A STUDY OF INFORMATION FRAGMENT ASSOCIATION IN INFORMATION MANAGEMENT AND RETRIEVAL APPLICATIONS

by

Demetrios Ioannides

B.A. Columbia University, 1984

M.A. Columbia University, 1986

M.S. Columbia University, 1990

Submitted to the Graduate Faculty of

The School of Information Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

UNIVERSITY OF PITTSBURGH

SCHOOL OF INFORMATION SCIENCES

This dissertation was presented

by

Demetrios Ioannides

It was defended on

August 3, 2007

and approved by

David B. Robins, PhD, Assistant Professor, Kent State University

Michael B. Spring, PhD, Associate Professor

Christinger Tomer, PhD, Associate Professor

Dissertation Director: Daqing He, PhD, Assistant Professor

A STUDY OF INFORMATION FRAGMENT ASSOCIATION IN INFORMATION MANAGEMENT AND RETRIEVAL APPLICATIONS

Demetrios Ioannides, PhD

University of Pittsburgh, 2007

As we strive to identify useful information sifting through the vast number of resources available to us, we often find that the desired information is residing in a small section within a larger body of content which does not necessarily contain similar information. This can make this Information Fragment difficult to find. A Web search engine may not provide a good ranking to a page of unrelated content if it contains only a very small yet invaluable piece of relevant information. This means that our processes often fail to bring together related Information Fragments. We can easily conceive of two Information Fragments which according to a scholar bear a strong association with each other, yet contain no common keywords enabling them to be collocated by a keyword search.

This dissertation attempts to address this issue by determining the benefits of enhancing information management and retrieval applications by providing users with the capability of establishing and storing associations between Information Fragments. It estimates the extent to which the efficiency and quality of information retrieval can be improved if users are allowed to capture mental associations they form while reading Information Fragments and share these associations with others using a functional registry-based design.

In order to test these benefits three subject groups were recruited and assigned tasks involving Information Fragments. The first two tasks compared the performance and usability of a mainstream social bookmarking tool with a tool enhanced with Information Fragment Association capabilities. The tests demonstrated that the use of Information Fragment Association offers significant advantages both in the efficiency of retrieval and user satisfaction. Analysis of the results of the third task demonstrated that a mainstream Web search engine performed poorly in collocating interrelated fragments when a query designed to retrieve the one of these fragments was submitted. The fourth task demonstrated that Information Fragment

Association improves the precision and recall of searches performed on Information Fragment datasets.

The results of this study indicate that mainstream information management and retrieval applications provide inadequate support for Information Fragment retrieval and that their enhancement with Information Fragment Association capabilities would be beneficial.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

1.1 BACKGROUND

1.1.1 The Information Users' Focus on Fragments

Even though the tools and the media available to us have changed dramatically over the centuries, our objective as we gather information has always been the same. Our struggle has always been to generate a set of interrelated information items useful to us for a specific project we are working on. We prefer to consult multiple resources both in order to verify the information we are gathering and in order to ensure that we have taken into consideration every element useful to us. At the same time, we make an effort to focus on the specific parts of the resources that are relevant to our project. In the process of selecting these specific information items we also draw associations between them in order to ensure that as we revisit and focus our attention on one of these items in the future we also have other closely associated items at our fingertips.

The scholar who piles up books in a dark and dusty corner of a library, inserting little bookmarks to identify the sections containing useful information, or the journalist who builds a file of clippings from various sources, are both aware of the fact that useful information is often found in specific passages within larger bodies of content. They are always in pursuit of innovative ways by which this process may be facilitated.

For the purpose of our discussion, we call these specific passages *Information Fragments* distinguishing them from the larger bodies of content which we call *Information Entities*. As for the associations established between Information Fragments, we call them *Information Fragment Associations*. Section 1.7 below provides more detailed definitions of these terms.

The concept of Information Fragment Association can be best demonstrated by a simple example. Suppose that a scholar working on Sappho (first passage in Figure 1 below) wishes to identify other instances in literature in which Love is compared to madness. The second passage, from Sophocles' Antigone includes such an allusion. We call Sappho's poem and Sophocles' tragedy *Information Entities*, since they are the large bodies of content produced as single works by the authors. We call the selections of specific parts of the content *Information Fragments*. By associating these two Information Fragments, a scholar makes sure that a visit to either of these two Information Entities will allow navigation to the other one. The association by itself makes some kind of a statement, but it can be further qualified by a description or by another association to an article discussing the issue

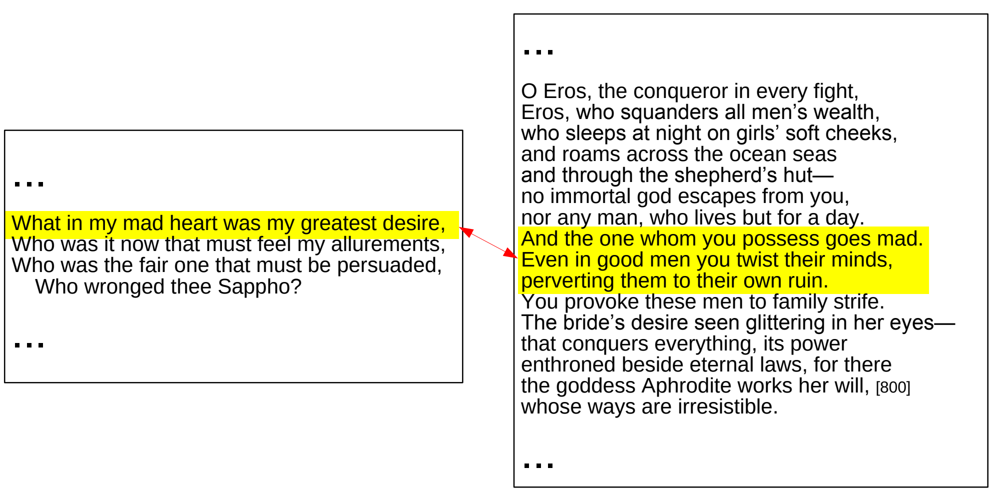


Figure 1 – Information Fragment Association

1.1.2 Information seeking techniques and their limitations when it comes to fragments – Fragment Association as a Statement

In general, we encounter information by employing two techniques:

a) Searching: In order to penetrate the content of the set of resources against which we apply our search and pinpoint the exact passages containing a discussion of the topic of interest to us, we employ free keyword searching. We try to translate our mental concepts into specific keywords likely to have been used in relevant passages.

b) Browsing: We identify a resource we consider highly relevant and we follow any references to other similar resources it may include. These potentially relevant resources can be pointed to by hypertext links in a web page or, traditionally, by simple references listed in a book's bibliography section.

In reality, these two techniques are used in tandem.

Unfortunately, when it comes to trying to identify specific Information Fragments within Information Entities, these techniques produce less than optimal results. It is particularly difficult in the current web environment to pinpoint information contained in a small Information Fragment. Navigation brings us usually to the top of an Information Entity or at best to a specific point within an Information Entity. It does not identify and display the boundaries of the Information Fragment. Similarly, keyword-in-context interfaces present us with a display of the keywords used in the search specification surrounded by an arbitrarily defined fragment, not an Information Fragment with exact boundaries. Moreover an Information Fragment may be highly relevant but it may still not contain the keywords we are using. Even if our search mechanism employs lookup vocabularies supplying matching synonyms, a match may still not be possible. As a result the Information Fragment may be missed by our query. The smaller the fragment, the more likely we are to miss it.

1.1.3 Difficulties with Fragment Associations

One of the beneficial consequences of the written word's moving from the scroll format to the codex format was that the codex format made pagination possible. A reference to a specific page in a book was perhaps not a perfect way of pointing to a useful fragment, but it was at least a way of focusing on the area around a useful fragment. With the advent of the web, we have for the most part moved back to the scroll format. Although it is not very common to have a body of text of the length of an entire book in a single web page, a very large number of web pages are too long to allow the reader to quickly scan them for a specific fragment reference. Even though HTML provides the capability of defining named anchors which can serve a function similar to pagination, their use has been very limited. Several web specifications[1-3] have dealt with the issue of granularity and have provided very satisfactory solutions[4]. Yet, the reality we are facing is that most web pages are still not built with any granularity defined, and even in the cases in which granularity exists there is lack of consistency. The majority of web page creators do not pay attention to such matters perhaps because they do not see any tangible evidence that doing so will be worth their effort.

Of course, even if web page creators were given facilities involving granularity to create references to fragments of their pages, our current environment would still not allow us to emulate the reader's mental process of identifying two fragments belonging to two different resources and creating an association between them. Accomplishing that would entail being able to define unambiguous references to both fragments and being able to record this information. Current web practice involves physically embedding the link in the referring web page. This makes it necessary that the creator of this link have the appropriate privileges to edit that page. Unfortunately, although this is possible in collaborative systems such as the wikis, it is not possible in the majority of cases in which the integrity of the original document is paramount. This is contrary to the spirit of anybody having the ability to associate any fragment with any other fragment. In other words, a scholar working with two primary sources created by two other authors has to be able to associate Information Fragments within these primary sources without altering them in any way. The question we are faced with is how to contrive a solution providing the capability of establishing XLink-like external links, ensuring persistency and incorporating creator ownership.

The lack of exact boundaries may not seem very important when it comes to casual browsing for information. It may be sufficient to point somebody in the general direction of useful information, but when it comes to creating associations between two Information Fragments, the inability to define boundaries may be crippling. An association between two Information Fragments is a kind of statement; perhaps a statement that needs further qualification, but still a statement. It is a statement expressing the relationship between two statement sets. Each statement set is contained within one of the two Information Fragments being associated. Boundary uncertainty can alter the interpretation of these statement sets thus generating uncertainty about the nature of this association.

1.1.4 Organization and Access to Fragment Associations

All of the issues listed above demonstrate that we lack a consistent and global way of making Information Fragment Associations available. Search engine results would potentially be enhanced if we were able to manage to capture the intellectual activity of Information Fragment Association creation and make it available to them. Current user interfaces for browsing also lack the functionality of allowing the user to easily navigate from one fragment to another. Browsing would be significantly enhanced by overcoming these shortcomings.

1.2 TYPES OF INFORMATION FRAGMENT ASSOCIATIONS

Information Fragment Association features as an indispensible characteristic of human communication regardless of the format in which this communication is recorded. Written communication strives to simulate mental activity. The tendency of carving a piece of content out of a larger body of content and extending it by associating it with something else is always present whether one uses papyrus scrolls or information systems. There is no difference in the way Information Fragment Association manifests itself in different formats. However, there are different *types* of Information Fragment Associations, the definition of which depends on the way the association is applied to the content. For the purposes of this Study we define three types of Information Fragment Association:

- Referential Information Fragment Association
- Annotative Information Fragment Association
- Relational Information Fragment Association

1.2.1 Referential Information Fragment Association

A *Referential Information Fragment Association* is defined as the association created in the process of composition of information content as a reference to information content found elsewhere. For example, a scholar writes an article in a paper publication and quotes verbatim, paraphrases or simply provides a reference to a section within another article. This act is considered to be a Referential Information Fragment Association. For the same reasons we consider Ted Nelson's *Transclusion* to be a form of Referential Information Fragment Association. In Figure 2 Wilson quotes an Information Fragment from Winkler.

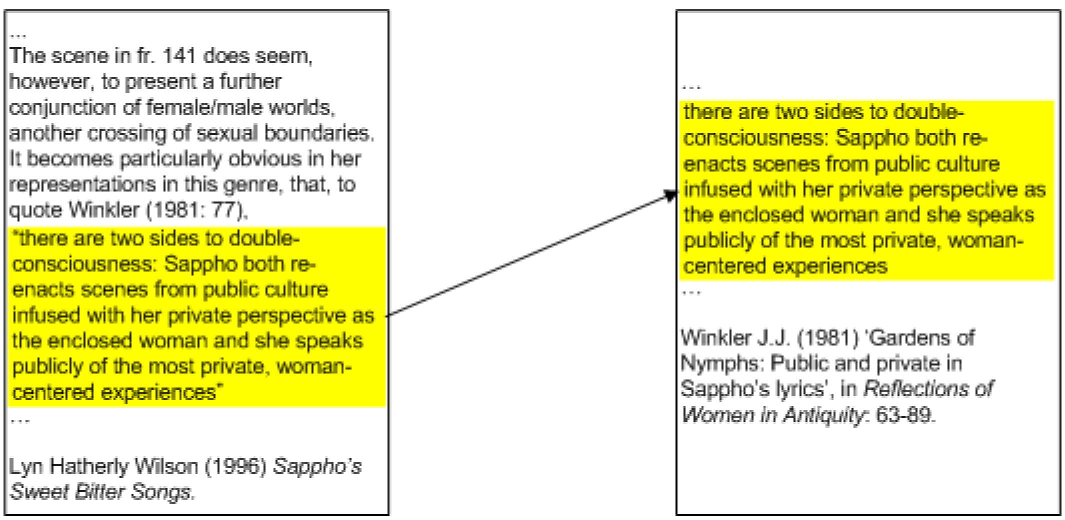


Figure 2 - Referential Information Fragment Association

1.2.2 Annotative Information Fragment Association

An *Annotative Information Fragment Association* is defined as the association created between an Information Fragment residing within an Information Entity and a comment made about that Information Fragment. This comment is composed with that specific Information Fragment in mind. It does not constitute an Information Entity nor is it an Information Fragment within a separate Information Entity. It is simply an Information Fragment attached additively to an Information Entity as an extension of its original content. The marginalia of a medieval manuscript or the entries of a traditional Commentary would constitute Annotative Information Fragment Associations. This is the type of Information Fragment Association supported by most modern annotation systems. Figure 3 shows an annotation made in Kenneth Quin's commentary on Catullus. Quinn's comment is attached to line 72 of poem 64.

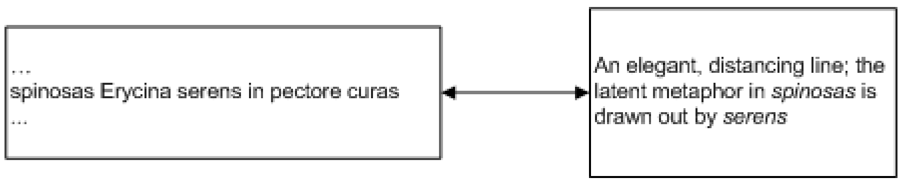


Figure 3 - Annotative Information Fragment Association

1.2.3 Relational Information Fragment Association

A *Relational Information Fragment Association* is defined as the association created between two Information Fragments residing within two Information Entities as a result of the realization that these two Information Fragments have some relationship which needs to be pointed out. Unlike the Referential Information Fragment Association, the Relational Information Fragment Association is not created during the composition of an Information Entity. It is the means by which a third-party associates information from two different sources. Traditional Commentaries make extensive use of this type of association when in addition to the annotations they provide about an Information Fragment, they point at similarities or differences between this Information Fragment and other Information Fragments. The left side of Figure 4 shows an entry in a commentary which refers to two other verses from two different poems of Catullus – poem 72, verse 2 and poem 55, verse 17. According to the commentator Kenneth Quinn these two verses are examples of use of the verb *teneo* having connotation similar to that of the same verb in poem 64 verse 28. These two references constitute two Relational Information Fragment Associations. These two Relational Information Fragment Associations are shown on the right side of Figure 4.

