the word selection, generating queries, downloading from the Internet, post-processing, composition assessment, and comparison of word lists (with the last two steps being optional depending on the researcher's intentions).

Rydberg-Cox, Jeffrey A. (2006). *Digital Libraries and the Challenges of Digital Humanities.* Oxford: Chandos Press.

This short text is very useful for any humanities scholar looking to collaborate with librarians to create and maintain a digital repository. By focusing on systems, data structures, and collaboration more than on highly technical details, the author has produced a neat resource for the creation and administration of digital objects. Sections include information on tagging and text encoding, parsing corpora, information retrieval, and new modes of scholarship enabled by these methods.

Baker, Paul. (2006). *Using Corpora in Discourse Analysis.* London: Continuum.

Paul Baker's 2006 book provides an excellent introduction to the state of the art in corpus linguistics, particularly as applied to social science and humanities methodologies such as discourse analysis. Much of the book is devoted to simple, powerful natural language processing techniques for analyzing text, such as the creation of dispersion plots, concordance frequencies, collocation frequencies, and word type/token ratios. In fact, Baker (2006) devotes remarkably little of the text to the theoretical underpinnings of discourse analysis, and at times the book more closely resembles an introductory computational linguistics textbook. Given that there is a relative dearth of such texts geared towards a humanities audience, this is, nevertheless, very much welcome.

Waugh, Andrew. (2007). The Design and Implementation of an Ingest Function to a Digital Archive. *D-Lib Magazine, 13*(11/12).

This article provides a detailed look at the design of an ingest function for a particular digital archive (in this case, the Public Record Office in Victoria, Australia) and the lessons learned, both managerial and technical, during this process. The core functionality of the ingest function includes: validation of digital objects, bulk object handling, support for object grouping, support for minor corrections to objects, and acceptance of responsibility for objects in the archive. Lessons learned included the need to re-evaluate the manual processing area (a sort of limbo between the original location of the object and its eventual space in the archive), and a desire to reduce administrative interaction while not overcompensating in automation.

Klavans, Judith, Tandeep Sidhu, Carolyn Sheffield, Dagobert Soergel, Jimmy Lin, Eileen Abels, Rebecca Passoneau. (2008). Computational Linguistics for Metadata Building (CLiMB) Text Mining for the Automatic Extraction of Subject Terms for Image Metadata. Proceedings from: *VISAPP Workshop Metadata Mining for Image Understanding*. Madeira, Portugal.

The authors report on the development of the Computational Linguistics for Metadata Building (CLiMB) project for automatically creating, associating, and indexing image metadata, addressing long-standing challenges in non-textual document retrieval. First, they begin by creating indexing terms for the images from a parallel text corpus. The computational workflow used by their tools include part-of-speech tagging of text, identifying maximally relevant phrases, associating terms using the WordNet thesaurus, and the further disambiguating of these terms. The output of this process is presented to a human cataloguer alongside potential matching images, and the system "learns" about the efficacy of its suggested indexing terms based on the accumulated results. The article concludes with a brief synopsis of the various libraries in which the software is currently being tested.

Mehler, Alexander. (2008). Large text networks as an object of corpus linguistic studies. In Anke Ludeling and Kyto Marja (Eds.), *Corpus Linguistics: An International Handbook* (pp. 328-382). Berlin: de Gruyter.

This book chapter details several potential applications of network analysis techniques to corpus linguistics. The author begins with relatively high-level graph theory, as would be useful to any novel software development efforts in corpora-driven network analysis. He follows this with an extraordinarily in-depth review of natural language processing techniques at the level of both syntax and semantics, as well as applications of graph theory to bibliometric mapping. Somewhat unusually for corpus linguistics, he provides a comprehensive overview of sources for mining network data (e.g., networked blogs, mailing lists), with relatively little attention given to traditional linguistic corpora. He concludes with an open-ended discussion of some unsolved problems in network analysis, in the hopes that they may be solved using mixed-method approaches from text analysis.

Flowerdew, Lynne. (2009). Applying corpus linguistics to pedagogy: A critical evaluation. *International Journal of Corpus Linguistics, 14*(3), 393-417.

The author of this paper seeks to address criticisms of using corpus linguistics in pedagogy, particularly in teaching English for Special Purposes (ESP). Among the primary reasons for this criticism, according to the article's author, is the tendency of corpus data to present language out of context, strongly encouraging an "atomistic" bottom-up approach. She reports on a case study addressing this problem from law education, where students were "inducted" into the structure of legal essays by reading through entire essays before being introduced to the corpora materials. She also notes that the "keywords in context" view returned by most lookup engines (displaying a fixed amount of the text immediately surrounding the search string) does in fact enable top-down thinking in consideration of the search results. With regard to ESP, it is true that a utilitarian, narrowly focused teaching philosophy might be further enabled by corpus linguistics, but the responsibility still falls squarely on the shoulders of the teacher, and any such concerns ignore individual differences in learning styles and cognition.

Gleim, Rüdiger, Ulli Waltinger, Alexandra Ernst, Alexnder Mehler, Tobias Feith, Dietmar Esch. (2009). eHumanities desktop: an online system for corpus management and analysis in support of computing in the humanities. Proceedings from: *The 12th Conference of the European Chapter of the Association for Computational Linguistics.* Athens, Greece.

This short paper presents a little-known toolkit that is both surprisingly complete and incredibly messy, taking an "everything but the kitchen sink" approach to the digital humanities that is nonetheless deserving of greater exposure. The *eHumanities Desktop* employs a client-server architecture, enabling it to run in a browser, and includes an easy-to-use graphical interface for part-of-speech tagging, lexical chaining, and text classification. As a document manager, it is designed to be extensible into repositories, and allows the user to create taxonomic links between documents, as well as set complex group permissions for collaborative reading and writing.

Honkapohja, Alpo, Samuli Kailaniemi, and Ville Marttila. (2009). Digital Editions for Corpus Linguistics: Representing manuscript reality in electronic corpora. Proceedings from: *The 29th International Conference on English Language Research on Computerized Corpora*. Oslo, Norway.

This paper introduces the Digital Editions for Corpus Linguistics (DECL) project, which "aims to create a framework for producing online editions of historical manuscripts suited for both corpus linguistic and historical research." The authors begin by discussing the theoretical underpinnings of their XML model, using a threefold division of *artifact, text, and context*. The first two correspond roughly to "expression" and "item" in the *work, expression, manifestation, item* or "WEMI" hierarchy imposed by the Functional Requirements for Bibliographic Records (FRBR) cataloguing standard. The third, that of *context*, encodes the "historical and linguistic circumstances" surrounding the artifact. It is important that access to these documents be as flexible as possible, while first and foremost still accurately preserving;

thus, editorial markup should be allowed, but must be able to be rolled back to a pristine original copy at any time. The authors speed through issues of problematic orthography and copyright, and conclude with a focus on the architecture of their system.

Rayson, Paul, and John Mariani. (2009). Visualising corpus linguistics. Proceedings from: *The Corpus Linguistics Conference.* Liverpool, UK.

The authors begin with a pitch of sorts: corpus linguistics is certainly not unlike related disciplines in having increasingly great mountains of data to sift through, and has in fact experienced this problem for at least as long as others, yet seems to have ignored information visualization as a solution. They then proceed to illustrate work-in-progress examples of this solution; articles such as this one are often useful insofar as they serve as effective advertisements for several smaller open-source projects, and this one is no exception. DocuBurst (2006) provides a graphical front-end to WordNet, the aptly-named Concgrams (2006) displays n-gram concordance, and Wmatrix (2008) generates a "key word cloud" for documents. The authors conclude with a prototype of their own work, a "Dynamic Tag Cloud" in which the size of represented words is dependent on a number of variables over which the user has full control.

Williams, Geoffrey. (2010). Many Rooms with Corpora. *International Journal of Corpus Linguistics, 15*(3), 400-440.

Much like Davies' (2010) article from the same issue of this journal, this article highlights some of the financial issues facing corpus developers, particularly with regard to large-scale collaborations. Williams (2010) reminds linguists, humanists, librarians, and technologists to know "where we have come from and why" so as not to push "in directions that undermine the very basis of the discipline" (p. 406). He calls for respect of each others' fields, toward the goal of working together to leverage lessons learned and, perhaps, grants awarded.