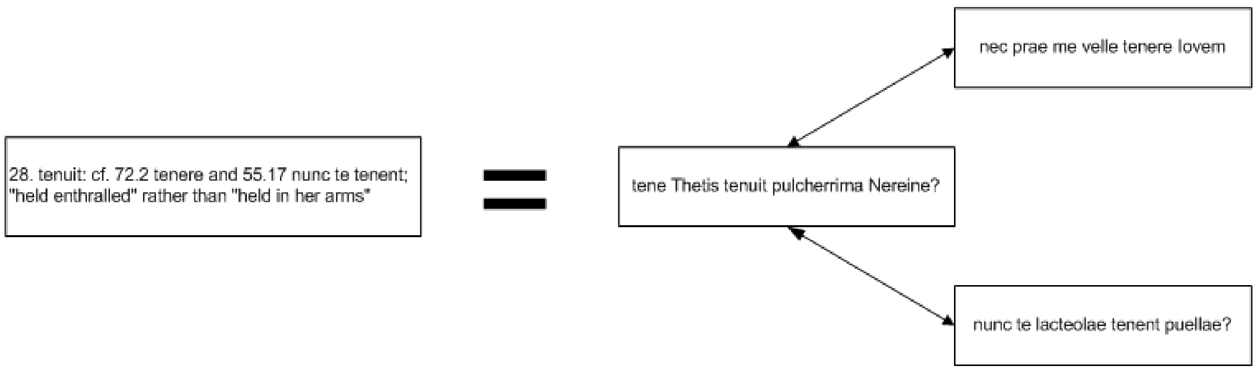


Figure 4 - Relational Information Fragment Association

1.2.4 Use of the Three Information Fragment Association Types in this Study

The distinction drawn here between these three types of Information Fragment Associations is made for the purpose of providing a more lucid outline of the scenarios in which Information Fragment Association is being used. Information Fragment Association is seen as a single model with three different manifestations. Each one of these types contains all of the characteristics of the model but it functions differently depending on its objective.

The terms used here have some similarity to terms used in DeRose's more detailed and comprehensive taxonomy of links (2.3.1) but they are not quite the same. DeRose does not draw the distinction between a reference and a relationship.

Having these three types in mind helps better understand the functionality of some of the systems presented in the Literature Review. Some of these systems support all three types, some of them only one. For example, the pure annotation systems support only the Annotative

Information Fragment Association. This dissertation deals with Information Fragment Association as a single entity. The discussion and conclusions of this Study apply to all three types, so none of these types is ever referred to individually below. The focus, however, is primarily on Referential and Relational Information Fragment Associations and far less on the Annotative.

1.3 STATEMENT OF THE PROBLEM

As the universe of information resources keeps expanding, users often feel overwhelmed with the quantity of content they have to consume in order to sieve out the specific information they need. The desired information is often residing within a larger body of content which does not necessarily contain similar information. On the other hand, similar information may reside in a multitude of small fragments residing within other, seemingly unrelated resources. Keyword searching is an invaluable tool used for penetrating through large bodies of content and identifying possibly useful fragments. However, although keyword searching is often an effective tool, we can easily conceive of two information fragments which according to a scholar bear a strong association with each other, yet contain no common keywords enabling them to be collocated by a keyword search. Associations can be drawn between thematic elements of otherwise completely different contents. For example, in spite of the lack of any plot similarities between Euripides' *Bacchae* and Shakespeare's *The Merchant of Venice*, some kind of association may be established between portions of a scene in the *Bacchae* in which a man is dressed like a woman and a scene in *The Merchant of Venice* in which a woman is dressed like a man.

Regardless of its format, an Information Fragment may contain both explicit and implicit content, the implicit content being evident to users with a common cultural or educational background. For example, an art critic may observe that one of the characters depicted in a Rafaello painting has the same melancholic countenance as one of the characters depicted in a Michelangelo painting. A user mentally associates an Information Fragment with another one judging that Information Fragment's implicit or explicit content as bearing a relationship with the implicit or explicit content of another Information Fragment. This judgment may be based on fact or on opinion. In either case, this mental association has the potential of being invaluable to other users. This Study attempts to estimate the extent to which the efficiency of information retrieval will be improved if users are allowed to capture their mental associations and share them with others using a functional design. Moreover, it attempts to evaluate the overall usefulness of Information Fragment Association to users and to determine what facilities should be employed in order to accomplish this.

1.3.1 Capturing and Re-using Users' Mental Associations

Search engines strive to bring some order to the chaos created by the vast accumulation of information resources on the Internet. Other than an often imperfect compliance with the HTML specification, most of these resources share very few structural similarities. As these engines sift through the mounds of available resources, they often benefit from any structured, commented or interlinked information they encounter. Since they continuously incorporate new features exploiting innovations in information generation, one cannot help but ponder of how much more needs to be done to present search engines with more hooks they can use to improve information retrieval.

The lack of such additional hooks has deprived search engines from information that users could have supplied with minimal effort and mostly as a result of their everyday routine. Information users have the ability of mentally establishing relationships between specific pieces of information as they encounter them. Ted Nelson points out that “links are made by individuals as pathways for the reader’s exploration”[5] (p.2/23). The current web environment allows the authors of web pages to provide links to be navigated by readers. Wouldn’t it be ideal if these links were created by the readers themselves? Is there a way to capture this information while at the same time facilitating the work of users instead of imposing an additional burden on them? Whatever these hooks may be, it is safe to assume that their success would be related to how simply they can be generated by users. Users have embraced the linking capabilities of HTML because they are very simple to create. If we introduce something new, it has to be as simple as that, or it will die of misuse.

These associations can serve as our additional hooks. How would these hooks be generated? Are existing technologies adequate to accommodate them or would there be a need for major changes? The answer to the last question is simple. Whatever is proposed here has to complement existing technologies, not replace them. The answer to the first question will be sought with further consideration of the issue.

1.3.2 Improving the User-friendliness of Currently Available Linking Technologies

Technical solutions to linking an information resource with an information fragment have been introduced. W3C’s XML Pointer [6], XML Base [7] and XML Linking [8] Specifications provide such capabilities. But what does this mean for the average user? Is the average user likely to embrace these technologies in their raw format in the same way that simple HTML linking has been embraced so far? Are users able to use these technologies to point to any information fragment in any information resource? The answer to both of these questions is no. Can these technologies provide a key to a solution allowing users to establish such links? The answer is yes. However, the particular techniques are not as important as the framework to which they are to contribute for this envisioned contribution to become a reality.

1.3.3 Providing Typed Link Capabilities

For expediency, this Study did not require subjects to use typed links. However the tools used were designed with that capability. The Study used the default typed link denoting “similarity” to associate the Information Fragments and the results are viewed with that restriction in mind.

It is desirable to have the option of attaching descriptive attributes to associations. Doing so may help search engines. Although anchor text is often invaluable as machines attempt to determine the general nature of an expert’s comments, the brevity of the comments and the ambiguity of the language used can lead to wrong readings. It is conceivable that an easy to use interface allowing users to create typed links using options available in general or domain-specific vocabularies would encourage some users to provide more “formal” information. Very interesting approaches at providing incremental formalization of informal user input have been proposed[9], albeit perhaps involving more steps than users would be willing to follow. If this formalization is not very ambitious, but rather focusing on detecting one or two important attributes such as negative vs. positive tone or the most descriptive term used in the anchor text, the creators of the links may not mind establishing these typed links if prompted for a simple verification.

1.4 THE STUDY OF INFORMATION FRAGMENT ASSOCIATION

The study of Information Fragment Association is being carried out with the ultimate goal of future development of a framework capable of handling Information Fragment Association and related issues at the web scale. However, the study presented in this dissertation has one modest goal: to assess the degree to which the enhancement of mainstream applications with Information Fragment Association capabilities is advantageous.

To achieve this goal, a study involving human subjects on several tasks was conducted around measuring the efficiency of the process of bookmarking and retrieving Information Fragments, estimating the accuracy of association, evaluating the user experience, determining the need for enhancement of results yielded by search engine when searching for Information Fragments and calculating the recall and precision of searches for Information Fragments in mainstream tools. Specifically, the study attempted to determine

- whether the total time needed for the entire process of bookmarking and then that of retrieving two Information Fragments within two Information Entities would be shorter if using a fragment association interface rather than using a mainstream social bookmarking tool.
- whether the overall usefulness and usability of the process of bookmarking and then that of retrieving two Information Fragments within two Information Entities would be better

if using a fragment association interface rather than using a mainstream social bookmarking tool.

- whether the rank difference of the two Information Fragments within the results of a search query designed to retrieve the first one is less than 10
- whether Information Fragment Association could improve the Recall, Precision and F-measure of keyword searches targeting a specific Information Fragment residing within an Information Entity

The Study consisted of 4 distinct tasks and was carried out by 3 subject groups each consisting of 6 University of Pittsburgh students. In Task 1 the subjects of Subject Group A and B were given twelve pairs of news stories and asked to define Information Fragments and establish an association between them. In Task 2, they were given the first Information Fragment and they were observed as they attempted to identify the second one. In Task 3, Group C was given the first Information Fragment of a pair and asked to compose a search query to be used to retrieve that Information Fragment from Live Search. By submitting this query the distance between the two Information Fragments within the search results was determined. In Task 4, the test administrator submitted the queries provided by the third group to the two tools under examination to determine Recall, Precision and F-measure.

The Study employed primarily quantitative research techniques but some qualitative analysis was also performed. The choice of the quantitative approach as opposed to the qualitative was based on the consideration of the objectives of the study and the nature of the tests feasible within the timeframe of the study. Since the main objective of the study was the examination of the effect that the presence or absence of Information Fragment Association functionality has on everyday information management and retrieval operations, the most suitable approach was a quantitative-experimental approach.

1.5 SIGNIFICANCE OF THE STUDY

Although one can argue that the benefits of Information Fragment Association seem to be self-evident, making such an assertion requires tangible proof.

This Study stemmed from the realization that by producing tangible evidence proving that retrieval performance and usability is significantly improved when Information Fragment Association is incorporated in mainstream applications, one can go a long way towards being able to make a strong argument for an integrated solution providing Information Fragment Association capabilities.

Another contribution of this Study was the fact that it provided an entirely different perspective for viewing the results of mainstream search engines. It has proven that contrary to our usual common assumptions it is often the case that two Information Entities containing two Information Fragments determined by users to have some strong relationship appear very far apart within the result sets of mainstream search engines. Sometimes a search query retrieving the one does not even retrieve the other one. This can be the case even if the entire Information Entities encompassing these two Information Fragments have related content. By demonstrating one area in which search engine results clearly need improvement and by doing so using data generated using a test prototype Information Fragment Association solution, this Study has produced results making a strong case for enriching search engine results with an additional result set dimension allowing the searcher to take advantage of Information Fragment Association data.

The outcome of this Study has given insights on the design of a comprehensive solution providing Information Fragment Association capabilities. Although it did not tackle every possible aspect of such a solution, it has provided a good basis for designing a framework for implementations providing these capabilities.

1.6 LIMITATIONS AND DELIMITATIONS

1.6.1 Limitations

- The news stories used in Task 1 and Task 2 of this test were pre-selected by the test administrator. It would have been preferable if these stories had been selected by the subjects themselves, but this would not have been possible within a reasonable amount of time. It was very important to allow the subjects to concentrate on the Information Fragment selection and submission rather than the preliminary task of identifying the stories. This time consuming task was undertaken by the test administrator. Attention was paid to making sure that the story pairs selected for this Study contained content with enough similarities. This made it easy for the subjects to identify related Information Fragments.
- In Task 3, instead of using the more popular Google search engine as originally planned this Study used the Live search engine. Google's December 2006 decision to reduce the resources available to its web services API affected the performance of the interface. For

this reason, the Live search engine API was selected because of its robust performance. This may have affected the findings of the Study slightly, but it is very unlikely that the outcome would be different otherwise.

- For practical purposes only the first 1000 results were considered in Task 3. Since several of these queries yielded more than 1000, in the cases in which the second Information Fragment was not found within these 1000 results it was not possible to determine whether it was at all present in the entire set of results. However the number was sufficiently high to provide a very high degree of confidence in the conclusions of the test. Any significant distance between the two Information Fragments was seen to support the notion that search engines have a lot to gain from the presence of Information Fragment Association capabilities.

1.6.2 Delimitation

The members of the test groups were primarily students with interest in news information resources and with some professed degree of comfort using online computer interfaces. Although advanced skills would have been beneficial to the test ensuring that they are highly motivated and diligent, this could have been a drawback as well. Average users may be far less capable of determining the topic of an Information Fragment than users with information management skills. The Study targeted students in the Arts and Sciences instead of Information Sciences. The varying degrees of commitment or competence of the subjects were taken into consideration for the determination of their assignment. For example, a subject with limited computer interface capabilities but with good searching skills was assigned to Group C which involved web searching instead of Group A or Group B.

1.7 DEFINITION OF TERMS

- ***Information Entity* --** Referring to instances of human communication can often be imprecise, as the definitions of information objects created as a result of human communication can vary. For this reason, a single term called "Information Entity" will be used throughout this discussion. We can informally define "Information Entity" as a single piece of human communication in a given state at a given moment. This means that an updated version of that piece of communication constitutes a different Information Entity. Specifically, as we are primarily talking about web pages, we can consider a web page as it is viewed in a browser by a user at a specific point in time to be an "Information Entity". The ability to cache a page at that specific moment would ensure the integrity of this Information Entity.
- ***Information Fragment* --** For the purposes of our discussion, an Information Fragment is a content portion of an Information Entity the boundaries of which have been defined by an Individual, the Fragment Association Creator.
- ***Individual* --** A person interacting with an Information Entity either as a creator or as a consumer (user).
- ***Entity Creator* –** An Individual who has created a given Information Entity
- ***Fragment Association Creator* --** An Individual who has created a given Information association between two Information fragments
- ***Identity* --** An Individual or an organized group of Individuals (e.g. a Professional Association) seen as a single entity. This Individual could be both an Entity Creator and a Fragment Association Creator.
- ***Identity Group* --** An Identity capable of containing multiple other Identities
- ***Information Fragment Association* --** The association between two Information Fragments created by an Identity. An Information Fragment Association may bear additional optional attributes describing this association.
- ***Information Fragment* Collection --** A Collection of Information Fragments brought together in a single unit by a single Identity. Its primary envisioned function would be to

bring together Information Fragments located in different locations within the same Information Entity and complementing each other in covering a particular topic.

- ***Relationship Qualifier*** **--** Any term or discussion used to describe the nature of a Fragment Association
- ***Descriptive Qualifier*** **–** Any term or discussion used to describe the content of an Information Fragment
- ***Creator-Specific Links*** **--** The links to be used in this study will be called “creator-specific links”. They will actually be “typed links” with the only required attribute being the creator.
- ***Collocation*** **–** The bringing together of two interrelated content items in the process of browsing or searching

1.7.1 The Relationship between Information Entity and Information Fragment

An Information Entity and an Information Fragment are both containers of some information content. An Information Entity is a container defined by the author and has some formal boundaries. An Information Fragment is defined by the consumer of the Information Entity and its boundaries are informal.

2 LITERATURE REVIEW

2.1 REVIEW AREAS

Frank Halasz, the great hypertext visionary in a 1988 article[10] reflecting on the lessons learned by his involvement in the development of the NoteCards hypertext system, lists seven issues:

1. Search and Query
2. Composites (dealing with relationships between notes and references)
3. Virtual structure for dealing with changing information
4. Computation over hypermedia networks
5. Versioning
6. Support for Collaborative Work
7. Extensibility and Tailorability

Although every single one of these issues has been tackled with, a lot of work still needs to be done for us to be in a position to say that we have really addressed them. This project attempts to tackle to some degree issues 1, 2, 3 and 5. Specifically the endeavor undertaken with this work has derived its inspiration from the study of literature covering the following main areas:

- Information Entity and Information Fragment (Halasz' issue #2)
- Information Fragment Association Theories and Reference Models (Halasz' issue #2)
- Systems Supporting Information Fragment Association (Halasz' issue #2 and #3)
- Versioning (Halasz' issue #3 and #5)
- Search Algorithms and Search Engines (Halasz's issue #1)

- Social Bookmarking (Halasz's issue #6)

Over the last 25 years or so we have seen remarkable developments in the world of information systems. The introduction of the World Wide Web and the exhilarating possibilities and daunting challenges it has brought forth have overshadowed some earlier work. During the quest for inspiration for this work, it seemed very attractive to look at some of these earlier ideas along with the new ones. The outcome was very rewarding.

2.2 INFORMATION ENTITY AND INFORMATION FRAGMENT

One of the major points that this work is dealing with is the fact that the current web environment is designed to handle Information Entities, which are entire documents, but it fails to adequately address Information Fragments within these Information Entities.