Document design and data classification

Between the need to provide a theoretical foundation for the technical architecture of a corpus and the iterative process of project development and testing, corpus developers and administrators must build in time to clearly design and classify the documents in the eventual data store. Articles in this section speak to this process and offer rich examples of the type and function of data classification as it pertains to corpus design.

Crane, Gregory. (2000). Designing Documents to Enhance the Performance of Digital Libraries: Time, Space, People and a Digital Library of London. *D-Lib Magazine, 6*(7/8).

Reports from the Perseus Project are both important and useful to any researchers planning to create and maintain corpora; important because of the breadth of information and lessons detailed therein, and useful because of the granularity of the information included within the texts. This essay is particularly interesting because of its publication date; namely, the attention paid to the geographical data (through Geographic Information Systems, or GIS) a full decade before major development work with GIS and the digital archive. In this essay, the authors discuss the development of a temporal-spatial front-end for digital libraries, so that users can better understand the full nature of the historical documents contained within. The descriptions of automated tagging of text and the aggregation of tagged data speak to the need for careful planning and construction of the documents within the corpora, as well as for the need to train scholars in the important tasks involved in corpus editing.

Luyckx, Kim, Walter Daelemans, and Edward Vanhoutte. (2006). Stylogenetics: Clustering-based stylistic analysis of literary corpora. Presented at: *The 5th International Language Resources and Evaluation Conference*. Genoa, Italy.

This is among the foundational papers in what has since become a popular subfield in its own: "stylogenetics." The term refers to the algorithmic clustering of similar *styles of authorship*, providing a quantitative measure of how much more similar, for example, David Foster Wallace's style of writing is to that of James Joyce than your own. This, as the authors explain, is accomplished through largely the same means as *topic* clustering – a familiar tool for bibliometric analysis – by iteratively sampling multi-word phrases (i.e., by learning which words are likely to follow after one another in a given style of writing), then plotting the nearness of these samples in a vector space, as is done in information retrieval. The authors isolate four features which they were able to use as "style markers" in this respect: token-level features such as word length; syntactic features such as parts-of-speech, usually entailing some degree of manual annotation in combination with an existing part-of-speech corpus; features based on vocabulary richness, such as type-token ratio; and, common word frequencies, such as function words. Their method eventually draws a hierarchical "family tree" to represent which authors are most similar to others, a helpful visualization trick at such an early stage of this research.

Ebeling, Signe O. and Alois Heuboeck. (2007). Encoding Document Information in a Corpus of Student Writing: The British Academic Written English Corpus. *Corpora,* 2(2), 241-256.

This essay details the transformation of user-submitted Microsoft Word documents (a proprietary file format) into XML-encoded corpus files. During the planning stages of this project, the researchers decided to apply TEI encoding standards to all documents in the corpus. As the authors describe, this decision had the following implications: files must be encoded in plain text; the XML will keep separate the text and metadata; and, the TEI standards will provide the basis for the subset selected for use. The bulk of the essay and its appendices outline the evaluation, conceptualization, and decision-making processes, and finally the process of transforming into XML the raw data submitted for inclusion in the corpus.

Kim, Yunhyong, and Seamus Ross. (2007). 'The Naming of Cats': Automated Genre Classification. *International Journal of Digital Curation,* 2(1), 49–61.

Kim and Ross' (2007) article is the lone paper in this review to use the word *genre*. In the years since, the term has risen sharply in popularity among researchers in digital libraries and information behaviour as a means of describing the miscellaneous attributes of a document that aid our understanding of it. This paper was selected for being (somewhat counter-intuitively) among very few to provide a lucid case study of how document genre can and cannot be deduced in certain circumstances, unlike many other genre studies which focus on eventual use cases for as-yet undefined characteristics. This aside, however, Kim and Ross' (2007) article is still, in a sense, about the *use* of genre artifacts to guide in the automated extraction of metadata from objects ingested into digital repositories. The enthusiastic reader is encouraged to look elsewhere for an in-depth description of metadata specificity, but their essential point is not a new one: just as there are certain "core" attributes which are common to nearly all catalogued documents (e.g. title, author), we can tailor individual attributes, or sets of attributes, to unique documents, or sets of documents (i.e. document genres). Here, Kim and Ross (2007) report on their efforts to use both stylistic and aesthetic features of documents to train an automatic genre classifier.

Ide, Nancy. (2008). Preparation and Analysis of Linguistic Corpora. In Susan Schreibman and Ray Siemens (Eds.), *A Companion to Digital Literary Studies* (pp. 289–305). Oxford: Blackwell.

In this chapter, Ide (2008) first gives a brief history of the creation of electronic corpora since the availability of computers, before segueing into best practices for the preparation of linguistic corpora. Specifically, the author describes in detail the all-important initial phase of corpus creation, data capture, and the multiple methods of doing so (manual entry, OCR, etc.). She then clearly states the reasoning behind the standard representation format for linguistic corpora (in XML) and the types of information encoded into the XML about the now-digital object. Once basic information has been encoded, the linguistic information is annotated; this includes morpho-syntactic, parallel alignment, syntactic, semantic, discourse-level, topic identification, co-reference, discourse structure, and speech and spoken data annotations. The article also includes discussion of common corpus annotation tools at the time of writing, notes on the future of corpus annotation specifically with regard to the Semantic Web, and types of analysis for linguistic corpora such as natural language processing, language learning, dictionary creation.

Sperberg-McQueen, C. M. (2008). Classification and its Structures. In Susan Schreibman and Ray Siemens (Eds.), *A Companion to Digital Literary Studies* (pp. 161–176). Oxford: Blackwell.

This incredibly rich chapter on classification structures and schemes, as well as the theoretical and practical questions of developing classification systems for objects, is indispensible for the novice researcher. Sperberg-McQueen (2008) begins by clearly defining classification and its purposes: first, to bring like objects together, and second, to distinguish between objects in a way that is relevant to their use. The author then outlines the ways in which classification schemes are used by various humanities practitioners, focusing the theoretical work that follows on the scheme best known in the humanities: libraries and bibliographies for classifying books and articles by subject. The attributes of a model classification scheme are defined, and the author reminds the reader of a crucial mantra: "perfect classification would require, and a perfect classification scheme would exhibit, perfect knowledge of the object" (p. 162).

Oard, Douglas. (2008). A Whirlwind Tour of Automated Language Processing for the Humanities and Social Sciences. Presented at: *Working Together or Apart: Promoting the Next Generation of Digital Scholarship, Council on Library and Information Resources*. Washington, DC, USA.

The author begins his paper with the truism commonly heard in the digital humanities and other interdisciplinary fields that the technological academic future has failed to become the present because those of us who need the future do not understand the technology, and those of us who understand the technology do not need the same future. While this is in many respects a rather plaintive (if not presumptuous) remark, the author soon justifies his "us and them" position with some fairly innovative concepts. Among them, the idea that optical character recognition (i.e., scanning and digitizing printed text) should not try to remove the "noise" from less-intelligible handwriting styles, but quantify it for archival study, as well as the (slightly inaccurate) idea that summarization research will be unremarkable so long as the state of the art involves algorithmically deriving the most relevant text from a passage and presenting it unaltered. Ultimately, this article's greatest value lies in its author's relatively successful attempts to describe where certain practice communities can be found (literally "conference X has many folks working on Y, but fewer Z than you might expect) —this is both rare and probably useful to novices in the field—and in his strongly-worded prediction that the dominance of statistics in fields such as machine translation will soon herald a re-education for some (digital) humanists.

Yu, Bei. (2008). An evaluation of text classification methods for literary study. *Literary and Linguistic Computing, 23*(2), 327-343.

This article, while a relatively simple application of natural language processing techniques to questions for literary study, is nevertheless interesting for precisely that simplicity, and a fine implementation of corpus linguistics to a straightforward problem. The author seeks to find which algorithm, a Naïve Bayes classifier or Support Vector Machines, is more efficient for the two tasks of measuring eroticism in Emily Dickinson and sentimentality in early American novels. Technical summaries of either method can be found elsewhere; for most purposes, it should be sufficient to keep in mind that both are popular machine learning techniques. Naïve Bayes is less computationally intensive and more modular (often applied to the task of detecting email spam), whereas Support Vector Machines are less prone to over-applying learned rules. While both algorithms performed sufficiently differently so that there was no clear winner, the author does offer some important lessons learned for literary text classification, such as the difficulty of removing stop-words (i.e., non-meaningful tokens such as "the" or "and") from verse.