The following sections are dealing with the concepts of Information Entity and Information Fragment and are attempting to incorporate some related theory. In the case of Information Fragment, some examples of practical use of this concept by systems and specifications are also presented.

2.2.1 Information Entity

Miksa[11] calls the universe of human knowledge representations the *Information Entity Universe*. He states that "every object ever used intentionally by humans to record and convey knowledge is theoretically included in this universe". He further provides a definition of Information Entity by describing the objects comprising this universe as "every formally produced thing, but also every scrap of paper, every occasional note and memorandum, in short, every object ever imbued by humankind with a message and set aside in order to preserve and convey the message it contains”.

For Schamber[12] such an object would constitute a “document”. However, for her a document may also be a form of a meta- document constituted by a number of heterogeneous items linked together, or a conceptual display of a number of related items. In general, what is usually defined as a “Document” is the same or similar to what we consider here to be an “Information Entity”. Of course as always, we have to be careful when we make generalizations. The word “Document” has been given many different definitions and some of them bear a weaker relationship with “Information Entity” than others. Buckland[13] presents various definitions offered by prominent scholars in the field. He quotes Briet’s definition of a document as "any physical or symbolic sign, preserved or recorded, intended to represent, to reconstruct, or to demonstrate a physical or conceptual phenomenon". He also presents Ranganathan’s view of a document as an "embodied micro thought" on paper "or other material, fit for physical handling, transport across space, and preservation through time". These definitions have a weaker relationship with what is conceived of as “Information Entity” here. For example, Ranganathan’s “micro thought” used without any further qualification can easily apply to an “Information Fragment” as well as an “Information Entity”. Nevertheless, in most cases, the definitions of “Document” would encompass the same sets of human communication objects as the ones an Information Entity is meant to describe.

2.2.2 Information Fragment

Ted Nelson, the visionary often credited with coining the word “Hypertext” [14], and whose work has been very influential in the Hypertext community, has often been very critical of the current state of the World-Wide-Web[15]. In fact, his work and lifetime pursuits are diametrically opposite in approach to the rapid and unbridled growth of the World-Wide-Web. Is his vision completely utopian or simply ahead of its time? The answer can be found in the numerous instances in which ideas articulated by him have found some kind of implementation albeit under a different incarnation. This is why hypertext theory is still extremely useful as we deal with the realities of today's interconnected information resources. Nelson introduced the concept of *Transclusion*[15, 16] which is essentially the ability of a document to include by reference sections of other documents. *Transclusio*n is in essence a Referential Information Fragment Association as defined in 1.2.1 above. This means that although Nelson never used the term *Information Fragment*, the most important aspect of his thinking was the realization that referencing small fragments within a larger body of content was an important issue that system design had to address. His elaborate tumbler addressing mechanism[5] aimed at being able to define exact spans of content within the entire "Docuverse". Even though the concept of a span was far more complex and ambitious than the concept of Information Fragment, the two share one fundamental characteristic, the fact that they represent an addressable slice of content within a larger content. The difference lies more on the mechanics of implementation rather than the concept. Nelson presents this concept masterfully, but when it comes to usage scope, he concentrates on quotation which is just one particular use of Information Fragment. In other words he has described the Referential Information Fragment Association but not the Annotative or Relational Information Fragment Association. Transclusion was certainly the inspiration for Information Fragment but the scope of Information Fragment is broader.

Nelson's emphasis was on reuse, and he has certainly been very influential with that. Modern website development capitalizes on that concept. Although Nelson's vision has proven to be very complex, to some degree it has been realized with a variety of technologies. The most simple implementation is the use of includes in web pages or the IFRAME tag in HTML, although that implementation runs short of what Nelson had in mind..

2.2.2.1 Information Fragments and W3C Specifications

The need to address a fragment within a larger content body has been recognized by the numerous web specifications available today. Many of these specifications are dealing with the issue of granularity. Specific sections within XML documents can be referenced with a variety of processes as long as the XML document is tagged in a way allowing for these specific sections to be referenced unambiguously. Each specification elaborately addresses a need, and they are roughly of two types: those used for manipulating or extracting information out of XML documents and those used to describe documents or define the relationships between them. Examples of the first type are those dealing with document presentation (XSLT[17]) , formatting (XSL-FO[18]), data processing (XQuery[19]) etc. Examples of the second type include specifications such as XLink[8], which allows for the description of extended links between resources, RDF[20] which allows for the description of the resources themselves. One may call this latter type of techniques as “external” since they do not alter the documents they describe but rather maintain external references to them.

One of these “external” techniques is XPointer[6], a specification describing a way of providing addressing for fragment identifiers. This particular specification recognizes and attempts to address one very important issue: defining, describing and navigating to distinct information fragment within a larger content body. Although it only deals with addressing and not the broader issues and it provides a solution only for XML tagged documents, presupposing that the desired fragment has been pre-tagged and defined as an XML element, this specification highlights an aspect of the web environment that deserves further examination. The main difference between XPointer and the approach taken here is that XPointer simply concentrates on the technique of defining the boundaries of Information Fragments and does not deal with their function. In other words, XPointer can very well be used with appropriate extensions as part of an overall framework providing Information Fragment Association functionality, but it does not constitute an integrated Information Fragment Association framework by itself.

2.3 INFORMATION FRAGMENT ASSOCIATION THEORIES AND REFERENCE MODELS

As the term implies a *Link* is an instrument used to connect two items. We have come to be very familiar with the ubiquitous HTML links, but those links are only a small simplified subset of the entire spectrum of links devised by traditional hypermedia systems. In his survey of Hypertext systems, Conklin[21] has outlined the several functions performed by links in hypertext systems:

- They can connect the reference to another document to the document itself.
- They can connect a comment or annotation to the text about which it is written.
- They can indicate that some text is a subsection of some other piece of text, or other kinds of organizational information (i.e. the link between a table of contents entry and its section).
- They can connect two successive pieces of text, or a piece of text and all of its immediate successors.
- They can connect entries in a table or figure to longer descriptions, or even sub-tables or figures.

In addition to these functions, hypertext theorists have introduced the concept of *typed links*. Trigg[22] provides an exhaustive outline of typed links, which incorporate another characteristic a link may have. These types are used to describe the relationship that these links define between the two linked items. Examples of some of the links proposed by Trigg are *refutation, support, solution, explanation, correction, update, continuation* etc. For example, in a specific passage in an article an author B refutes a specific passage of an article by author A. The link between these two passages (which we call Information Fragments) is a *refutation link*.

Yet another characteristic of more advanced links is directionality. Ordinary HTML links are unidirectional, allowing the user to jump from one place to the other. In more full-featured hypertext systems links are bidirectional, providing more navigation options.

Links can therefore incorporate far more functionality than ordinary web links. The concept of Information Fragment Association is based on some of these additional functionalities. In order to get a better picture of the nature of Information Fragment Association, we need to examine some attempts at shedding some light on this complex concept. Links can be categorized in many ways depending on their functionality or structure. Even though many categorizations such as the one introduced by DeRose[23] and listed below are possible, most scholars agree with Davis[24] that there are two primary kinds of links: *Embedded* and *External*. *Embedded links* are those containing the linking information embedded within one of the two documents being linked. *External links* employ some external storage component where the linking information is being recorded.

The following sections are dealing with some theories regarding linking. First, a general taxonomy of links is presented in order to offer a broader and more comprehensive view of the topic. After that, we focus on the more specific issue of Information Fragment linking, presenting previous approaches. Finally, since we are committed to supporting a simple form of typed links, a discussion of the concept is provided which includes a reference to the way typed links can be used as descriptive attributes, the conditions for their use and their differences from what has been traditionally proposed.

2.3.1 Taxonomy of Links

DeRose[23] has provided a comprehensive taxonomy of links to help us distinguish between the variety of linking usages shown in Figure 5. According to DeRose, "the destination of an intensional link is defined by some function that finds the desired ends, rather than being a list of known ends". The link is thus inferred from the structure rather than being defined. The fact that intensional links are predictable allows them to lend themselves to be used by automated systems. However, they are not as sophisticated as the extensional links which have the capability of being defined as links from any location (or fragment) to any location (or fragment). These are the links normally created by humans after some content processing, or by very advanced automated systems. The Information Fragment Associations being envisioned here are “Extensional Links” of any possible subcategory, but they are primarily “Associative Links”.

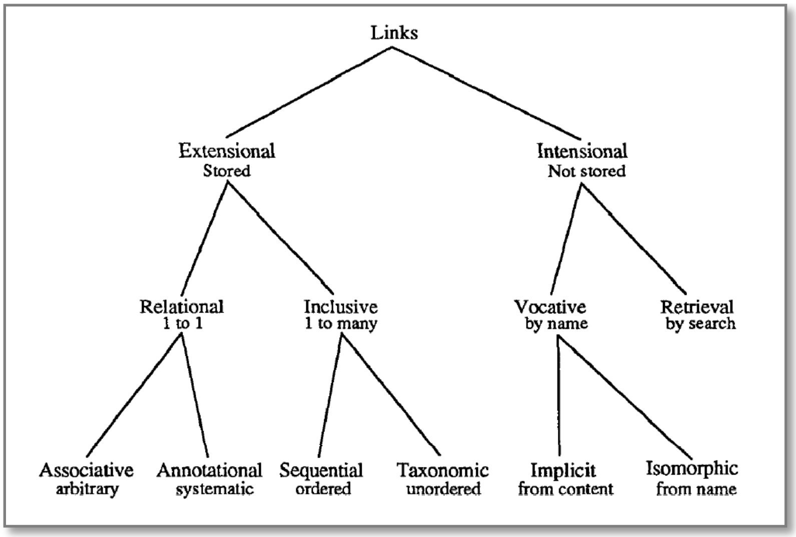


Figure 5 - DeRose's Links[1]

2.3.2 Linking to an Information Fragment

The concept of linking to an information fragment is not novel. The *ReferenceSpecifier* defined by Grønbæk and Trigg [25] in their extension to the Dexter Hypertext Reference Model [26], was based on a similar concept. *ReferenceSpecifier* was introduced as a specific type of *LocationSpecifier*. *LocationSpecifier* was defined as a model of the capability of a hypermedia system "to specify a location in a body of electronic material". It specifies more than just a location though. It contains an attribute called the "within-component structure descriptor" which can specify among others position, span, video frame, chapter, paragraph etc. *ReferenceSpecifier* had an additional attribute for specifying the parent *component*. The *component* was the "wrapping" for a document. In other words, *ReferenceSpecifier* theoretically as a model contains the attributes allowing it to establish an Information Fragment Association. However, this has

[1] From [23] S. J. DeRose, "Expanding the notion of links," in *Proceedings of the second annual ACM conference on Hypertext*, Pittsburgh, Pennsylvania, United States, 1989, pp. 249-257. http://doi.acm.org/10.1145/74224.74245.

some inadequacies when it comes to adapting this within today's environment. *ReferenceSpecifier* evolved out of hypermedia systems that were for the most part within the control of their designers. Even though the framework that Grønbæk and Trigg presented was part of the transition to open hypermedia systems, the functionality which would have allowed Information Fragment Associations to be created could not be applied to the Web. Even though the Web and its first search engines were already in existence at the time *ReferenceSpecifier* appeared, the only relationship between that model and the Web involved the consideration that the standard embedded URL could be expressed using this model as a simple *LocationSpecifier*. This means that it was never intended to be a complement of the Web. The Web was considered a simple implementation of this model, lacking most of its functionality.

However, this concept did appear as a desirable complement to the Web infrastructure when issues like referential link integrity are considered. Davis [24] outlines all issues and possible approaches to the *content reference* issue. He defines *content reference* as some kind of pointer to the destination of a link and he tackles the *content reference* problem which occurs when a document is edited but the references into the document are not updated causing the link to point to the wrong point in the document. He does not attempt to present any solutions. Instead, he presents several possible options, some of them better than others, but with the awareness than none of these options provides a perfect solutions. In the interest of being comprehensive, the first option he presents is doing nothing about the problem, just making sure that the users are aware of this problem. The second option is to attempt to express links declaratively (as some kind of queries). This of course, does not guarantee their accuracy. His third option is realistic and practical. He suggests leaving the burden of fixing the links to the users but at least providing them with some tools to make this easy. The forth option is the slightly utopian solution providing accurate just-in-time automated adjustment of a link when the content changes. His last option is a solution requiring all hypertext objects, including the node content to be part of the hyperbase which maintains the integrity of the links. Davis concludes that none of these options are ideal and that all of them compromise the integrity of open hypermedia systems. He suggests that depending of the scope of an application different degrees of implementations of these options may be necessary. During the discussion he mentions the versioning option, a variant of which is the approach taken here, imperfect as it may be.

The XLink[8] and XPointer[6] specifications are also comprehensive solutions aimed at utilizing the current Web infrastructure to accomplish similar results. However, they are not designed to handle existing information. They are a well-structured solution presented as the prescribed way for offering external link capabilities and content referencing. This means that it relies on the authors of the documents to provide some structure. We can expect to see more and more structured documents in the future, but currently there are a lot of unstructured but valuable documents on the Web, for which XLink and XPointer may not be the best solution.

2.3.3 Typed Links

Introduced by Randall Trigg[22], this concept allows for assigning attributes to describe the relationship between linked items. Trigg divides typed links into two categories: *Normal Links* and *Commentary Links*. *Normal Links* are the links describing the relationship between two nodes. For example a *Support* link, which is a Normal typed link, can specify that A is a node containing information supporting B. The implication here is that A and B can be independent of each other, and they are simply associated with each other with a link describing a relationship between them. Trigg's *Commentary Links* are links between a node and statements about that node. This implies that these statements are dependent on the original node and are applicable only to that node. They are therefore similar to what we call Annotative Information Fragment Associations. Although the distinction between *Normal Links* and *Commentary Links* is very helpful, it is too general and it fails to take into consideration the most important information about a link, which is the creator of the link. To demonstrate this deficiency, we can consider the following example:

a is the creator of A

b is the creator of B

The statement "A supports B" is insufficient unless we know who is making the statement because we may trust the first scenario more than the last two:

a says that A supports B

b says that A supports B

c says that A supports B

Therefore it can be established that that at least as far as an Information Fragment Association solution is concerned typed links are only part of the answer. It is important to be able to have a structure allowing for these descriptive attributes to be attached to a link between two Information Fragments. In order to provide a good Information Fragment Association solution it is necessary to give users the capability of attaching descriptive attributes but not require it. Complexity may be perceived by users as an impediment therefore, populating these typed links with explicit attributes needs to be promoted with a simple interface which would encourage users to select some terms describing the nature of these Information Fragment Associations. The one absolutely required attribute would be the creator. The above example demonstrates its importance and in a solution in which Identity tracking is an integrated component this issue is easily and seamlessly being addressed. It is unclear whether Trigg's framework had a provision for identifying the creator. The concept, however, has appeared prominently in Neto, Pimentel and Truong's work[27].

According to Trigg[22], typed links are explicit relationships between two nodes. The function of a typed link is not to provide the information per se but rather, as Kopac[28] says "what the link should be doing is telling us *how* the content of the destination node (modifying information) is intended to alter our understanding or interpretation of the source node (the object information)." The lack of typed links on the web has been bemoaned ever since it started gaining popularity[29]. Typed links were envisioned and well-articulated within the Hypertext community[26], but were never successfully applied broadly to web applications. As a consequence of the lack of typed links, the vast majority of web resources provide no qualification for the linked site. Annotation and Typed links provide an additional dimension to information retrieval. An information entity has several attributes or associations with other entities such as individuals or other information entities. A traditional information entity has basic associations such as its association with its creator and its "description", an example being the Dublin Core basic elements. Typed links can only be useful if there is a certain formality in their use, i.e. using consistent labels. This is why a successful implementation involving end-users has to contain only a very limited number of link types.