Gow, Jeremy, George Buchanan, Ann Blandford, Claire Warwick, and Jon Rimmer. (Forthcoming). *User-Centred Requirements for Document Structure in the Humanities.*

This article appears to have sat unpublished for a couple of years despite its presence on its authors' respective preprint archives, not to mention its high quality. The authors present a medium-technical review of important factors in document display and design, based on user studies with humanities scholars. As an exemplar, they note the traditional table of contents, theoretically more powerful in a hypertext environment, but nevertheless often overlooked. Inspired by this, and with special attention to document retrieval, they expound on the idea of an XML hierarchy that subdivides all documents into "high, medium, and low" content tiers, suggesting that these could represent chapters, paragraphs, and sentences or individual words, respectively. This, they claim, would make the creation of tables of contents, article abstracts, and indices a virtually automatic process, all but ignoring the ample effort involved in marking up such documents in the first place. While this abstract mode of categorization seems somewhat unrealistic, it is, if nothing else, made rather intuitive by an example set of rules that the authors lay out for indexing these content tiers in digital libraries; for example, "search[ing] over high- and medium-content nodes, and browsing over high-content ones should be supported" (n.p.).

Scifleet, Paul, and Susan P. Williams. (2009). Practice theory & the foundations of digital document encoding. Proceedings from: *The 27th Annual ACM Conference on Design of Communication*. Bloomington, Indiana, USA.

This article from the SIGDOC conference dissects the technology of digital document design as it relates to the real-world needs, intention, and practice of document encoders. The focus here is not on the study of document *genre,* which has its trappings in communication, but on the challenges faced by the two-decade old Text Encoding Initiative (TEI) community. As the authors note, there are ontological questions about what, in a given document, is actually represented as *text*; certain document viewers ignore certain tags. When analyzed in terms of documentary practice, this question of representation is one of normative versus empirical practice: are document encoding structures self-describing, and how closely are these structures related to the context in which they are encoded and later used? In order to explore these questions, the authors outline plans for a survey of cataloguers and other information professionals working with the new Functional Requirements for Bibliographic Records model for library classification.

Examples and lessons learned in corpus-building projects

These and other sources produce a knowledge base of lessons learned in corpus-building projects of any kind. Elements of the design, management, and management processes can of course be independent of the content contained in the eventual corpora.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. (1998). The Berkeley FrameNet Project. Proceedings from*: The 17th International Conference on Computational Linguistics.*

This article outlines the key features and early technological implementations of FrameNet, an NSF-supported corpus-based lexicographic project. This particular paper is important as it provides an example of the technologies implemented in the early days of the Web and before contemporary scripting languages were created. Specifically, the FrameNet project contained software modules with user interface components written in Perl and accessed via CGI, and the data structures themselves were implemented in SGML. The limitations of the interface and the data models are acknowledged by the authors who note the intention to move to an XML data model and to provide a more flexible interface and processing suite of tools. All of this information is important to prevent future researchers from repeating past mistakes, and ensuring they evaluate the possibilities and availability of more modern languages and tools.

van Deemter, Kees, Ielka van der Sluis, and Albert Gatt. (1998). Building a Semantically Transparent Corpus for the Generation of Referring Expressions. Proceedings from: *The Fourth International Natural Language Generation Conference*. Morristown, NJ: Association for Computational Linguistics.

This report discusses the need for creating a semantically transparent corpus, specifically for the purpose of evaluating algorithms that generate referring expressions, related to using corpora for experiments related to natural language generation. Natural language generation systems are used to understand how concepts are put into words, a type of linguistic analysis that is not too far afield from the types of analytical processes performed on existing texts. To that end, the idea elaborated here, that "the necessary contextual information and the [knowledge of] conditions under which the texts in a corpus were produced" (p. 132) are crucial elements of corpus creation and maintenance.

Ebeling, Jarle. (2007). The Electronic Text Corpus of Sumerian Literature. *Corpora, 2*(1), 111–120.

Although this is a rather narrowly focused case study, researchers would be wise to review the lessons learned about the methods for compiling, lemmatizing, and making available digitized objects that exist originally in cuneiform and on clay, as there are some carry-over considerations for the classification, storage, and retrieval of born digital objects. In this paper, the authors explain the nuanced methods for marking up the cuneiform in TEI P4, paying close attention to the need to annotate transliterations but not translations, all the while marking up both in an effort to make clear the relationships between the two. In doing so, the researchers found it necessary to introduce new attributes to existing mark-up tags. The management and evaluation process of new attribute creation is also described and provides a useful model for other researchers.

Davies, Mark. (2009). The 385+ Million Word Corpus of Contemporary American English (1990-2008+): Design, Architecture, and Linguistic Insights. *International Journal of Corpus Linguistics, 14*(2), 159–190.

Much like technical and theoretical insights from the Perseus Project, any published reports regarding the creation and ongoing administration of the Corpus of Contemporary American English (now into its twentieth year) will prove useful for anyone working with information management and specifically linguistic corpora. This report in particular is interesting and relevant for information management work as a result of its detailed discussion of relational database architecture and acquisition of corpora from Web sources. The premise of the argument for the use of an underlying relational database architecture for this corpus is that the architecture "allows for a wide range of queries that are not available (or are quite difficult) with other architectures and interfaces" (p. 169). Specifically, Davies (2009) notes that the relational database architecture is an updated version of similar relational architectures in use for at least the previous seven years. The main table in the database contains 385+ million rows (one for each word in the corpus), in which the position, type, and parent text of each word is indicated in separate fields. That the interface to the corpus does allow for a wide range of searches and functionality with relative ease, while not taking advantage of XML data structures or Solr-based indices, deserves further study and evaluation.

Stewart, Gordon, Gregory Crane, and Alison Babeu. (2007). A New Generation of Textual Corpora. Proceedings from: *The 7th ACM/IEEE Joint Conference on Digital Libraries.* Vancouver, Canada.

The authors report on the first successful endeavour to undertake Optical Character Recognition (OCR) on Classical Greek texts, which previously required labour-intensive manual input. Their enthusiasm owes in part to the revelation that OCR systems can easily output many competing translations, the effective equivalent of hiring multiple translators. They believe that this new advance is comparable to the advent of unlimited file storage in terms of its eventual benefit for the digital humanities, and they reinforce this statement with a brief history of text digitization. Finally, they briefly explain the advantages of using large corpus linguistics to cross-reference two or more texts with OCR.

Rydberg-Cox, Jeffrey A. (2009). Digitizing Latin Incunabula: Challenges, Methods, and Possibilities. *Digital Humanities Quarterly, 3*(1), n.p.

Much like the discussion of digitizing and storing textual representations of Sumerian cuneiform, many researchers can gain from the lessons learned working with the digitization and storage of non-standard typographical glyphs, such as those produced by medieval handwriting. The author clearly outlines the requirements-gathering process for this project, which included the evaluation of multiple approaches to digitizing these documents with non-standard typography, as well as the development of a data entry methodology. Determining the data entry methodology required the creation of a custom catalog of entry keystrokes, as the characters and glyphs found in the original texts do not exist in any current encoding systems. The authors conclude, via a detailed explanation of the various resources entailed in such a project, that the greatest dollar expense would be in human resources (multiple levels of editors, entry operators, etc.) rather than technological resources.

Davies, Mark. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing, 25*(4).

This article introduces the Corpus of Contemporary American English as a new high standard for what are called *monitor* corpora, which are designed to be constantly updated to reflect changes in the language. The corpus is claimed to consist of a near-unquantifiable 400 million words from 1990 to the present, divided evenly between spoken dialogue, academic journals, books, newspapers, and magazines. The author explains that the relatively small size of stalwarts like the Brown corpus and the British National corpus make them uniquely unsuited to tracking language change, and so it follows that enormity would be a *monitor* corpus' greatest asset. Most of the article is devoted to detailing the exact provenance of some corpus resources, and expounding on its utility for study of all branches of linguistics (phonology, morphology, etc.).

Analytical facilitators: A review of select bibliographic sources

The scholarship reviewed in this section ranges from the exploration of historical sites via information presented on mobile phones, to the vast body of work on the ever-growing Text Encoding Initiative (TEI), to several neatly conflicting perspectives on data-driven research. The reader would do well to remind him or herself that much of the work in digital humanities to date has focused on experimentation, including developing novel systems for novel modes of analysis and implementation, and officiating complex cyber-infrastructures.

The Oxford Journal of *Literary and Linguistic Computing* can be relied upon to produce relevant material, as can the long-running *Computers and the Humanities,* now called *Language Resources and Evaluation*. Additionally, the fast-paced nature of work in this field, combined with the relatively slow publication schedule of some scholarly journals, helps to highlight the surprising relevance of presentations and reports to associated groups of the Institute of Electrical and Electronics Engineers (IEEE), such as the *Conference on Digital Ecosystems and Technologies,* as well as the annual *Digital Humanities Conference*.

Data visualization and geographical information

Although there is an enormous body of work on developing novel interfaces for work in the digital humanities, this research shifts in tone somewhat dramatically when dealing specifically with visualized information. Here, the digital humanities provide a curious counterpoint to data visualization in the sciences in that the articles summarized below are specifically concerned with the representation and interpretation of epistemology, and nearly all of them mention the humanities' distrust of non-textual information. While by no means an unknown topic in computer science, the question of trust in visualization is an important one, and the treatment it receives here decidedly unique.

Jessop, Martyn. (2007). The Inhibition of Geographical Information in Digital Humanities Scholarship. *Literary and Linguistic Computing, 23*(1), 39–50.

The author initially laments the underuse of spatial and spatiotemporal data in the humanities, claiming with a humanist's fervor that "landscape has been likened to a palimpsest as all past human activities have left their signature upon the land each partially overwriting whatever has gone before." He is quick to note, however, that spatial data need not necessarily be limited to GIS, particularly in the humanities where place names originate from gazetteers as often as from geometry. Problematically, these place names are an exemplar of the often diachronic data in the humanities, and lending structure to this data is easier said than done when scholars have spent centuries learning to live with small contradictions across time and place. It is also worth noting that the vast majority of structured geographic data originates from the natural sciences, where both the existing infrastructure and typical use cases are radically different. The author concludes by suggesting that shifting trends in digital scholarship, such as the migration to online journals, may eventually help humanities scholars, and their respective funding sources, to realize the advantages of geographical data.

Jessop, Martyn. (2008). Digital visualization as a scholarly activity. *Literary and Linguistic Computing, 23*(3), 281-293.

This article, a call for increased use of visualizations in the digital humanities, begins with a philosophical discussion of visualization itself as a medium of information transmission. According to the author, a "humanistic" visualization is not simply a unidirectional data communication channel; rather, concretized "data" itself betrays this definition of the humanities as tools of "rational enquiry in areas that are mostly too complex to yet be treated by science." He then argues semantics, highlighting that a visualization ceases to be merely an *illustration* – i.e., an *illustrative* complement to some text – when it becomes the principal medium of communication. Of course, text itself can be "visualized," both in terms of NLP metrics such as text concordance, and in the aesthetic variations of experimental fiction. There is also much to be said about representation, particularly in the old-fashioned McLuhan sense where even the arrangement of visual objects in a museum gallery is in a way "visualized." The author concludes by putting forward transparency and documentation as the necessary components of a successful turn towards visualization in the digital humanities.

Matei, Sorin Adam, Eric Wernert, and Travis Faas. (2009). Where Information Searches for You: The Visible Past Ubiquitous Knowledge Environment for Digital Humanities. Proceedings from: *The IEEE International Conference on Computational Science and Engineering*. Hong Kong, China.

The Visible Past knowledge environment, as described by the authors, is a sort of scholarly "augmented reality" system, which maps contextual, real-world data onto a digital object. They claim to have produced fully annotated models of "several dozen Asian, Pre-Columbian, African, Middle Eastern, and European models of UNESCO cultural heritage sites" (p. 1043), each accessible from a lightweight MediaWiki back-end. They compare their project to MIT's OpenCourseWare system for portable, digitally-enabled education, with the distinction that theirs is a rich, non-static, multimedia environment. Although they boast that their 3D models are of sufficient quality for projection in "virtual reality theatres", the greatest achievement of the Visible Past project is probably its potential for crowd-sourcing by end-user digital humanists carrying mobile phones to annotate their own locales.

Battino, Paolo, and Tarcisio Lancioni. (2010). Visualization and Narrativity: A Generative Semiotics Approach. Proceedings from: *The Digital Humanities Conference*. London, UK.

With this article, the authors seek to highlight the potentially many underestimated differences in visualizing textual *semantics*, via some mark-up embedded in an XML text, versus textual *syntax.* Although they devote surprisingly little attention to the great labour of providing accurate semantic mark-up for any narrative text, they are very much attentive to the idea of carefully expressing *function* in narrative (à la predicate logic). They consider the difficulty of accurately "visualizing" a phrase such as "The Princess awoke when kissed by the Prince" in its many possible iterations, particularly in light of *valency grammar,* which focuses on the verb as the central object of the sentence, allowing readers to consider syntax and semantics separately from one another. From this, they use the similarly-intentioned *actantial model* (of more verb-driven predicate logic) to illustrate the difficulty in reproducing the relationship between languages' deep structures and surface structures given the available repertoire of XML analysis tools.

Nooy, Wouter de. (2010). Network analysis of story structure. Proceedings from: *The Digital Humanities Conference.* London, UK.

Although, as the author of this article notes, the use of network analysis methods to analyze connections between characters and places in fiction is not new, there are many social network analysis techniques that have *not* yet been exported to the humanities. The author uses the example of structural balance (from social psychology), which assumes that individuals favour social networks that have a relatively even degree of friendship and antagonism, which sounds as though it could as easily be a recipe for a satisfying narrative when called by a different name. He reminds us that while these networks must operate on somewhat rigidly defined attributes (of character, etc.) in order to be computationally viable, those attributes can represent abstract concepts so long as they do so systematically.

Drucker, Johanna. (2011). Humanities Approaches to Graphical Display. *Digital Humanities Quarterly, 5*(1).

Drucker (2011) makes herself an example of the humanist's distrust of images early on, with the caveat that she is specifically concerned with observation and representation of the world in scholarly practice. Because, she notes, an objectivist approach to visualization acts "if the phenomenal world were self-evident and the apprehension of it a mere mechanical task, [it is] fundamentally at odds with approaches to humanities scholarship." Data is given; *capta* is taken. Much of the rest of the article is devoted to exploring this fundamental subjectivity, with extensive visual aids. She concludes with a reference to Edward Tufte's famous Choleric water pump visualization, noting that, to a humanist, each of the data points is of infinitely more interest as a history and as a life than a statistic.

Text encoding and analytic tools

The Text Encoding Initiative (TEI) is perhaps the single most visible research project in the digital humanities, and certainly the longest running. Although only a handful of the articles in this section were written specifically to address TEI, text encoding has long been at the forefront of research concerns for digital libraries and computational linguistics alike.

Schreibman, Susan, Amit Kumar, and Jarom McDonald. (2003). The Versioning Machine. *Literary and Linguistic Computing, 24*(3), 339-346.

This eight-year old article describes the Versioning Machine, an open-source tool designed to facilitate looking at multiple version of a document alongside each other. The authors expound on their decision to build on the Text Encoding Initiative standard, as well as the REST-like architecture – software that runs directly in the browser, supporting client-side transformation – some years in advance of the widespread adoption of REST in enterprise software development. They insist, however, that XSLT – then also fairly new – is the "heart" of the software, enabling multiple version *layers* to be stored on the same document using a XML tags which the interface is then able to transform into successive HTML views.

Bradley, John. (2008). Thinking about interpretation: Pliny and scholarship in the humanities. ***Literary and Linguistic Computing, 23*****(3), 263-279.**

This article details Pliny, a new software tool for the digital humanities. The author is curiously direct about Pliny's failure in its design to support any novel methods of textual analysis; rather, "the computer is meant to sit seemingly in the corner and to be almost invisible, helping the researcher do things the way s/he always does." The need for such a tool, according to him, is largely owing to the lack of any consensus over research *methodologies* in the pure humanities, which seldom extend outside of the library walls. The author goes on to elaborate on the precise capabilities of Pliny, and it is acknowledged that the software prototype, as it currently exists, is not much more than an annotation suite, and a proprietary one at that. Though the system at least deserves praise for its support of multiple file formats (e.g., both images and text), this article is primarily valuable for its discussion of computing tools in the humanities.