2.3.4 Versioning

"Version" is a broad term used in everyday life in a variety of contexts and bears a variety of connotations. For this reason Conardi and Westfechtel [30] divide versions into three types, depending on the intention of its use. They call *revision* a version intended to supersede its predecessor, *variant* a version intended to coexist with other versions, and *workspace* a version belonging to a set of versions maintained to support collaboration. Whitehead [31] goes further to distinguish between a *variant*, a *rendition* and an *alternate version*. He defines a variant as "a snapshot of an instant in the evolution of a work or entity, whose differences from other snapshots can be precisely specified, or parameterized, in a form other than a delta" and a rendition as a mechanically derivable variant, such as the PDF version of a Word document. As an alternate version he defines a variant that is "sufficiently different from other instances of a work or entity that causes it to have … a change in identity". He provides several useful examples elucidating these sometimes misunderstood terms.

In spite of the fact that these terms are often used loosely and sometimes interchangeably, the concept of a version and specifically the concept of a revision, which is what we are mainly interested here, is rather simple and easily understood. However, the processes and techniques involved in maintaining versions can be fairly complex and the various existing version models differ substantially. According to Conardi and Westfechtel [30], a "version model defines the items to be versioned, the common properties shared by all versions of an item, and the deltas, that is the differences between them". They go on to define a version as a representation of a state of an evolving item. The item possesses some unchanged properties that are common in all versions of the item. These properties are called "invariants", and at the minimum they include a unique identifier for the item, called OID. Each version is seen as having a unique version identifier, called VID. Conardi and Westfechtel then define a versioned item as a container for a set V of versions and they distinguish two types of versioning, *extensional versioning* and *intensional versioning*. Extensional versioning defines the version set V by enumerating its members: $V = \{v_1, \dots v_n\}$while intensional versioning employs a predicate to define the version set: $V = \{v|c(v)\}$ where c are the constraints that must be satisfied by all members of V. In intensional versioning, versions are constructed as a result of a condition, such as a specific query for some required attributes.

Extensional versioning is more often called state-based versioning and intensional versioning is called task-based or change-based versioning. Haake and Hicks [32] draw the attention to the fact that the basic state-based versioning approach does not allow for tracking coordinated sets of changes. They present the example of a bug-fix in a software development project. The fix may involve changes in several files, yet it constitutes a single change. The advantage of task-based versioning is in its ability to keep information regarding any existing complex relationships between distinct files, thus becoming very suitable for a hypermedia environments, as Vitali [33] points out.

2.3.5 Dexter Hypertext Reference Model

During the halcyon days of hypertext system development in the 1980s, developers came to a realization that it was necessary to establish a common ground for all of these hypertext systems. The goals and functionality of these systems were within the same realm, but the techniques and terminology they were using were different. In an effort to create a common reference model to help both with the comparison of the various functionalities offered by these systems and the establishment of an interoperability platform, several developers put together the Dexter Hypertext Reference Model[26] in 1998.

The Dexter Hypertext Reference Model represents the first comprehensive attempt to standardize Information Fragment Association functionality. It was successful in providing a set of common syntax and semantics for expressing Information Fragment Associations which it called "Span-to-span" links. The Model recognized that such functionality depends "on a mechanism specifying substructure within components". Information Fragments are represented in the Model by *Anchors*. An Anchor contains both the information needed to retrieve the Information Fragment and a unique identifier for it. A link simply associated two anchor IDs. Even though this Model is based on Open Hypermedia Systems some of which predate the web, and itself predated the significant advances of Web Services of the last few years, it lays the foundation for future frameworks evolving around Information Fragment Associations.

One of the goals of the Dexter Hypertext Reference Model was to resolve terminology conflicts. For this reason it uses the term *component* to describe the model abstraction containing data. Every component has a unique identifier. The model defines three layers for a hypertext system, shown in Figure 6; the *run-time layer*, the *storage layer* and the *within-component layer*.

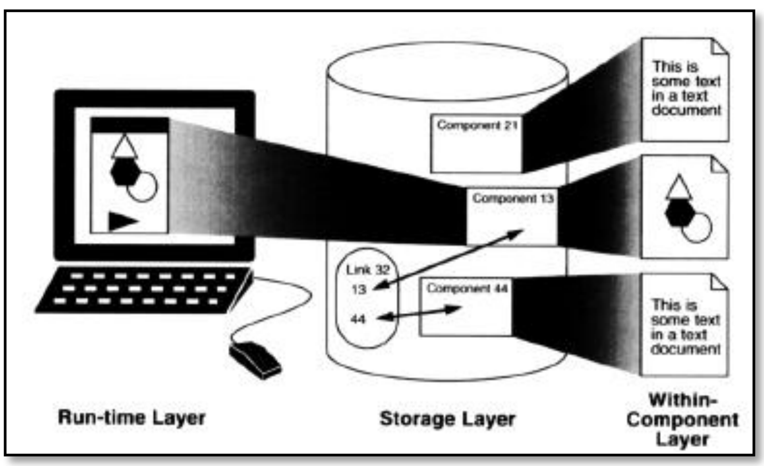


Figure 6 - The Three Layers of the Dexter Model as Embedded in an Actual System[2]

The node/link structure is defined for and stored in the storage layer. It defines the structure of the links but not the content or format of the components. The within-component layer defines

[2] From [26] F. Halasz and M. Schwartz, "The Dexter hypertext reference model," *Communications of the ACM*, vol. 37, pp. 30 - 39, 1994. http://doi.acm.org/10.1145/175235.175237.

the structure within the components. The model does not define any of these particular structures leaving them to the individual system design or to other more specific models. The run-time component uses information encoded in the links for the presentation of the content. The model defines three terms involved in the linking process, the *link*, the *anchor* and the *specifier*. The function of these terms and the relationship between them is demonstrated in Figure 7. Individual functions were defined each one of them conceptually handling a specific functionality of a hypertext system.

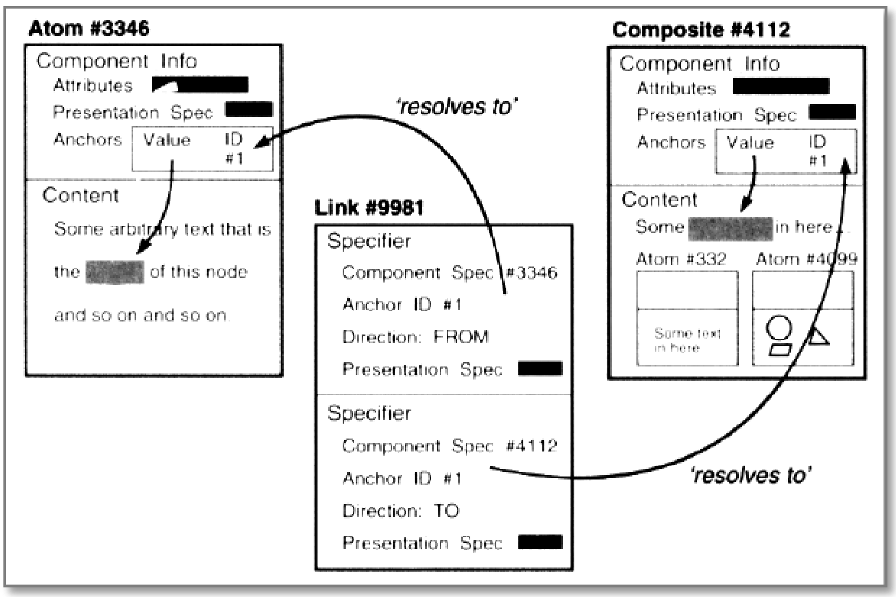


Figure 7 - The Dexter Storage Layer[3]

The Dexter Hypertext Reference Model did not offer a specific solution but it provided a way of referring to the feature sets of different systems. None of the systems of its time supported all of the functionality specified in the model, but it offered a way of determining the strength of a system by the number and kind of Dexter functionalities it supported.

[3] From [26] Ibid.

2.3.6 Open Hypermedia Protocol and the OHSWG Unified Data Model

The Open Hypermedia Protocol[34] is the product of the work of the Open Hypermedia Systems Working Group (OHSWG). This group was formed in 1995 to address issues of interoperability between OHSs. It met for several years during the International Workshop on Open Hypermedia Systems. The framework was designed to support navigational hypertext, spatial hypertext, taxonomic hypertext, hypertext by transclusion as embodied by Xanadu[5] etc.

The Open Hypermedia Protocol as originally designed was defining interoperability between hypermedia services and client applications. It was soon realized, however, that this was not sufficient to achieve real-life interoperability. Therefore, what was originally defined as the Open Hypermedia Protocol was called the *Content Handler Interface* (CHI), and a new interface was added to the mix called *Hypermedia Database Interface* (HDBI). The function of HDBI is to enable interoperability between hypermedia services and hypermedia databases. The two interfaces together are shown in Figure 8.

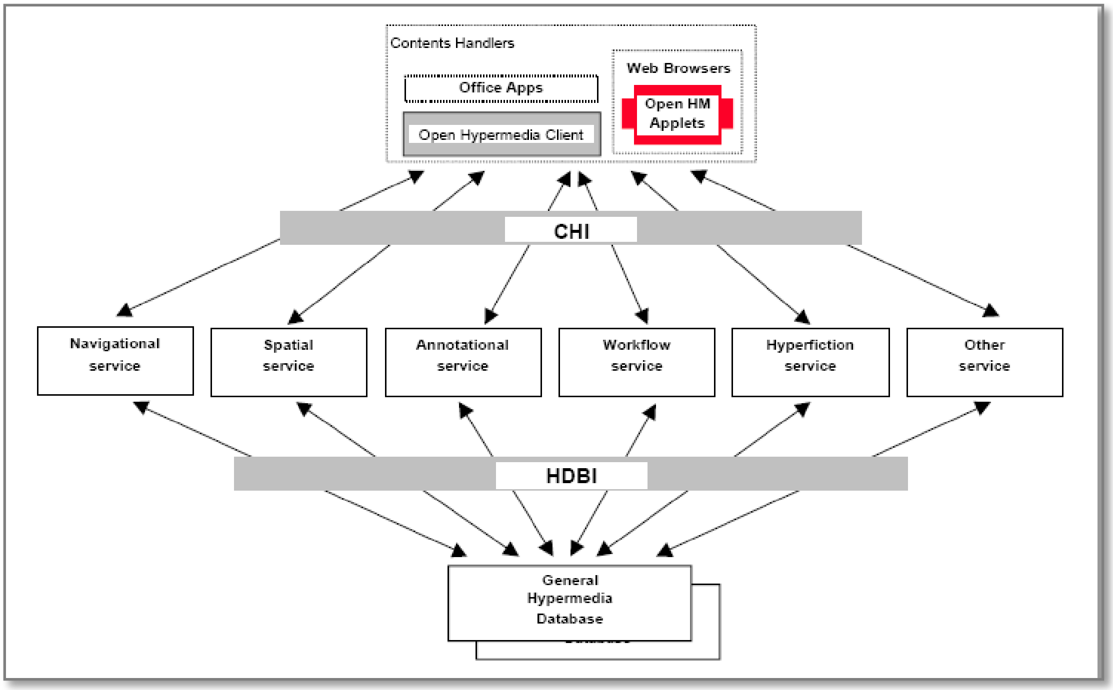


Figure 8 - The OHSWG Unified Data Model[4]

OHSWG put together a detailed data model. One of the key elements of this model is *HMObject* which corresponds to the component in the Dexter Hypertext Reference Model, but it is wider in scope. Subclasses of *HMObjects* are *Context, Link, Endpoint, Anchor, Node, Computation* and *PSpec*.

OHSWG carried the concepts of the Dexter Hypertext Reference Model into a web-aware environment. In spite of its complexity, it is a good model and it still maintains its support for Information Fragment Association. It simply adds an additional abstraction to the mix by introducing *EndPoint*, which consists of an *Anchor* identifier and a *Link* identifier. There is only one *Anchor* and only one *Link* in an *EndPoint*, but the same *Link* and same *Anchor* may appear in multiple *EndPoints*.

---

[4] From [34] S. Reich, U. K. Wiil, P. J. Nuernberg, H. C. Davis, K. Groenbaek, K. M. Anderson, D. E. Millard, J. M. Haake, and K. Groenbaek, "Addressing Interoperability in Open Hypermedia: the Design of the Open Hypermedia Protocol," *The New Review of Hypermedia and Multimedia* vol. 5, pp. pp. 207-248, 2000.

2.3.7 Spatial Hypertext

Spatial hypertext[35, 36] differs from navigational hypertext in that it does not use explicit links to express content relationships. Instead, it uses maps of content structure. As Cathy Marshall puts it "spatial hypertext is ... a way to take advantage of human perceptual abilities in hypertext navigation and to provide users with a fairly intuitive medium through which they may express new structures and manipulate existing structures".

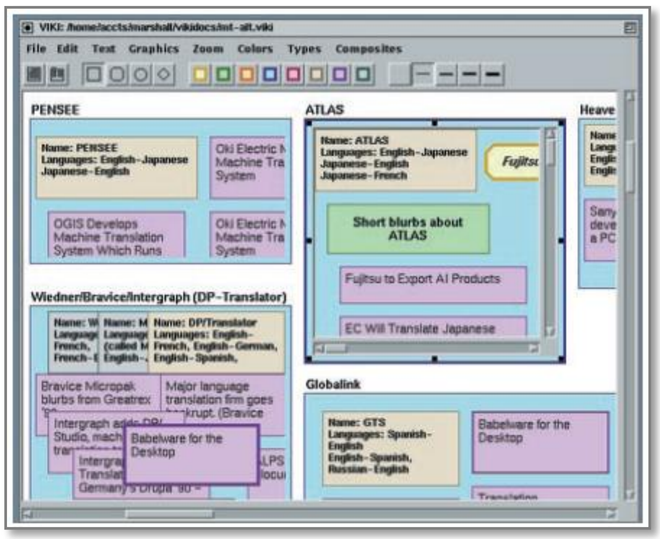


Figure 9 - Collections at Different Stages of Organization[5]

One of the most important characteristics of spatial hypertext is that its structure can be implicit and informal. Structure can start being built in the mind of the user as the user attempts to make sense of content by moving symbols around in a manner similar to the way users handle pieces of paper containing related information.

Spatial hypertext tools allow users to move content around and record the action, thus providing their interpretations of the information they consume. In a system like VIKI

[5] From [36] C. C. Marshall and I. Frank M. Shipman, "Spatial hypertext: designing for change," *Communications of the ACM*, vol. 38, pp. 88 - 97, 1995. http://doi.acm.org/10.1145/208344.208350.

collections can be organized by the user. Figure 9 shows five collections, some of which are more organized than the others.

When it comes to Information Fragment Association, Spatial Hypertext presents a unique way for its construction. Information Fragments can be associated together with visual techniques such as position and color. For some users this may be a preferable and more productive way of establishing Information Fragment Associations. As long as the visual representations can be formalized and saved using common schemata in a common registry, spatial Information Fragment Association tools can co-exist with more traditional Information Fragment Association tools. The flip side of this can be conceived for browsing purposes. Information Fragment Associations created with traditional tools would potentially be browsed using spatial tools, as long as the spatial tools know how to represent every possible type of relationship.

2.4 SYSTEMS SUPPORTING INFORMATION FRAGMENT ASSOCIATION

The term "Registry" has been used loosely over the years, but in general most would agree that a Registry is a mission-specific compilation of information aimed primarily at providing a consistent central reference for the purpose of enhancing some functionality usually contained outside the Registry. A Registry is mission-specific in the sense that it has to accommodate the specific needs it has been compiled to meet. Even the most universal registry of everything known to mankind would probably not help a small hardware store owner locate items on the store's shelves.

Registries are often part of larger systems, usually systems providing some kind of service (with the term "system" used in a sense broader than that of an information system). They can be either absolutely essential to that system or essential to just one piece of functionality within the system. When a student registers for a course, most of the information regarding the course and the student's progress is registered in a registry, but the classes take place outside the registry. In this particular case, the registry is very important but not absolutely essential. An unregistered student auditing a course would still benefit equally from the course as a registered student. The lack of a registry entry would simply deprive the student from receiving academic credit. Other registries however, are tightly woven into a system and the system cannot function without them. An example of such a registry is the one used by the Windows operating system.