Craig, Hugh, and R. Whipp. (2009). Old spellings, new methods: automated procedures for indeterminate linguistic data. ***Literary and Linguistic Computing, 25*****(1), 37-52.**

This article serves as an introduction to probabilistic natural language processing for the humanities. The central object of study is a frequency table containing the entire vocabulary of an archive of early modern English plays and poems from the period 1580-1640. With this data, the authors sought to begin solving, once and for all, the Oxford English Dictionary's centuries-old problem of distinguishing variant *spellings* from variant word *forms*. Here, they lament the fact that many modern translations of their corpora are not in the public domain, and working from open access volunteer archives such as Project Gutenberg only pollutes the spelling data further. They present the compression of this old text as a four-level process, similar to the Phonology-Morphology-Syntax-Semantics interface commonly taught in linguistics, specifically: 1) the first orthographic appearance of a word; 2) the early codified form of that word; 3) the dictionary lemma; and 4) aggregation into word classes or semantic groups. This work is concerned largely with the second level, disambiguating by frequency. In closing, the researchers echo Noam Chomsky, noting that their disambiguated corpus seems remarkably small given the many theoretical linguistic mutations available to its early modern authors.

Jannidis, Fotis. (2009). TEI in a crystal ball. ***Literary and Linguistic Computing, 24*****(3), 253–265.**

This article provides an assessment of the state of the ever-growing Text Encoding Initiative (TEI) in the Digital Humanities at the time of publication. Although the author refers to TEI as a Web "standard," governed by a consortium, not unlike any other mark-up language, its largely academic home(s) have positioned it as, more appropriately, a *community.* As of the writing of this article, TEI had eighty-four institutional members from eighteen different countries. Anecdotally, the community has grown harder to penetrate, if not less accepting, with the heavyweight contributors to the listserv characterized more and more by a power law. TEI, like other mark-up languages, needs to strike a balance in the number of elements it incorporates, as it is constantly being mutated by its various stakeholders to serve different aims. The author believes that the best way to minimize fragmentation going forward is by migrating much of TEI to a cloud-server architecture, helping to support the shrinking contributor base while maximizing use.

Jong, F. De. (2009). NLP and the humanities: the revival of an old liaison. Proceedings from: *The 12th Conference of the European Chapter of the Association for Computational Linguistics*. Athens, Greece.

This article reviews current research in the digital humanities and natural language processing, particularly in the field of probabilistic language modeling, with the hope of harmonizing disciplinary research agendas for the coming years. The author offers the example of speech recognition models that can be "trained" on the available metadata and supplementary data sources. This neatly supports research interests in multiple domains; humanists can use expert knowledge to compile ever-greater language corpora, which computational linguists can then use to iteratively refine their analysis tools. These analysis tools can then be reflexively applied to *surprise* datasets captured in the field, speeding new analysis in the humanities. The author articulates a now very frequently-heard need for common platforms, repositories, and standards, ideally *open* standards. From there, evaluation metrics would be helpful.

Wulfman, Clifford E. (2009). The Perseus Garner: Early Modern Resources in the Digital Age. *College Literature, 36*(1), 18–25.

The Perseus Garner is an experimental system for visualizing hypertextual connections between primary and secondary research texts from the Early Modern period. Wulfman (2009) begins with the casual admittance that "many [digital humanities] tools and texts are incunabula, and their designs and obsessions already seem quaint" (p. 19). While his modesty is laudable, the Perseus Library, and the Perseus Garner, are clearly more than the sum of their parts; he notes that the fact of the library materials being co-located, and thus conforming to the same digital standard, is infinitely more valuable than their having been digitized in the first place. As such, the system is very, very good at visualizing hyperlinks; references can be automatically extracted from the text and made interactive within the article margins. However, this capability has informed the design of the system to such an extent that the links are foregrounded against the text itself, making them quite difficult to ignore as a reader might otherwise be free to do in a traditional reading environment. While this development is disturbing, it does remind us that such an extensive hypertextual network does not, in the author's words, "relieve the reader of the burden of judgment" (p. 23).

Jewell, Michael. (2010). Semantic Screenplays: Preparing TEI for Linked Data. Proceedings from: *The Digital Humanities Conference*. London, UK.

This straightforwardly-named article provides a brief research-oriented summary of the author's efforts to augment the existing Text Encoding Standard to support linked data using the RDFa (Resource Description Format in Attributes) schema. Jewell (2010) notes that linked data is growing in popularity from the early days of theorizing the semantic Web through complex modern ontologies for linking attributes of bibliographic data (in this case, films). What he does not note, and what should not go ignored, is that linked data is meanwhile rapidly gaining attention in the electronic publishing sphere, a development that is of particular relevance to digital humanists seeking to enhance our capabilities for document sharing and distribution. Still, the fact that this article is only formatted as plain text does not preclude the author from providing descriptive examples of his mark-up translation tool, dubbed *tei2onto.*

Rockwell, Geoffrey, Stéfan G. Sinclair, Stan Ruecker, and Peter Organisciak. (2010). Ubiquitous Text Analysis. *Poetess Archive Journal, 2*(1), n.p.

This experimental review article from the *Poetess Archive* is largely composed of annotated text analysis prototypes, some of which are literally interspersed into the HTML article view using javascript. The authors discuss usability issues that plagued early prototypes, such as not offering separate interfaces to developers and end-users. It then took several years more to abandon the notion that novel reading and writing interfaces should necessarily take the form of a "workbench," which was not, it turns out, how many humanists characterized their workspaces. Later, the advent of extensible Web browser environments confused development resources between ubiquitous but minimally useful "bookmarklet" add-ins and fully-featured Firefox or Chrome plug-ins. From these experiences, the authors put forth several recommendations for the future: avoid opaque, non-scalable toolsets like Flash; avoid immature or closed plug-in models; and, above all, work to harmonize the cultures of text analysis and digital libraries.

Reviewing humanities computing

The articles in this section represent a snapshot of stakeholder views on the digital humanities research ecosystem in the past half-decade. Although the digital humanities are technically many disciplines, and there is appropriately little effort made here to *summarize* the breadth of research currently being undertaken across the globe, these papers provide a helpful view of some especially fruitful pursuits informed by qualitative and quantitative retrospective analysis.

Smith, Martha Nell. (2004). Electronic Scholarly Editing. In Ray Siemens, Susan Schreibman, and John Unsworth (Eds.), *A Companion to Digital Humanities* (pp. 306–322). Oxford: Blackwell.

This entry from the edited volume *A Companion to Digital Humanities* reflects on the new opportunities and challenges occasioned by the move to digital editing, eight years ago. It is interesting to note how little has changed in the intervening near-decade as the author makes mention of online commenters' ability to effect *immediate* "peer review" after an article is published, noting that this still has not dramatically shifted editorial practices away from an overwhelming focus on preening an article for its debut. There are also some remarkably charming discussions of lexicography: "Strong positions have been taken about whether the shorter en- or the longer em-mark most faithfully translates [Emily] Dickinson's mark into print, and [editor R.W.] Franklin has resolved the matter with the authoritarian stance that neither the en- nor the em-suffice: according to him, the shorter-than-either hyphen best conveys Dickinson's practice of using a horizontal or angled mark rather than a comma" (p. 310). There is also a rich discussion of encoding formats, and of representation and truth, certainly echoed elsewhere but rarely so richly and with broad acknowledgments of metadata standards and structures outside of the Text Encoding Initiative (TEI).

Schreibman, Susan, Ray Siemens, and John Unsworth (Eds.). ***A Companion to Digital Humanities*****. Oxford: Blackwell.**

This edited volume remains a landmark in the Digital Humanities, a now seven-year-old "story so far" containing contributions from many leading scholars that is hardly any less relevant now than when it was published. The book's breadth of content speaks to its varied audience; there are individual chapters recounting the computing history of art history, lexicography, and performing arts to name only a few, opposite a deceptively artful chapter titled *How the Computer Works*. Hundreds of pages in the latter half of the book are dedicated to exploring various applications of the digital humanities, ensuring that this remains an excellent starting point for new researchers. It is perhaps only the digital humanities whose core inter-disciplinarity could be reviewed so completely in a single monograph without appearing to be divorced from the "digital." It makes sense, then, that as an added bonus the complete volume is freely available online.

McCarty, Willard. (2005). ***Humanities Computing*****. Basingstoke: Palgrave MacMillan.**

McCarty, long a preeminent voice in the digital humanities, offers this solo authored volume very much in the humanist's spirit; digital libraries and digital publishing are a distant second-fiddle to an exhaustive rendition of current modes of analysis. He alternately discusses textual modeling, new avenues for scholarly commentary, and research agendas, all but the latter in broadly theoretic terms. As such, the book fulfills a very different purpose from the Blackwell *Companion;* where the *Companion* is a broadly grounded review of ongoing research, McCarty's (2005) book will probably never be dated because he makes the digital subservient to the humanities, a rare feat that is nothing if not interesting.

Warwick, Claire, Melissa Terras, Paul Huntington, and Nikoleta Pappa. (2007). If You Build It, Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data. ***Literary and Linguistic Computing, 23*****(1), 85–102.**

This article reports on the Log Analysis of Internet Resources in the Arts and Humanities (LAIRAH) project, assessing the long-term use of digital resources. The study in question used deep log analysis; that is, literally, analyzing all of the available data, without sampling, to provide as complete a representation as possible. The researchers also used a parallel interview study to understand why resources were or were not being well-used. They note that they were able to do so efficiently by giving participants a very short time (5-10 minutes) to assess the utility of a resource, on the grounds that users typically spend comparably little time in making utility judgments in naturalized contexts. The study showed, perhaps unsurprisingly, that name recognition was an important factor contributing to the use of the resource. This name recognition, however, took two forms: resources were not only better used when they contained easily navigable, Wikipedia-type proper noun hyperlinks inward and outward, but also when they were clearly titled. For example, simply having the name

"Exeter" in the Exeter Cathedral Keystones and Carvings Project seemed to measurably increase traffic. The findings also suggested "that there is a scholarly bifurcation between those who create specialist digital resources as part of their research, but do not tend to reuse those of others, and those who prefer to use more generic information resources, but are less concerned with deposit and archiving" (p. 97).

Juola, Patrick. (2007). Killer Applications in Digital Humanities. *Literary and Linguistic Computing, 23*(1), 73-83.

This article serves as a response to recent surveys of non-digital humanists who have proven unmoved by research into digital methods of analyses, and ignorant of what the digital humanities can do to expand their horizons to their disciplinary benefactors. The author begins with the troubling evidence of what he calls the "rock-bottom" (p. 75) impact factor of *Computers and the Humanities,* the field's longest-running journal. Although this metric hardly spells doom on its own (and the author notes this in a somewhat self-defeating fashion, worrying about the significance of participation in the digital humanities by prestigious research universities), it does indicate that digital humanities content is not often reused. The article then highlights several projects that the author believes may become "killer applications" (i.e., individual implementations, theories, or methodologies that drive the adoption of the entire field) for the digital humanities, selling the whole of the field single-handedly. While it is not at all clear that this "killer application" philosophy is suitable to academic research, nor to the digital humanities in particular, the article nevertheless provides an accurate summary of promising research avenues in this field.

McCarty, Willard. (2008). Can we build it? Lessons and speculations on literary computing. Seminar: *An Foras Feasa.* Maynooth: National University of Ireland.

This lecture script reviews different perspectives on digitization in the humanities over the last half-century. The author makes no secret of his belief that "literary computing has had very little to say in response to critical discourse" (n.p.) over this time period. One exception to this, for him, has been the digital humanities' willingness to study implementation, providing a much-needed complement to the typical language of criticism. Without delving too deeply into epistemic questions of whether criticism *is* implementation or vice versa, McCarty (2008) concludes, somewhat abruptly, with a call for the development of more "play spaces" in the digital humanities. He offers the example of the IVANHOE project: situating, visualizing, and preserving discourse around a single digital object. It is probably not lost on him that such a system is perhaps the *de facto* digital humanities tool of the past decade, iterated tens of times, and yet there clearly remains work to be done.

Sculley, D., and Bradley M. Pasanek. (2008). Meaning and mining: the impact of implicit assumptions in data mining for the humanities. *Literary and Linguistic Computing, 23*(4), 409–424.

This article's title is rather self-explanatory, and no less enlightening for it. The authors seek to inform a data mining-naïve audience in the humanities about the problematic breadth of data evidence that can be counted as "positive," made all the more dangerous, they claim, by the humanist's pursuit of novel subjectivity. Data mining often carries with it the promise of a *fixed,* empirical distribution that can continue to provide new evidence over time but should not become any less true for it. This can be taken for granted in the humanities, where data is ironically more often incomplete. To test their prejudices, the article conducts a neat study of eighteenth-century political metaphors, à la George Lakoff. While the results of this study seem to get very far away from the discussion of metaphors, and are not reported on in any particularly clear manner, the plain-language review of machine learning techniques that the authors provide instead is more than worthwhile.

Schreibman, Susan, and Ray Siemens (Eds.). (2008). *A Companion to Digital Literary Studies*. Oxford: Blackwell. Oxford.

This edited volume, a spiritual follow-up to 2004's *Companion to Digital Humanities*, does not attempt to duplicate its predecessor's breadth; instead, it takes a focused look at literary studies through the lens of the digital humanities. The result is a book with a much more experimental character than the *Companion to Digital Humanities* which aimed to review the history of digital humanities subfields, with much of the work documented here nascent or speculative. There is, for example, a retrospective of *Private Public Reading: Readers in Digital Literature Installation*, and an ample body of work on digital worlds, analyzing representation and play. Still, more than a third of the volume is dedicated to clearly illustrating various methodologies currently in use in the digital humanities, ensuring that the book's relevance will not be diminished should all of its open questions manage somehow to be solved.

Siemens, Ray, John Willinsky, Analisa Blake, Greg Newton, Karin Armstrong, Lindsay Colahan. (Forthcoming). A Study of Professional Reading Tools for Computing Humanists." *Digital Humanities Quarterly*.

This comprehensive write-up of an in-progress qualitative study seeks to document how the design and structure of scholarly reading environments affect the experience of novice and expert users. The study focuses primarily on the tool suite embedded in the Public Knowledge Project's *Open Journal Systems* (OJS) platform, and particularly hypertext-enabling features within OJS. Much of the feedback obtained through interviews centred on the limited information resources made available through OJS. Although "expertise," in both information assessment and the domain itself, may enable users to search the open Web for supplementary information outside the narrow snapshot that is available to the interface, the OJS reading tools were thought to be particularly helpful for novice users, providing a finite and trustworthy set of links. Among the requested future developments were a more sophisticated open access content recommender system, as well as fewer tool categories, more of which should be contextual and keyword-driven.

Note

The authors would like to thank Julie Melone, who began this work by gathering 21 annotations to works essential to those working in the area. This kernel was built upon considerably by Alex Garnett, working consultatively with Cara Letich, Ray Siemens, and members of the Implementing New Knowledge Environment (INKE) and Public Knowledge Project (PKP) research groups.

References

Allinson, Julie, Sebastien, François, Sebastien, & Stuart Lewis, Stuart. (2008). SWORD: Simple Web-service offering repository deposit. *Ariadne, 54*.

Aschenbrenner, Andreas, Blanke, Tobias, Flanders, David, Hedges, Mark, & O'Steen, Ben. (2008). The future of repositories? Patterns for (cross-)repository architectures. *D-Lib Magazine, 14*(11/12).

Baker, Collin F., Fillmore, Charles J., & Lowe, John B. (1998). The Berkeley FrameNet project. Proceedings From*: The 17th International Conference on Computational Linguistics.*

Baker, Paul. (2006). *Using corpora in discouse analysis*. London: Continuum.

Balnaves, Edmund. (2005). Systematic approaches to long term digital collection management. *Literary and Linguistic Computing, 20*(4), 399–413.

Battino, Paolo, & Lancioni, Tarcisio. (2010). Visualization and narrativity: A generative semiotics approach. Proceedings from: *The Digital Humanities Conference*. London, UK.

Blanke, Tobias, & Hedges, Mark. (2010). A data research infrastructure for the arts and humanities. In Simon C. Lin and Eric Yen (Eds.), *Managed Grids and Cloud Systems in the Asia-Pacific Research Community* (pp. 179–191). Boston, MA: Springer.

Blanke, Tobias, & Hedges, Mark. (2008). Providing linked-up access to cultural heritage data. Proceedings from: *ECDL 2008 Workshop on Information Access to Cultural Heritage.* Aarhus, Denmark.

Borgman, Christine. (2009). The digital future is now: A call to action for the humanities. *Digital Humanities Quarterly, 3*(4).

Bradley, John. (2008). Thinking about interpretation: Pliny and scholarship in the humanities. *Literary and Linguistic Computing, 23*(3), 263–279.