Open Hypermedia Systems[37] are systems designed to operate between application interfaces and the resources that these applications are utilizing. The intermediate nature of these systems allows for the creation of external links establishing relationships between resources without having to alter these resources. Some of them were too ambitious to be successful, a fact which served as a valuable lesson during the undertaking and design of this project.

A variety of different approaches and applications of the principles of OHS have been introduced. Systems like HyperDisco[38] and Microcosm[39] support the integration of a variety of distinct third-party applications.

Some of these systems are being presented here. They are by no means the only ones, but they are representative of their respective categories.

2.4.1 Traditional Open Hypermedia Systems

2.4.1.1 Microcosm

Microcosm was an open hypertext system developed at the University of Southampton[40]. It consisted of several viewers, each one supporting a different information format. Microcosm could use any viewer, as long as it was capable of incorporating a menu for follow-link, compute-link, start-link, and end-link, sending the user selection to Microcosm and allowing the user to specify the buttons (anchors) for the links.

The viewers, as Figure 10 demonstrates, would interact with Microcosm which would send messages through several filters. A filter could respond to messages, pass them on or block them. The most important filters were the linkbases which responded by finding the link information. There was no internal markup of documents in Microcosm. Links resided in external linkbases, and the viewers communicated with the linkbases to find and display the links.

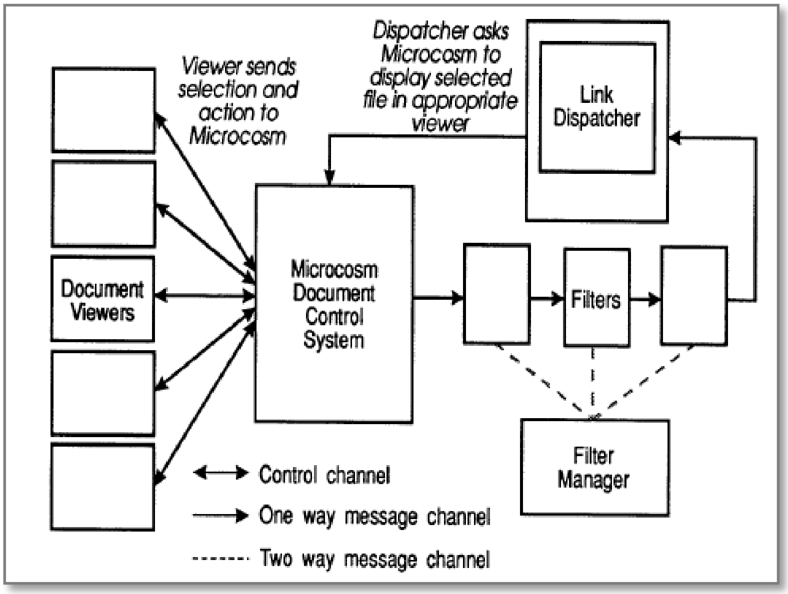


Figure 10 - Microcosm[6]

There were different kinds of links in Microcosm[41]:

- *Specific* Links - from a particular object at a specific point in a source document to a particular object in a destination document.
- *Local* Links - from a particular object at any point in a specific document to a particular object in a destination document.
- *Generic* Links - from a particular object at any position in any document to a particular object in a destination document.

Microcosm also supported Computed Links, which were essentially links generated by various types of queries and not manually created by an individual.

Microcosm's Generic Links make indiscriminate navigation inevitable by enabling a situation such as the one described by Fountain[42] in which if a generic link to a word has been created, any new document containing that word would contain that link as well. The generic

[6] From [41] H. Davis, W. Hall, I. Heath, G. Hill, and R. Wilkins, "Towards an integrated information environment with open hypermedia systems," in *Proceedings of the ACM conference on Hypertext* Milan, Italy: ACM Press, 1992. http://doi.acm.org/10.1145/168466.168522.

link approach arbitrarily interjects links within contexts which may very well be completely unrelated.

Information Fragment Association as conceived of in this Study aims at enriching the information universe with unique relationships between Information Fragments. The goal of these relationships is both to augment the content of the Information Fragments *and* to qualify them, thus maintaining precision as it improves recall. In other words, Information Fragment Associations are not conceived as instruments of indiscriminate navigation. The approach taken by this Study makes multiple associations to an Information Fragment possible, but it ensures that each association is made by users with conscious deliberation.

2.4.1.2 DeVise Hypermedia (DHM)

DeVise Hypermedia[43] was a system developed by Kaj Grønbæk and Randall Trigg as an application of the Dexter Hypertext Reference Model. It also added functionality to the Model as it provided support for long-term transactions, locking and event notification. Event notification is one of the features envisioned by Halasz[10]. It involves the ability of users to subscribe to a feature allowing them to be notified of events (such as changes to a document) occurring and involving shared content. Object access and locking is an indispensible feature that had to be emphasized in these early systems and it was well handled by DHM using the LocSpec parameter. The system was introduced around the time the Web had just started becoming popular. A few years later, with the Web browsers maturing with the introduction of Java applet and ActiveX capabilities, DHM was adapted for the Web[44]. This web adaptation of DHM inserts external links to a web page by feeding an applet with LocSpec information[45]. Figure 11 shows an example of link information as it is passed to the applet the function of which is to apply the link to the page. This particular example shows two links. This information consists of the URL of the document, the text to be selected, the context of this text, the position of the selected text, the offset of the selected text and the last modification date of the document.

```
(1,2, ("http://www.authors.jp/oe.html""Oe""Kenzaburo
Oe""70""2"
"850650404"), (''http://www.daimi.aau.dk/japan.html"Ke
nzaburo Oe" "Yukio Mishima, Kenzaburo Oe and Musashi
Miyamoto are my""230" "12""852100133"))
```

Figure 81
A response generated by FollowLink.

Figure 11 - DHM FollowLink Response with LocSpec[7]

The very structure of DHM's links points to the fact that its main concern is the ability to apply links externally. This structure does not preclude the association of sizable Information Fragments, but it is not very suitable for them. The current approach strives to put more emphasis on the content of Information Fragments.

2.4.1.3 Chimera

Chimera was an open hypermedia system developed at the University of California, Irvine, with an emphasis on the modeling of heterogeneous software engineering environments. It is a client-server system with the server providing external link capabilities to the clients. Multiple users on different machines could access a hyperweb from a dynamically changing set of viewers. Hypertext events propagate from the one viewer to the other via the server. Client applications could be written in any language accessing the server with the particular API provided for that language.

[7] From [45] K. Grønbæk and R. H. Trigg, *From Web to Workplace : Designing Open Hypermedia Systems*. Cambridge, Mass.: MIT Press, 1999.

Since Chimera is geared to support software development environments it incorporates several *Concepts* (Figure 12) into which tools can map. These concepts are *Objects, Viewers, Views, Anchors, Links, Attribute-Value Pairs and Hyperwebs. Objects* are named, persistent entities whose internal structure is unknown and irrelevant to Chimera. *Viewers* are named active entities that display objects. *Views* denote a pair (ν,o) where ν is a viewer for an object. *Anchors* are portions of a view defined as items of interest. A *Link* is a set of anchors. Links relate portions of views.

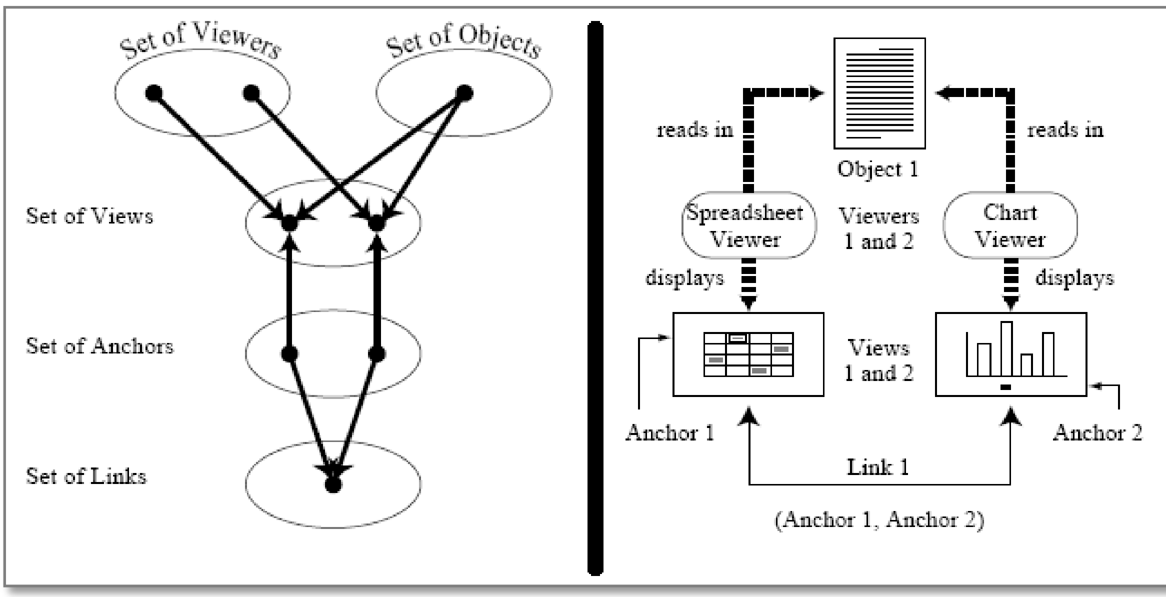


Figure 12 - Chimera Concept Example. Chimera's hypertext concepts are shown on the left. Two viewers are combined with one object to produce two distinct views. An anchor is added to each view and then combined in one link. On the right, an example hyperweb presents a data file (stored as a file in the operating system) being displayed by two different viewers. One viewer displays the data as a spreadsheet, creating a spreadsheet view of the data file. The other viewer displays the data as a chart, creating a chart view of the same data. The two distinct anchors are indicated by a black box in the spreadsheet, and a black underline in the chart. The anchors are stored in the Chimera database, not in the data file. The two anchors are members of the link. Attribute-value pairs are not indicated to avoid visual clutter. The Chimera architecture consisted of the Chimera Server, Clients, Process Invoker and External tools.[8]

[8] From [46] K. M. Anderson, R. N. Taylor, and J. E. James Whitehead, "Chimera: hypertext for heterogeneous software environments," in *Proceedings of the 1994 ACM European conference on Hypermedia technology* Edinburgh, Scotland: ACM Press, 1994.

Links can link to other links. An *Attribute-Value Pair* consists of two associated strings where one string contains the attribute's name, the other its value, providing run-time semantics, such as the creator etc. A *Hyperweb* is a collection of objects, viewers, views, anchors, and links along with their attributes.

Chimera's architecture, as shown in Figure 13 consists of the *Chimera Server*, the *Process Invoker* and the *Chimera Client*. The *Chimera Server* was designed to ensure the persistence of a hyperweb by storing the hypertext *Concepts* and to receive, route, and generate hypertext events. The *Process Invoker* was designed to be used whenever a hypertext event had to be sent to a viewer which was not running at that moment. The *Chimera Server* would send the *Process Invoker* the information about the specific viewer and the *Process Invoker* would launch it. The *Chimera Client* encompassed the various clients used. Chimera also supported the use of any External Systems the Viewers were using.

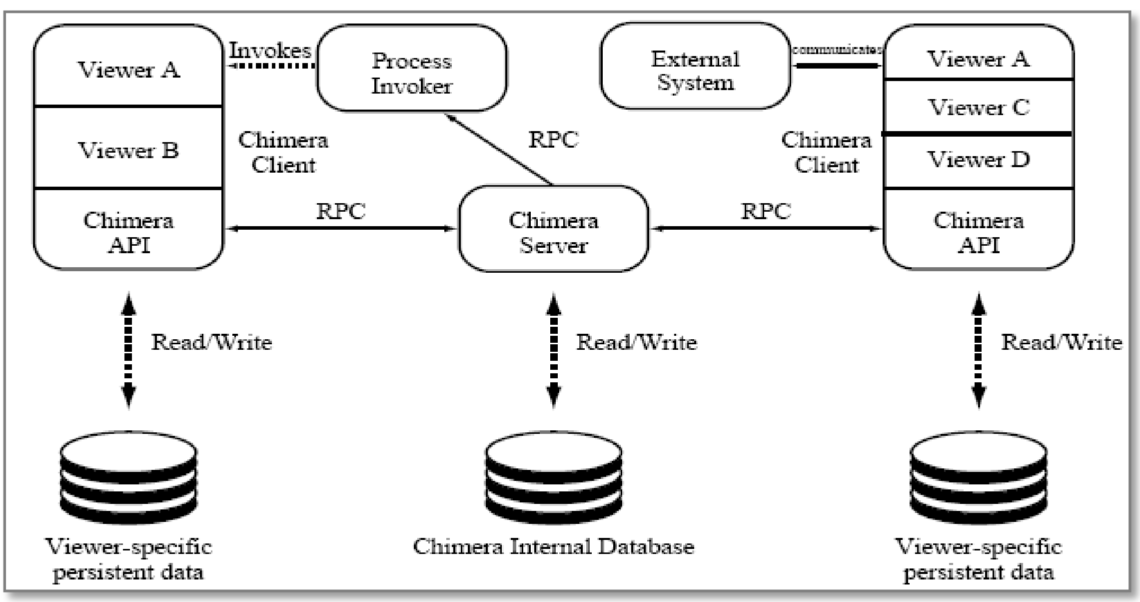


Figure 13 - Chimera's Architecture[9]

---

[9] From [46] Ibid.

In 1997 Chimera was reconfigured to work with the Web. While many systems prior to the domination of the WWW were at the time in the process of attempting to be integrated into the WWW, Kenneth Anderson argued for taking the integration into another direction i.e., integrating the WWW into an Open Hypermedia System[47]. The approach used the Chimera Presence CGI script and involved changing the reference in the anchors of the links in a retrieved web page and inserting an applet tag at the end of a page (see Figure 14). The applet was downloaded into the web browser and interacted with the hyparweb manager and the client server.

```
Before:

<HTML><HEAD><TITLE>Example One</TITLE></HEAD>
<BODY>Please visit the <A HREF="http://www.ics.uci.edu/pub/chimera/">Chimera web page</A>.
</BODY></HTML>

After:

<HTML><HEAD><TITLE>Example One</TITLE></HEAD>
<BODY>Please visit the <A HREF= "http://www.some.domain/chimera/chimeraPresence?http://
www.ics.uci.edu/pub/chimera/">
Chimera web page</A>.
<applet CODE="presence.class" CODEBASE="/applets/presence" WIDTH=500 HEIGHT=250>
<param name="page" value="http://www.ics.uci.edu/pub/chimera/">
<param name="user" value="kanderso">
The Chimera Presence Applet</applet>
</BODY></HTML>
```

Figure 14 - The Chimera Presence Script[10]

Chimera focuses on the presentation of Information Fragments through different viewers and less on their association. Moreover, in spite of the obvious benefits of Chimera's approach involving multiple presentations of an Information Fragment by multiple viewers, the current approach defines different presentations as different Information Fragments. The formatting of an Information Entity and the presentation of the context within which an Information Fragment occurs is sometimes crucial to its interpretation.

---

[10] From [47] K. M. Anderson, "Integrating open hypermedia systems with the World Wide Web," in *Proceedings of the eighth ACM conference on Hypertext* Southampton, United Kingdom: ACM Press, 1997. http://doi.acm.org/10.1145/267437.267454.