Brase, Jan. (2009). DataCite-A global registration agency for research data. Proceedings from: *Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology.* Beijing, China.

Burnard, Lou. (2005). Metadata for corpus work. In Martin Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 30–46). Oxford: Oxbow.

Buyya, Rajkumar, & Venugopal, Srikumar. (2005). A gentle introduction to grid computing and technologies. *CSI Communications, 29*(1), 9-19.

Chowdhury, Gobinda. (2002). Digital divide: How can digital libraries bridge the gap? *Lecture Notes in Computer Science – Digital Libraries: People, Knowledge, and Society, 2555*, 379–391.

Coyle, Karen. (2006). Mass Digitization of Books. *Journal of Academic Librarianship, 32*(6), 641–645.

Craig, Hugh, & Whipp, R. (2009). Old spellings, new methods: automated procedures for indeterminate linguistic data. *Literary and Linguistic Computing, 25*(1), 37–52.

Crane, Gregory. (2000). Designing documents to enhance the performance of digital libraries: Time, space, people and a digital library of London. *D-Lib Magazine, 6*(7/8).

Crane, Gregory, Babeu, Alison, & Bamman, David. (2007). eScience and the humanities. *International Journal on Digital Libraries, 7*(1-2), 117-122.

Crane, Gregory, & Rydberg-Cox, Jeffrey A. (2000). New technology and new roles: The need for 'corpus editors.' Proceedings from: *The fifth ACM Conference on Digital Libraries*. New York: ACM.

Crane, Gregory, Seales, Brent, & Terras, Melissa. (2009). Cyberinfrastructure for classical philology. *Digital Humanities Quarterly, 3*(1).

Crane, Gregory, & Wulfman, Clifford. (2003). Towards a cultural heritage digital library. Proceedings from*: The 3rd ACM/IEEE-CS Joint Conference on Digital Libraries.* Washington, DC: IEEE Computer Society.

Crane, Gregory, Wulfman. Clifford E., & Smith, David A. (2001). Building a hypertextual digital library in the humanities: A case study on London. Proceedings from*: The first ACM/IEEE-CS Joint Conference on Digital Libraries*. New York: ACM Press.

Davies, Mark. (2009). The 385+ million word corpus of contemporary American English (1990-2008+): Design, Architecture, and Linguistic Insights. *International Journal of Corpus Linguistics, 14*(2), 159-190.

Davies, Mark. (2010). The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing, 25*(4).

Davis, Boyd. (2000). Taking advantage of technology. Language and digital technology: Corpora, contact, and change. *American Speech, 75*(3), 301-303.

Drucker, Johanna. (2011). Humanities approaches to graphical display. *Digital Humanities Quarterly, 5*(1).

Ebeling, Jarle. (2007). The electronict corpus of Sumerian literature. *Corpora, 2*(1), 111-120.

Ebeling, Signe O., & Heuboeck, Alois. (2007). Encoding document information in a corpus of student writing: The British Academic Written English Corpus. *Corpora, 2*(2), 241-256.

Flowerdew, Lynne. (2009). Applying corpus linguistics to pedagogy: A critical evaluation. *International Journal of Corpus Linguistics, 14*(3), 393-417.

Gleim, Rüdiger, Waltinger, Ulli, Ernst, Alexandra, Mehler, Alexnder, Feith, Tobias, & Dietmar Esch,Dietmar. (2009). eHumanities desktop: An online system for corpus management and analysis in support of computing in the humanities. Proceedings from: *The 12th Conference of the European Chapter of the Association for Computational Linguistics*. Athens, Greece.

Gold, Anna. (2007). Cyberinfrastructure, data, and libraries, part 1: A cyberinfrastructure primer for librarians. *D-Lib Magazine, 13*(9).

Gold, Nicolas. (2009). Service-oriented software in the humanities: A software engineering perspective. *Digital Humanities Quarterly, 3*(4).

Gow, Jeremy, Buchanan, George, Blandford, Ann, Warwick, Claire, & Rimmer, Jon. (2009). *User-centred requirements for document structure in the humanities*. In Preparation.

Green, Richard, & Awre, Chris. (2009). Towards a Repository-enabled Scholar's Workbench: RepoMMan, REMAP and Hydra. *D-Lib Magazine,* 15(5/6).

Groth, Paul, Gibson, Andrew, & Velterop, Johannes . (2010). The anatomy of a nanopublication. *Information Services and Use, 30*(1-2), 51-56.

Han, Yan. (2004). Digital content management: The search for a content management system. *Library Hi Tech, 22*(4), 355-365.

Hedges, Mark. (2009). Grid-enabling humanities datasets. *Digital Humanities Quarterly, 3*(4).

Hicks, Diana, & Wang, Jian. (2009). Towards a bibliometric database for the social sciences and humanities. URL: http://works.bepress.com/diana_hicks/18/ [July 15, 2011].

Honkapohja, Alpo, Kailaniemi, Samuli, & Marttila, Ville. (2009). Digital editions for corpus linguistics: Representing manuscript reality in electronic corpora. Proceedings from: *The 29th International Conference on English Language Research on Computerized Corpora*. Oslo, Norway.

Ide, Nancy. (2008). Preparation and analysis of linguistic corpora. In Susan Schreibman & Ray Siemens (Eds.), *A Companion to Digital Literary Studies* (pp. 289-305). Oxford: Blackwell.

Jannidis, Fotis. (2009). TEI in a crystal ball. *Literary and Linguistic Computing, 24*(3), 253-265.

Jessop, Martyn. (2008). Digital visualization as a scholarly activity. *Literary and Linguistic Computing, 23*(3), 281-293.

Jessop, Martyn. (2007). The inhibition of geographical information in digital humanities scholarship. *Literary and Linguistic Computing, 23*(1), 39-50.

Jewell, Michael. (2010). Semantic screenplays: Preparing TEI for linked data. Proceedings from: *The Digital Humanities Conference*. London, UK.

Jong, F De. (2009). NLP and the humanities: the revival of an old liaison. Proceedings from: *The 12th Conference of the European Chapter of the Association for Computational Linguistics*. Athens, Greece.

Juola, Patrick. (2007). Killer pplications in digital humanities. *Literary and Linguistic Computing, 23*(1), 73-83.

Kim, Yunhyong, & Seamus Ross, Seamus. (2007). 'The naming of cats': Automated genre classification. *International Journal of Digital Curation, 2*(1), 49-61.

King, Gary. (2007). An introduction to the Dataverse Network as an infrastructure for data sharing. *Sociological Methods & Research, 36*,173-199.

Klavans, Judith, Sidhu, Tandeep, Sheffield, Carolyn, Soergel, Dagobert, Lin, Jimmy, Abels, Eileen, & Passoneau, Rebecca. (2008). Computational linguistics for metadata building (CLiMB) Text mining for the automatic extraction of subject terms for image metadata. Proceedings from: *VISAPP Workshop Metadata Mining for Image Understanding*. Madeira, Portugal.

Kucsma, Jason, Reiss, Kevin, & Sidman, Angela . (2010). Using Omeka to build digital collections: The METRO case study. *D-Lib Magazine, 16*(3/4).

Lagoze, Carl, Payette, Sandra, Shin, Edward, & Wilper, Chris. (2005). Fedora: An Architecture for Complex Objects and their Relationships. *International Journal on Digital Librarie*s, *6*, 124-138.

Levy, David M., & Marshall, Catherine C. (1995). Going Digital: A look at assumptions underlying digital libraries. *Communications of the ACM, 38*(4), 77–84.

Luyckx, Kim, Daelemans, Walter, & Vanhoutte, Edward. (2006). Stylogenetics: Clustering-based stylistic analysis of literary corpora. Presented at: *The 5th International Language Resources and Evaluation Conference*. Genoa, Italy.

Marchionini, Gary. (2000). Evaluating digital libraries: A longitudinal and multifaceted view. *Library Trends, 49*(2), 304-333.

Marcial, Laura, & Hemminger, Brad. (2010). Scientific data repositories on the Web: An initial survey. *Journal of the American Society for Information Science and Technology, 61*(10), 2029-2048.

Marcus, Mitchell P., Marcinkiewicz, Mary Ann, & Beatrice Santorini. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*(2), 313-330.

Matei, Sorin Adam, Wernert, Eric, & Faas, Travis. (2009). Where information searches for you: The visible past ubiquitous knowledge environment for digital humanities. Proceedings from: *The IEEE International Conference on Computational Science and Engineering*. Hong Kong, China.