2.4.1.4 HyperDisco

HyperDisco is an open hypermedia system designed to integrate and extend other tools. It provides a very flexible model for integration, allowing each tool to select which hypermedia services to use. HyperDisco has two layers of hypermedia functionality, the *integration model* and the *data model* (see Figure 15). Both of these models have some build in classes. The *integration model* layer has basic linking services (anchors and links). The *data model* layer has basic hypermedia storage services for hypermedia objects, such as nodes, composites, links and anchors. The *tool intergators* reside in the in the *integration model* layer and the *hyperbase management systems* (HBMS) reside in the *data model* layer. This example demonstrates HyperDisco's ability to handle diverse tools performing diverse functions through these integrators which interact with the data stored in the *hyperbase management system.*

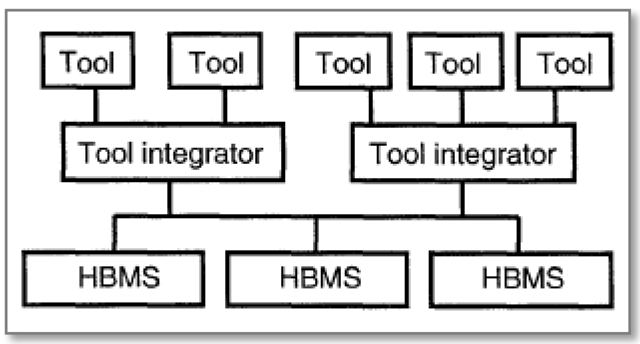


Figure 15 - The HyperDisco Architecture[11]

2.4.1.5 Hyper-G and HyperWave

Hyper-G[48] provides an excellent example of the power of a well-designed scalable infrastructure accompanying the content of the World Wide Web and making up for some of its

[11] From [38] U. K. Wiil and J. J. Leggett, "The HyperDisco approach to open hypermedia systems," in *Proceedings of the the seventh ACM conference on Hypertext*, Bethesda, Maryland, United States, 1996, pp. 140-148. http://doi.acm.org/10.1145/234828.234842.

inadequacies. Hyper-G was developed at the Gratz University of Technology in Austria. Concepts such as the use of replication for scalability purposes and the ability to link to a destination anchor within another document are powerful contributions of this system which was perhaps ahead of its time. Hyper-G also provides support for multiple protocols and languages. Documents can be arranged in collections which themselves may belong to other collections and they can reside on different Hyper-G servers. A collection presents as a single unit physically disparate resources.

A Hyper-G server responds to HTTP requests and through CGI it returns pages with some functionality such as Menus, Collections and attributes added to them. One of the menu items is search, which brings up a page allowing the user to submit a search request that could be scoped to the current collection or subcollections.

Hyper-G eventually evolved into a commercial document management product called HyperWave[49]. HyperWave uses an interesting approach[50]. It stores the documents and the links separately. When a document is inserted into the database, the hyperlinks embedded in the document are removed from the document and stored as individual objects. The object record contains all of the information necessary for describing and recreating the link. When the document is retrieved, it is retrieved as a plain document and the hyperlinks are added to it from the object records.

One of Hyper-G's main features is its support and integration of different protocols, such as HTTP, Gopher FTP etc. This specific functionality is obviously outdated today, both because some of these protocols are no longer in use, and because interoperability can be accomplished with Web Services using HTTP in conjunction with any other protocol that may be used in a given application. However, it presents a model for heterogeneity which can be emulated in the continuing struggle for bridging the various data formats available today. Hyper-G does not seem to handle Information Fragment Association very well. The approach used in this dissertation tackles this issue much better, but as it evolves in the future it can benefit from Hyper-G's vision.

2.4.1.6 Distributed Link Service (DLS)

Distributed Link Service (DLS)[51] evolved out of the Microcosm hypertext system and it attempted to extend the functionality of World Wide Web links.

The client is designed as a set of menus on the top of any application used as a document viewer. It allowed the user to select a section of the viewed document by submitting a request to the server (see Figure 16), and it is able to get a listing of the links available for that section (see Figure 17).

The server component is a pseudo-server which acts like an ordinary web server interacting with the web browser. The difference is that it does not store any documents. It allows for creating and editing links which are stored in several link databases. These databases keep information such as the source and destination attributes of the link, the type of the link, its creation time and a link description.

DLS provides the capability of passing queries from a proxy server to a link server and then to another link server if necessary (see Figure 18). Linkbase data could also be downloaded and cached instead of being visited every time.

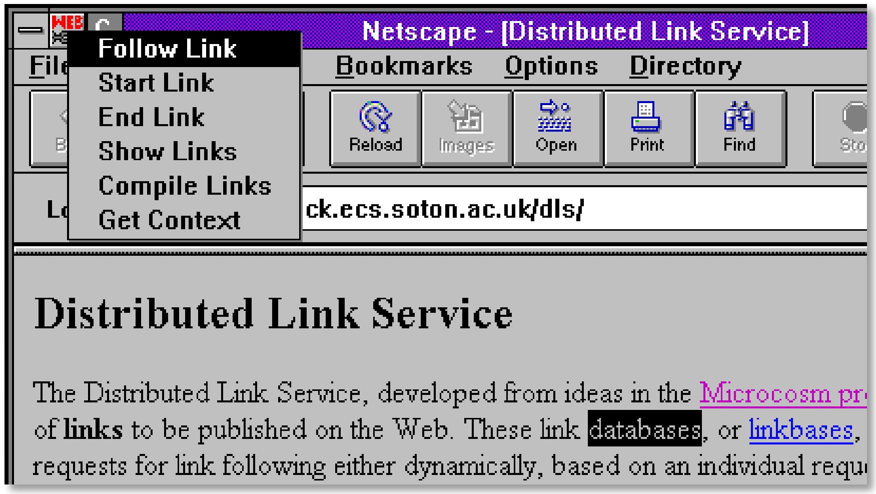


Figure 16 - DLS - A user Requests a Link from the Link Service Using the Client Interface[12]

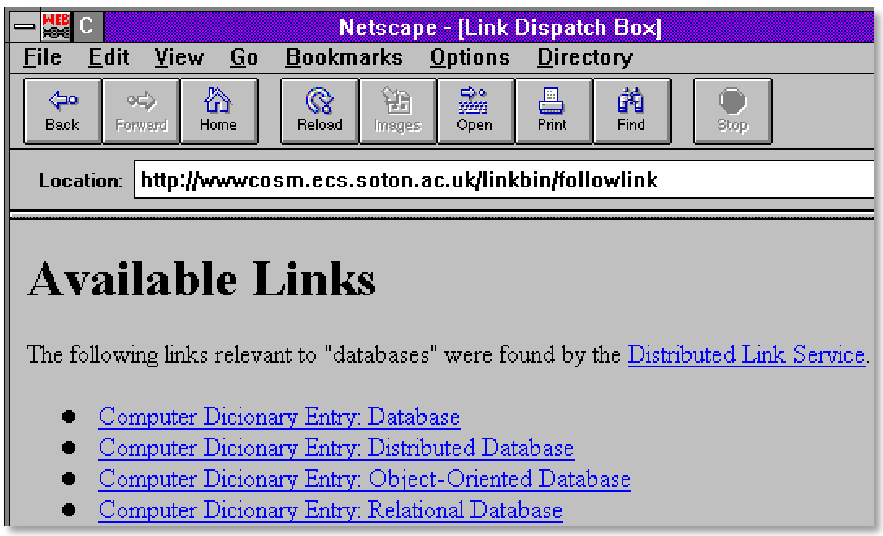


Figure 17 - DLS - The Link Service Responds with a List of Available Destinations[13]

It is unclear whether the links in DLS are bi-directional. Their schema calls for *source* and *destination*. Since the information is stored in the Linkbase, reverse navigation may still be possible, but the labeling is problematic. Information Fragment Associations involve implicit directionality in the case in which the author of the association is one of the authors of the two associated Information Fragments, however, in other cases they are purely bi-directional. The

[12] From [51] L. A. Carr, D. De Roure, W. Hall, and G. Hill, "Implementing an open link service for the World Wide Web," *World Wide Web*, vol. 1, pp. 61-71, 1998. http://dx.doi.org/10.1023/A:1019251328413

[13] From [51] Ibid.

approach used in this dissertation ensures that directionality is far less important than the semantics used in the typing of the relationship. For example, the type “supports” implies a direction.

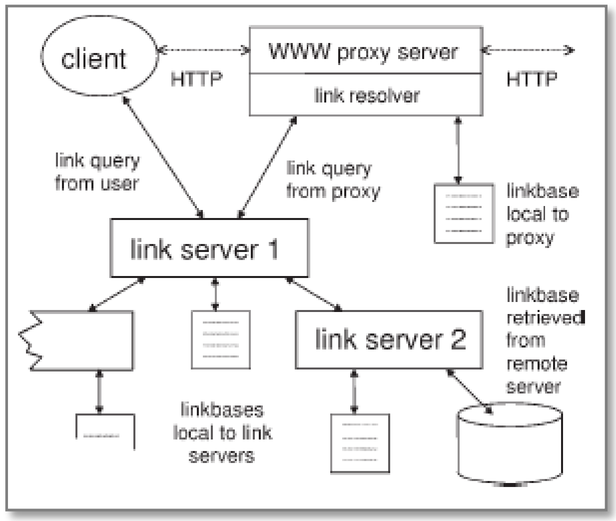


Figure 18 - DLS Network Model[14]

2.4.1.7 Webvise

Webvise[52] is an open hypermedia system developed at the University of Aarhus in Denmark, which provides external linking and other external extension capabilities to standard web pages and other documents. It is based on DeVise Hypermedia (DHM) framework[44] designed by Kaj Grønbæk and Randall Trigg[53]. It provides support for standalone client applications as well as for extensions to other applications such as Internet Explorer, Word and Excel. Webvise introduces the notion of a *global link* which unlike an ordinary *anchored link* (embedded link) is created externally. These global links are created within the displayed body of the client application by highlighting the passage to be linked to and right-clicking, as shown in Figure 19.

---

[14] From [51] Ibid.

After creation, the links are available in the document though automatic insertion of HTML tags through manipulation of the DOM[54]. This is achieved by accessing the given document through a proxy. When the browser is set to go through the proxy, the HTML pages called are altered to contain the links from the structure server. The changes to the HTML involved the addition of HTML anchor tags "<A>" as well as JavaScript code generating popup windows displaying the nodes created with the Webvise client. The Webvise client, which features a separate interface consisting of a *node browser* provides the capability of creating *guided tours*, the composite nodes containing other nodes and the graph connecting them.

Figure 20 shows the architecture of the Webvise Open Hypermedia Service. The Webvise client plays a key role in the communication between the end-user application, which can be a web browser, Word or Excel and the various Structure Servers. The communication between the Webvise client and the Structure Servers is achieved through an XML implementation of the Open Hypermedia Protocol specification (2.3.6).

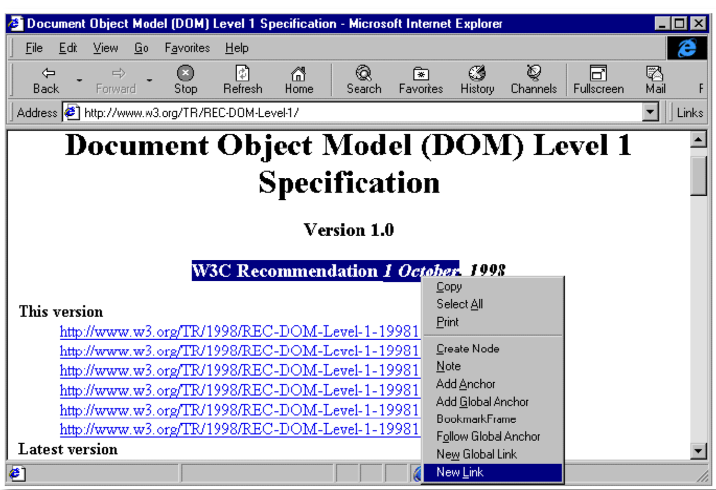


Figure 19 - Webvise - Microsoft Internet Explorer extended with open hypermedia services[15]

---

[15] From [52] K. Grønbæk, L. Sloth, and P. Ørbæk, "Webvise: browser and proxy support for open hypermedia structuring mechanisms on the World Wide Web," in *Proceeding of the eighth international conference on World Wide Web* Toronto, Canada: Elsevier North-Holland, Inc., 1999.

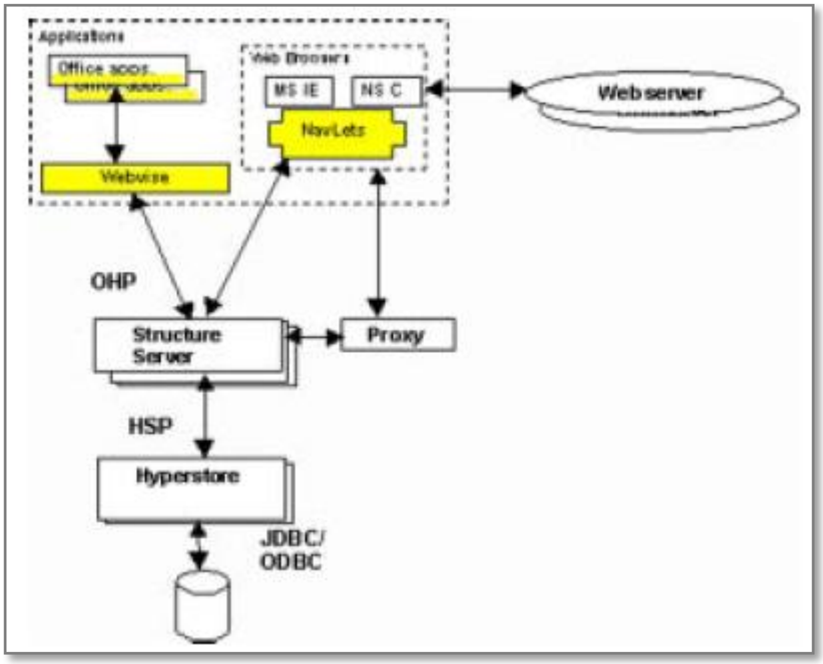


Figure 20 - The Architecture of the Webvise Open Hypermedia Service[16]

Webvise provides valuable functionality which can be used for establishing Information Fragment Associations. Just like most other tools, however, it does not handle fragments very well. It is more geared toward creating external equivalents of web links. An entire fragment placed within an HTML anchor element would look awkward at best. The current approach focuses on Information Fragments and employs side-by-side displays to make the association creation a productive experience.

2.4.1.8 Arakne

Just like Webvise (Section 2.4.1.7), Arakne[55] is a collaborative component-based system based on the Open Hypermedia Protocol (Section 2.3.6). It was developed by Nils Bouvin at the University of Aarhus in Denmark. Originally, unlike DLS (Section 2.4.1.6), Arakne would

[16] From [52]   Ibid.

modify a web page after the page is retrieved. This involved the modification of links. Later, instead of handling this locally Arakne was configured to interact with DHMProxy and at that point all link decoration was handled through that proxy (see Section 2.4.1.2 above). Its functionality involves detecting a web page or other file that the document display interface (i.e. the web browser) displays, retrieving the pertinent information from the server and modifying the web page by adding LocSpec information retrieved from the proxy (the Decorator) (see Figure 21). In a fashion similar to that of Webvise, Arakne accesses the Internet Explorer COM component and obtains access to the DOM object of the page. This and other activities are handled by the *Render Engine* a DLL interacting with the Java interface.

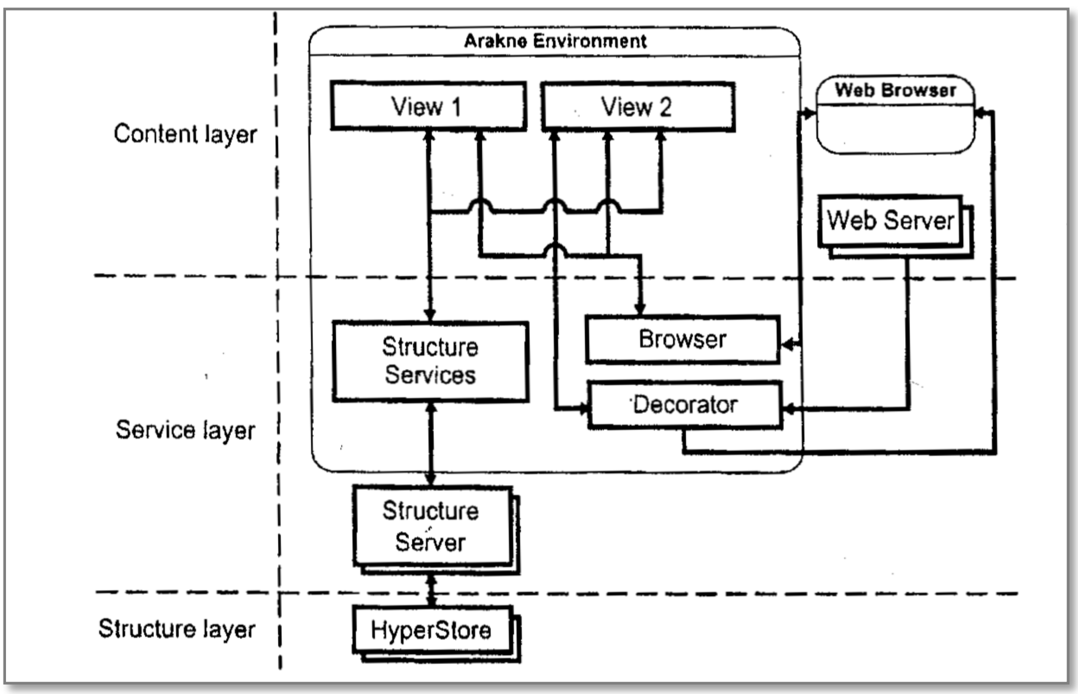


Figure 21 – Arakne[17]

Arakne also supports *fluid annotations*[56]. A *fluid annotation* is an animated annotation attached to the page, which can be displayed or not displayed as desired. *Fluid annotations* do

---

[17] From [55] N. O. Bouvin, "Augmenting the web through open hypermedia," *The New Review of Hypermedia and Multimedia*, vol. 8, pp. 3-25, 2003.

not change the formatting of the page and it does not hide any part of the page. A *fluid annotation* consists of an *anchor* and a *gloss*. An *anchor* is the primary material, i.e. the content originally found on the page to which the *fluid annotation* is applied. A *gloss* is the supporting material, i.e. the content added to the web page and attached to a specific *anchor*. Figure 22 shows the CNN page with four *fluid annotations*.

Figure 22 - Arakne - Fluid Annotations on the CNN Page[18]

The *anchors* of these *fluid annotations* are “Weather”, “Sports”, “Entertainment” and “rubble”. These *anchors* appear underlined with a dotted line. Each one of these *anchors* has a *gloss* attached to it containing specific information or personal links to allow quick access to custom information of interest. For example, under “Weather” which is an ordinary link on the

[18] From [56] P. T. Zellweger, N. O. Bouvin, H. Jeh, and J. D. Mackinlay, "Fluid annotations in an open world," in *Proceedings of the twelfth ACM conference on Hypertext and Hypermedia* Århus, Denmark: ACM Press, 2001. http://doi.acm.org/10.1145/504216.504224.

CNN page, a *gloss* with three additional specific city weather links appears. *Glosses* can be open or closed. In this example the "Sports" *gloss* is closed, whereas the "Weather" and "Entertainment" glosses are open.

Arakne offers an interesting approach, but just like Webvise it does not handle Information Fragments very well. It offers a very skillful and user-friendly insertion of links. However the current approach offers the ability to deal with an Information Fragment's content and not merely treat it as an anchor.

2.4.2 RDF-based OHS Systems

2.4.2.1 Annotea

One of the best examples of an annotation system is Annotea [57, 58]. The system consists of two components, the Annotation Server and Amaya, which is a web browser developed to incorporate special features for annotation creation, editing and browsing. An add-in for Mozilla/Firefox browser is also available called Ubimarks. The Annotation Server is an RDF[59] database which stores metadata in addition to storing link information. XPointer[3] is used to point to specific positions within the document being annotated. The body of the annotation is a URI referenced resource editable through the Amaya browser. Annotea goes a long way towards providing a means of linking to specific fragments within information entities. An example of how Annotea bookmarks are created is shown in Figure 23.

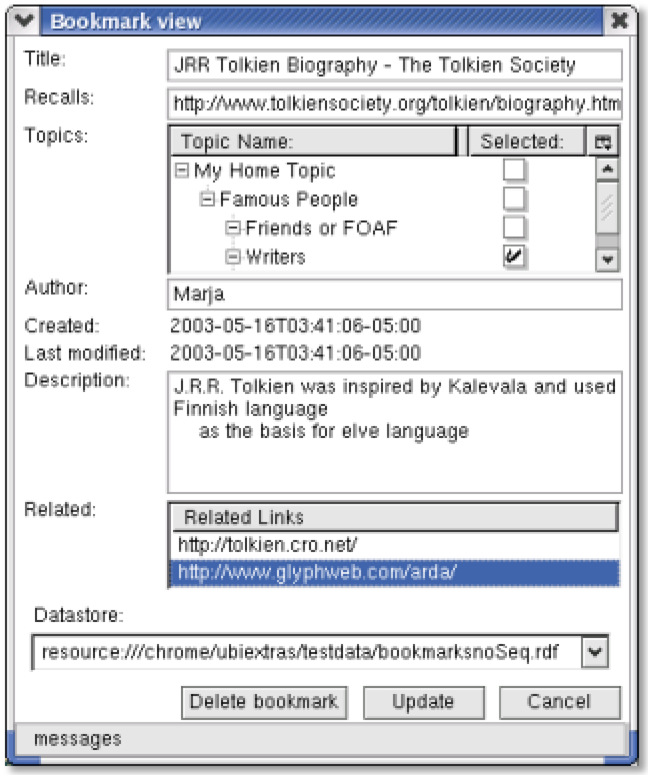


Figure 23 - Annotea - Properties of a Bookmark Presented in a Bookmark view[19]

Unfortunately, Annotea does not readily allow the association of the Information Fragments that it so elegantly defines. The association is between an Information Fragment and an annotation residing within its database. The approach used in this dissertation provides full Information Fragment Association capabilities.

2.4.2.2 WLS (Web Linking Service) and WebNote

WLS (Web Linking Service)[27] is an open hypermedia system which uses RDF[59] to store and exchange information about hypertext structure. The WLS conceptual model (Figure 24) defines the relationships between the several classes used by WLS. The *Anchor* class handles the location within a document. That location is recorded in the *expression* property.

---

[19] From [58] M.-R. Koivunen, "Annotea shared bookmarks: Semantic Web at your fingertips," in *International Semantic Web Conference*, Hiroshima, Japan, 2004. http://www.annotea.org/ISWC2004/annoteademo.html.

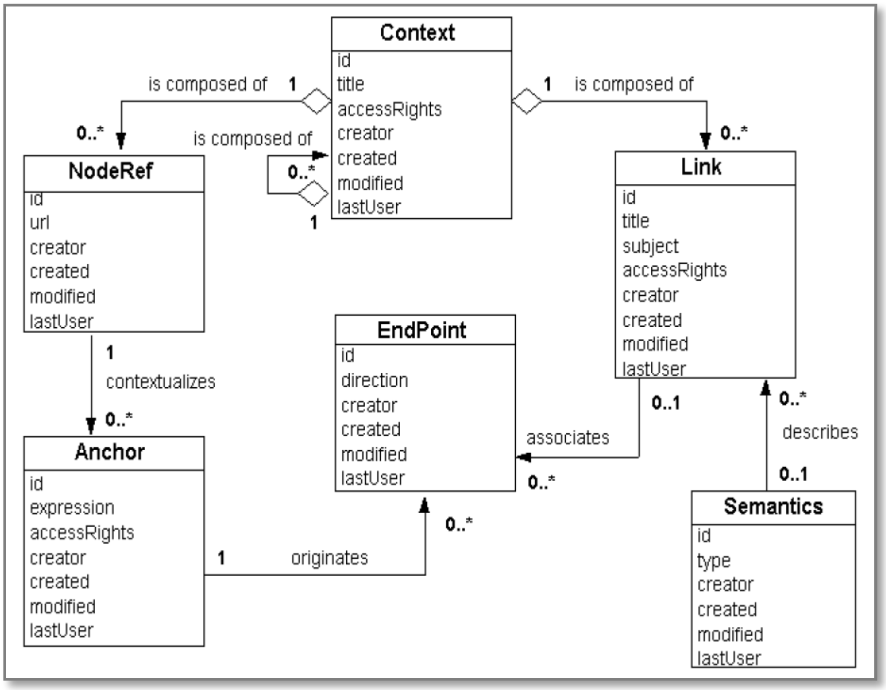


Figure 24 - WLS Conceptual Model[20]

An *Anchor* may have many *EndPoints*. The *EndPoint* class handles the direction of a link. The *Link* class handles the associations between the *EndPoints*. The *URL* property within the *NodeRef* class points to documents on the Web.

The classes represented in this model are used by server-side scripts. The server is accessible to client applications. When a link is created because a user selects a content section in a client application, an Anchor is defined in XPointer[3] syntax and the code for that Anchor along with the URL of the resource is passed on to the server in an XML message. The information is saved in WLS's linkbase and the client application marks the content accordingly.

WebNote is a client application for WLS. This application uses the paradigm of folders for organizing stored annotations. It offers several features, one of the most interesting of which is the capability of linking one annotation to another annotation. As Figure 25 shows, a user can

[20] From [27] R. B. Neto, C. A. Izeki, P. Maria da Graca, R. P. Fortes, and K. N. Truong, "An open linking service supporting the authoring of web documents," in *Proceedings of the 2002 ACM symposium on Document engineering*, McLean, Virginia, USA, 2002, pp. 66-73. http://doi.acm.org/10.1145/585058.585071.

initiate a linking operation from one annotation, browse the contents of another folder and select another annotation to link to.

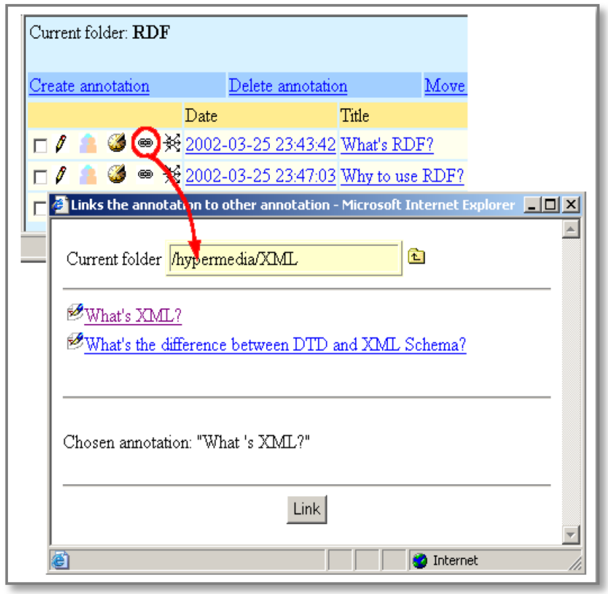


Figure 25 - WebNote - Linking two Annotations[21]

It is unclear whether WLS allows direct association of two Information Fragments. It allows the creation of annotations stored within the linking service. These annotations can be linked, which means that an indirect association at least is possible. The approach used in this dissertation allows for a more direct approach.

2.4.3 Web Services-based OHS Systems

2.4.3.1 Babylon

Babylon is a system supporting the integration of Open Hypermedia Services with Web Services. It aims at providing a process for creating Web Services and mapping functions of hypermedia services to operations of Web Services and vice versa.

---

[21] From [27] Ibid.

As Figure 26 shows, Babylon has three layers, the *storage layer*, the *taxonomic management layer* and the *client layer*. The *storage layer* is where the structural and non-structural information is stored and managed. The *taxonomic management layer* provides taxonomic creation and manipulation services. The *client layer* contains the applications. Some of these applications can have tree-structure providing tree service development capabilities.

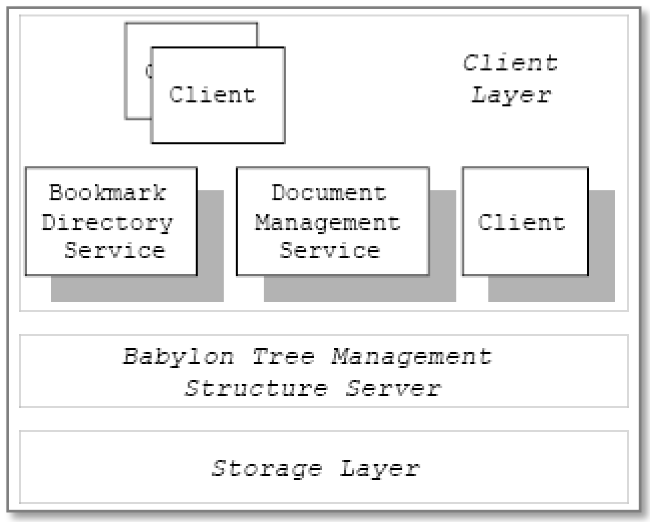


Figure 26 - Babylon Architecture[22]

This architecture is further enhanced with the incorporation of Web Services functionality. Babylon's OHS services can be exposed as web services as Figure 27 demonstrates. An interesting and innovative approach offered by Babylon is its introduction of Hypermedia Service Description Language (HSDL) which is based on WSDL, the Web Services Description Language[61, 62], which is a W3C Recommendation. HSDL provides service seekers with the essential information needed for determining what an Open Hypermedia System's has to offer in terms of functionality and how other systems can interact with it. It

[22] From [60] N. Karousos, I. Pandis, S. Reich, and M. Tzagarakis, "Offering open hypermedia services to the WWW: a step-by-step approach for developers," in *Proceedings of the 12th international conference on World Wide Web*, Budapest, Hungary, 2003, pp. 482-489. http://doi.acm.org/10.1145/775152.775221.

provides general service information, service interface information, service behavior information and general service comments.

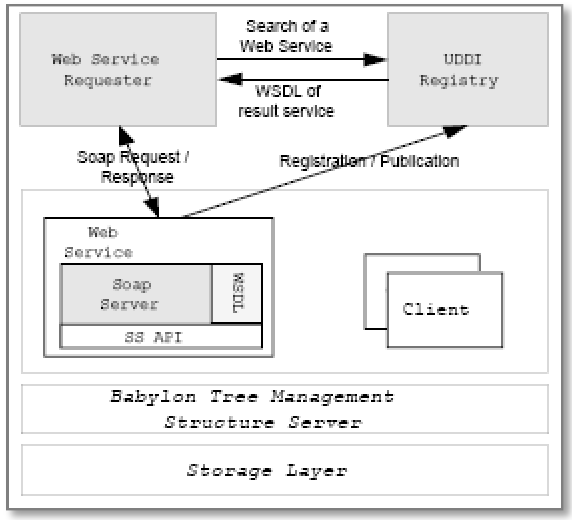


Figure 27 - Babylon Web Service[23]

Part of the process is publishing the service on the UDDI[63] Registry so it can be easily discovered and utilized by other systems.

Babylon is an invaluable system taking advantage of Web Services. Unfortunately, the currently available literature does not provide many details regarding its internal functionality as an individual open hypermedia system. The discussion concentrates on the discovery and interoperability of open hypermedia systems and as such it provides unique insights into possible new directions for open hypermedia systems. It is unclear, however, how it exactly handles Information Fragment Association. The approach used in this dissertation provides a comprehensive solution for Information Fragment Association and since this solution is Web Services based, it can very well fit well within the larger service discovery organization offered by Babylon.

[23] From [60] Ibid.

2.4.4 XLink-based OHS Systems

2.4.4.1 XLinkProxy

XLinkProxy[4] is a system providing a hyperbase service built with the W3C standards XLink and XPointer. XLinkProxy acts as an intermediary. When a document is requested from the web XLinkProxy checks to see if links for this document exist in any of its linkbases. If it finds links, it adds it to the document and returns it to the user with the links (Figure 28).