McCarty, Willard. (2005). *Humanities computing*. Basingstoke: Palgrave MacMillan.

McCarty, Willard. (2008). Can we build it? Lessons and speculations on literary computing. Seminar: *An Foras Feasa*. Maynooth: National University of Ireland.

Mehler, Alexander. (2008). Large text networks as an object of corpus linguistic studies. In A. Ludeling and K. Marja (Eds.), *Corpus Linguistics: An International Handbook* (pp. 277). Berlin: de Gruyter.

Nooy, Wouter de. (2010). Network analysis of story structure. Proceedings from: *The Digital Humanities Conference*. London, UK.

Oard, Douglas. (2008). A whirlwind tour of automated language processing for the humanities and social sciences. Presented at *Working Together or Apart: Promoting the Next Generation of Digital Scholarship, Council on Library and Information Resources*. Washington, DC, USA.

Rayson, Paul, & Mariani, John. (2009). Visualising corpus linguistics. Proceedings from: *The Corpus Linguistics Conference*. Liverpool, UK.

Reilly, Sean & Tupelo-Schneck, Robert. (2010). Digital object repository server: A component of the digital object architecture. *D-Lib Magazine, 16*(1/2).

Rimmer, Jon, Warwick, Claire, Blandford, Ann, Gow, Jeremy, & Buchanan, George. (2008). An examination of the physical and the digital qualities of humanities research. *Information Processing and Management, 44*(3), 1374-1392.

Rockwell, Geoffrey, Sinclair, Stéfan G., Ruecker, Stan & Peter Organisciak. (2010). Ubiquitous text analysis. *Poetess Archive Journal, 2*(1).

Rosenthal, David S. H., Robertson, Thomas, Lipkisi, Tom, Reich, Vicky, & Morabito, Seth. (2005). Requirements for digital preservation systems: A bottom-up approach. *D-Lib Magazine, 11*(11).

Rydberg-Cox, Jeffrey A. (2006). *Digital libraries and the challenges of digital humanities.* Oxford: Chandos Press.

Rydberg-Cox, Jeffrey A. (2009). Digitizing Latin incunabula: Challenges, methods, and possibilities. *Digital Humanities Quarterly, 3*(1).

Schreibman, Susan, Kumar, Amit, & McDonald, Jarom. (2003). The versioning machine. *Literary and Linguistic Computing, 24*(3), 339-346.

Schreibman, Susan, & Siemens, Ray. (Eds.). (2008). *A companion to digital literary studies*. Oxford: Blackwell.

Schreibman, Susan, Siemens, Ray, & Unsworth, John. (Eds.). *A companion to digital humanities*. Oxford: Blackwell.

Scifleet, Paul, & Williams, Susan P. (2009). Practice theory & the foundations of digital document encoding. Proceedings from: *The 27th Annual ACM Conference on Design of Communication*. Bloomington, Indiana, USA.

Sculley, D., & Pasanek, Bradley M. (2008). Meaning and mining: the impact of implicit assumptions in data mining for the humanities. *Literary and Linguistic Computing, 23*(4), 409-424.

Seadle, Michael, & Greifeneder, Elke. (2007). Defining a digital library. *Library Hi Tech, 25*(2), 169-173.

Sefton, Peter. (2009). The fascinator: A lightweight, modular contribution to the Fedora-commons world. Proceedings from: *Fourth International Conference on Open Repositories*. Atlanta, Georgia, USA.

Sharoff, Serge. (2006). Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics, 11*(4), 435-462.

Siemens, Ray, Willinsky, John, Blake, Analisa, Newton, Greg, Armstrong, Karin, & Colahan. Lindsay. (Forthcoming). A study of professional reading tools for computing humanists. *Digital Humanities Quarterly*.

Sinclair, John. (2005). Corpus and text – basic principles. In Martin Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 1-16). Oxford: Oxbow.

Smith, Martha Nell. (2004). Electronic scholarly editing. In Ray Siemens, Susan Schreibman, & John Unsworth (Eds.), *A Companion to Digital Humanities* (pp. 306-322). Oxford: Blackwell.

Smith, MacKenzie, Barton, Mary, Bass, Mick, Branschofsky, Margret, McClellan, Greg, Stuve, Dave, Tansley, Robert, & Walker, Julie Harford. (2003). DSpace: An open source dynamic digital repository. *D-Lib Magazine, 9*(1).

Sperberg-McQueen, C. M. (2008). Classification and its structures. In Susan Schreibman & Ray Siemens (Eds.), *A Companion to Digital Literary Studies* (pp. 161-176). Oxford: Blackwell.

Staples, Thornton, Wayland, Ross, & Payette Sandra. (2003). The Fedora Project: An Open-source Digital Object Repository Management System. *D-Lib Magazine, 9*(4).

Stewart, Gordon, Crane, Gregory, & Babeu, Alison. (2007). A new generation of textual corpora. Proceedings from: *The 7th ACM/IEEE Joint Conference on Digital Libraries*. Vancouver, Canada.

Terras, Melissa M. (2009). The potential and problems in using high performance computing in the arts and humanities: The Researching e-Science Analysis of Census Holdings (ReACH) project. *Digital Humanities Quarterly, 3*(4).

Thaller, Manfred. (2001). From the Digitized to the Digital Library. *D-Lib Magazine 7*(2).

Thelwall, Mike. (2005). Creating and using Web corpora. *International Journal of Corpus Linguistics, 10*(4), 517-541.

van Deemter, Kees, van der Sluis, Ielka, & Gatt, Albert. (1998). Building a semantically transparent corpus for the generation of referring expressions. Proceedings from: *The Fourth International Natural Language Generation Conference*. Morristown, NJ: Association for Computational Linguistics.

Viterbo, Paolo Battino, & Gourley, Donald. (2010). Digital humanities and digital repositories. Proceedings from: *28th ACM International Conference on Design of Communication*. Sao Paolo, Brazil.

Voss, Alex, Mascord, Matthew, Fraser, Michael, Jirotka, Marina, Procter, Rob, Halfpenny, Peter, Fergusson, David, Atkinson, Malcolm, Dunn, Stuart, Blanke, Tobias, Hughes, Lorna, & Anderson, Sheila. (2007). e-Research infrastructure development and community engagement. Proceedings from the *UK e-Science All Hands Meeting 2007*. Nottingham, UK.

Warwick, Claire, Galina, Isabel, Rimmer, Jon, Terras, Melissa, Blanford, Jeremy, & Buchanan, George. (2009). Documentation and the users of digital resources in the humanities. *Journal of Documentation,* 65(1), 33-57.

Warwick, Claire, Terras, Melissa, Huntington, Paul, & Pappa, Nikoleta. (2007). If you build it, will they come? The LAIRAH study: Quantifying the use of online resources in the arts and humanities through statistical analysis of user log data. *Literary and Linguistic Computing, 23*(1), 85-102.

Waugh, Andrew. (2007). The design and implementation of an ingest function to a digital archive. *D-Lib Magazine, 13*(11/12).

White, Hollie. (2010). Considering personal organization: Metadata practices of scientists. *Journal of Library Metadata, 10*(2), 156-172.

Williams, Geoffrey. (2010). Many rooms with corpora. *International Journal of Corpus Linguistics, 15*(3), 400-440.

Witten, Ian H. (2003). Examples of practical digital libraries: Collections built internationally using Greenstone. *D-Lib Magazine, 9*(3).

Witten, Ian H., & Bainbridge, David. (2007). A retrospective look at Greenstone: Lessons from the first decade. Proceedings from*: 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 147-156). New York: ACM.

Witten, Ian H., Bainbridge, David, & Boddie, Stefan J. (2001). Greenstone: Open-source digital library software. *D-Lib Magazine, 7*(7).

Witten, Ian H., Bainbridge, David, Tansley, Robert, Huang, Chi-Yu, & Don, Katherine J. (2005). StoneD: A bridge between Greenstone and DSpace. *D-Lib Magazine, 11*(9).

Wulfman, Clifford E. (2009). The Perseus Garner: Early modern resources in the digital age. *College Literature, 36*(1), 18-25.

Wynne, Martin. (2005). Archiving, distribution and preservation. In Martin Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 71-78). Oxford: Oxbow.

Yu, Bei. (2008). An evaluation of text classification methods for literary study. *Literary and Linguistic Computing, 23*(2), 327-343.