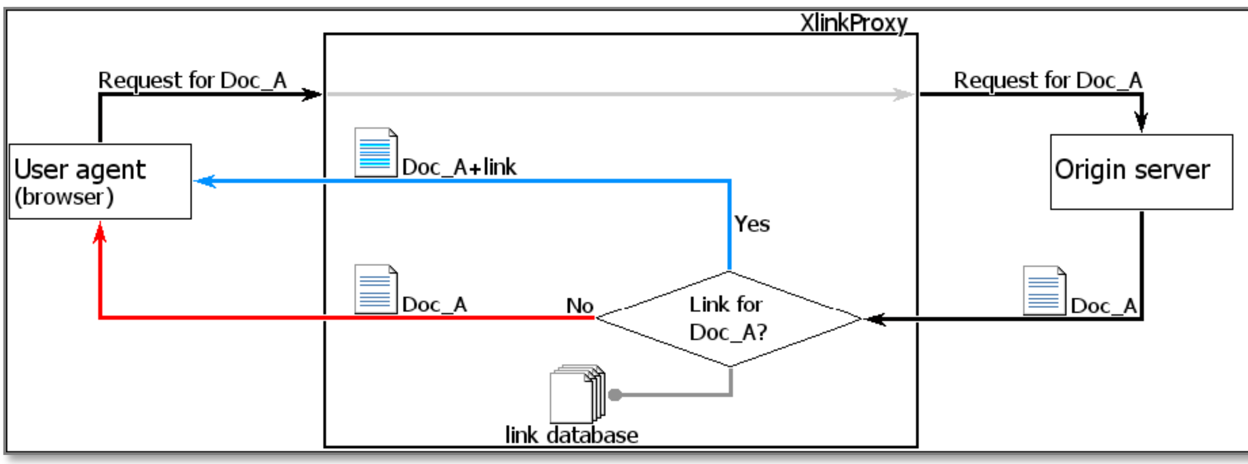


Figure 28 - The Base Process of the XLinkProxy Server[24]

XLinkProxy supports multiple links for one single location on the document. A user can right-click and navigate to any one of the available links. As Figure 29 shows, a user can add a link by selecting text and adding it as the "current startpoint" or the "current destination". At that point, the correct XPointer is calculated. The user has the capability of selecting the linkbase in which the link is to be stored.

[24] From [4] P. Ciancarini, F. Folli, D. Rossi, and F. Vitali, "XLinkProxy: external linkbases with XLink," in *2002 ACM symposium on Document engineering*, McLean, Virginia, USA, 2002, pp. 57 - 65. http://doi.acm.org/10.1145/585058.585070.

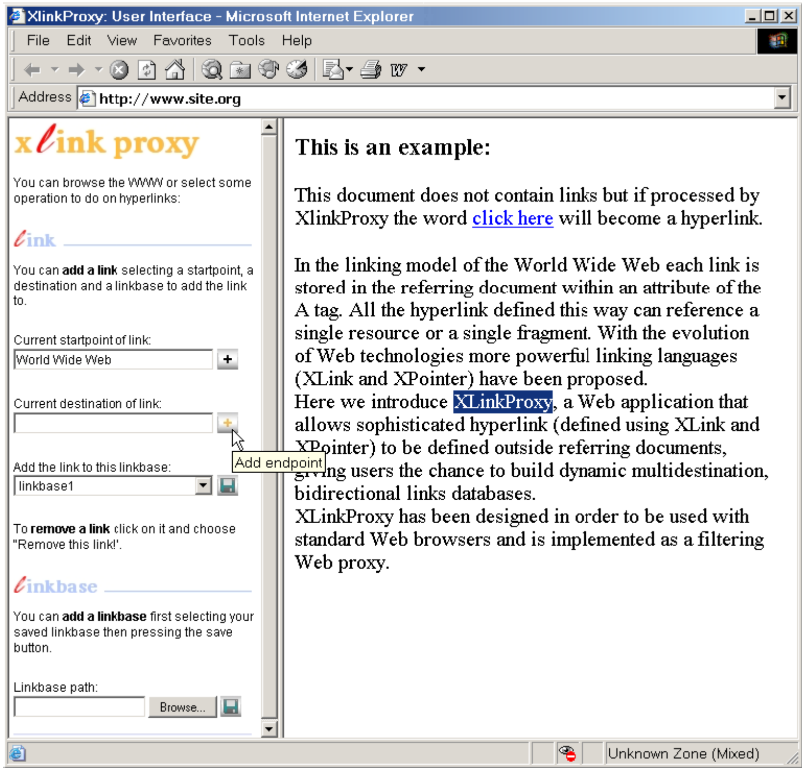


Figure 29 - XLinkProxy - Adding Links[25]

XlinkProxy provides a sound technical solution that can handle Information Fragment Association well. Unfortunately it does not go far beyond the mechanics of association functionality to describe how the relationships between these Information Fragments can be qualified and how these associations can be used to improve current practices.

2.4.5 Other Examples of Annotation Systems

Annotation systems are a type of open hypermedia systems whose primary, and often the only goal is to provide annotation capabilities. They provide a good example of Open Hypermedia Systems on which some of the characteristics of the Information Fragment are based. However, just like in the case of Nelson's Transclusion, these systems concentrate only on one aspect of

[25] From [4] Ibid.

the Information Fragment. As implemented, most annotation systems are even more restrictive in their nature than Transclusion. An annotation is an Information Fragment Association between an Information Fragment and a comment. This comment is some sort of Information Fragment but in most cases it does not reside within the content of another Information Entity. It is an independent fragment composed by the same person who creates the Fragment Association for the sole purpose of providing a commentary or some extension to the Information Entity being commented. One can reasonably argue that a comment is not an Information Fragment but a distinct Information Entity, albeit one dependent on another Information Entity. It does not have a context of its own. Its only context is that of the Information Fragment with which it is associated. An annotation is therefore a very specific subset of Information Fragment Association. Any work involving annotation can be very useful for this project, but it is inadequate to address most of the fundamental requirements for a framework providing a solution to the outlined problems.

The following are representative of the various types of annotation systems. All of them exhibit weaknesses when it comes to Information Fragment Association.

2.4.5.1 CritLink

CritLink[64] is an annotation system that goes beyond merely providing annotation capabilities. It provides a flexible linking model supporting bi-directional links. It is also designed to be used with any web browser, which makes it more attractive for implementation purposes, and more convenient to the user. CritLink distinguishes between *coarse-grained links* and *fine-grained links*. The *coarse grained links* are the ordinary web links addressing an entire document. The *fine-grained links* are those addressing specific fragments. The identification of the fragment is rather simplistic, depending on words on the text of the document, but this is sufficient to make the system functional.

CritLink addresses the challenge of annotation and external links by providing a panel with a secondary location space in which the user enters the desired URL. CritLink retrieves the entered resource embedding annotation markers on the content (Figure 30).

At the bottom of a page displayed though CritLink, a list of items linking to this page is generated and displayed (Figure 31). This list is generated as the result of a query to the CritLink hyperlink database.

CritLink also provides notification capabilities. A user could register on a page and be notified by e-mail every time a new annotation was made to this page.

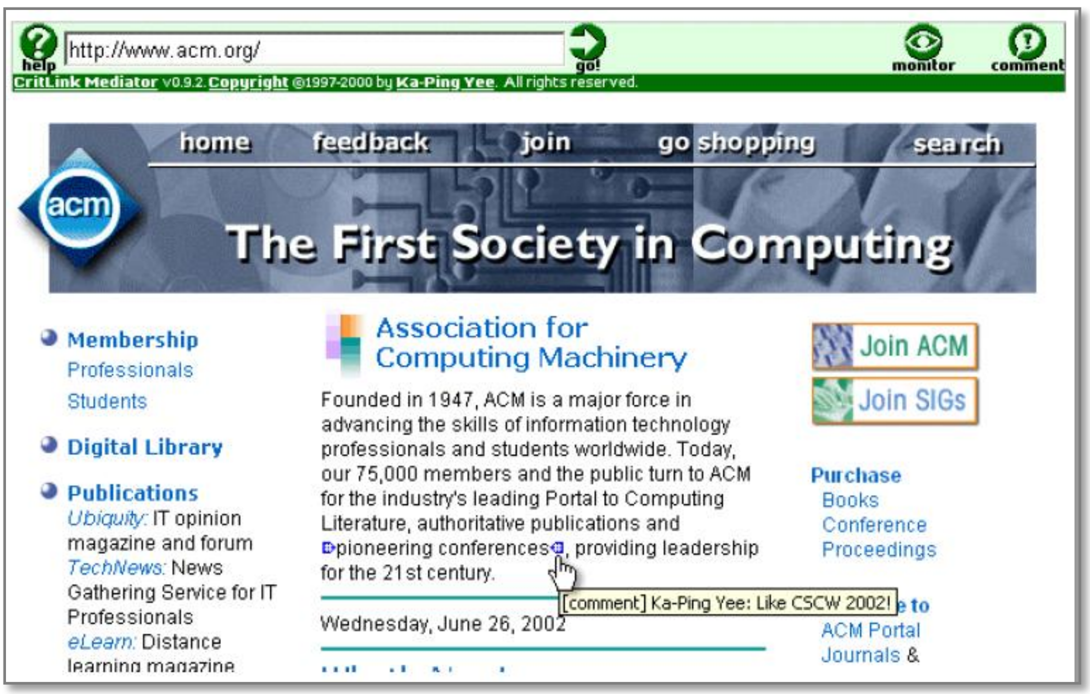


Figure 30 - CritLink Annotation[26]

[26] From [64] K.-P. Yee, "CritLink: Advanced Hyperlinks Enable Public Annotation on the Web," in *Computer Supported Cooperative Work (CSCW)*, New Orleans, Louisiana, USA, 2002. http://zesty.ca/crit/yee-crit-cscw2002-demo.pdf.

Figure 31 - CriLink Fine-Grained and Coarse Grained Backlinks appearing at the bottom of the page[27]

2.4.5.2 Annotation Engine

Annotation Engine[28] is a simple annotation proxy written by Wendy Seltzer of the Berkman Center for Internet & Society at the Harvard School of Law. It is inspired by CritLink (2.4.5.1) and simulates some of its functionality. It is strictly a simple annotation system, however, because it does not provide linking capabilities between resources. The annotations appear in a frame on the left. By clicking on an annotation in this frame the page scrolls down to the position of the annotation. When the content changes the annotations appear in the bottom as "orphaned nodes".

[27] From [64] Ibid.
[28] http://cyber.law.harvard.edu/projects/annotate.html

Figure 32 - Annotation Engine[29]

2.4.5.3 VIKI

VIKI is built as a Spatial Hypertext tool[65, 66] developed at Xerox PARC. It provides users with visual and spatial capabilities for organizing information. Content segments are manipulated as objects placed in hierarchically nested spaces. The data model used by VIKI has three types of elements: *objects*, *collections* and *composites*. The *Objects* are the nodes containing content. The *Collections* contain an arbitrary spatial arrangement of *objects* or other *collections*, forming a hierarchy. The *Composites* are combinations of two or more *objects* or *collections* in a particular visual configuration. VIKI handled visual and spatial relationships between objects. Examples of these relationships can be seen in Figure 33. These examples demonstrate the idea behind the development of this tool. Each spatial arrangement presents a set of relationships between content fragments without explicitly defining these relationships.

[29] From http://cyber.law.harvard.edu/projects/annotate.html

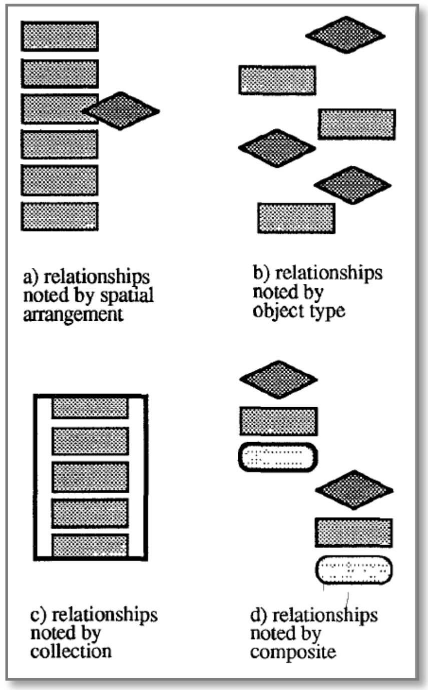


Figure 33 - VIKI Relationships[30]

One of the features provided by VIKI is that the user cannot choose to have the system provide suggestions regarding the generation of collections or composites. This could save the user a lot of time.

2.4.5.4 XLibris

XLibris[67] uses a display similar to a paper document and it allows users to mark the displayed document in a fashion similar to the user's traditional interaction with paper documents. There is a document viewer, a reader's notebook and a margin links area allowing the create links to related passages. In spite of its innovative approach and its unique, this system does not quite fit with the vision of a large distributed linking environment.

---

[30] From [65] C. C. Marshall, I. Frank M. Shipman, and J. H. Coombs, "VIKI: spatial hypertext supporting emergent structure," in *1994 ACM European conference on Hypermedia technology*, Edinburgh, Scotland, 1994, pp. 13 - 23. http://doi.acm.org/10.1145/192757.192759.

2.4.6 Commercial Annotation Systems

2.4.6.1 Third Voice

Third Voice is a commercial annotation system which had the distinction and misfortune of being one the first web page annotation systems to find its way out of the Academic environment into the wider public[68]. It provides the capability of attaching notes to web pages. This operation is handled with a sidebar (Figure 35) added by a browser add-in installed on client machines. Its introduction to the wider public created a controversy when web site providers protested against its use as “defacing” their web pages[69]. They even created an organization to fight it. Third Voice eventually went out of business for financial reasons and not because of the opposition it faced. Its contribution was far less technological than social.

Figure 34 - Third Voice Annotated Page[31]

[31] From [70] V. Wielbut, "Third Voice," *Spotlight: Online Newsletter of the Alliance for Community Technology*, June 28, 1999 1999. http://www.communitytechnology.org/newsletter/no2.html#thirdvoice.

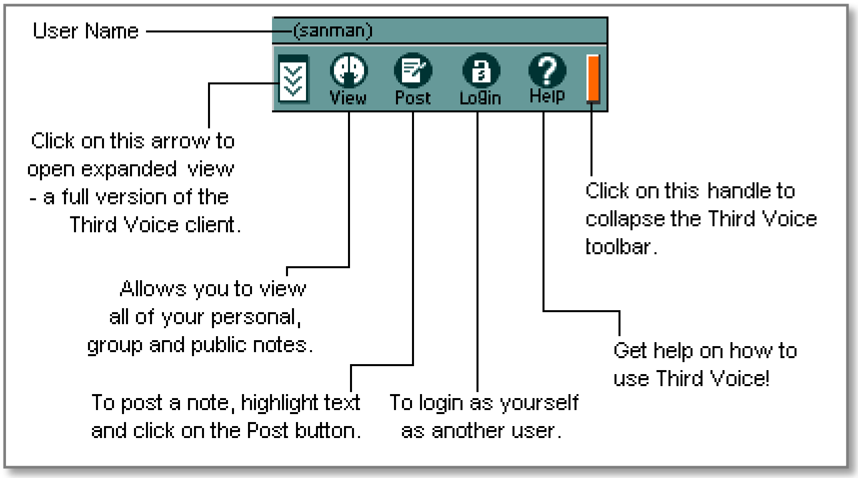


Figure 35 - Third Voice Sidebar Functions[32]

2.4.6.2 Fleck, Stickis, Diigo, Trailfire

Several new annotation systems have recently appeared resurrecting the functionality of Third Voice. They are trying to ride on the popularity of other social tools. Their functionality is very similar. They are being mentioned here without much discussion of their functionality and their differences in an effort to present a trend rather than discussing the details of potentially ephemeral tools.

Fleck[33] allows the addition of notes anywhere in the page. The note can be moved easily around. It supports new trends such as blogs which were not available with Third Voice. Stickis[34] is another recent annotation system similar to Fleck. One of the differences between the two tools is that Stickis has a browser toolbar that needs to be downloaded and installed, while Fleck has the tool appear within the page. This downloadable toolbar has some significant advantages. A user can sign in and open the toolbar, and every time s/he visits a page, a list of annotations available for that page appears in the toolbar. The user can then select an annotation

---

[32] From [70] Ibid.
[33] http://fleck.com/
[34] http://www.stickis.com/

in the toolbar and display it in the page (see Figure 37). Trailfire[35] has an interesting feature allowing a user to group an annotation s/he is creating with other existing annotations. This grouping creates a *trail* which can be followed using arrow buttons which appear in a toolbar and in the annotation box (see Figure 38). This allows the user to navigate from one page to the other and at the same way see the annotations about each one of those pages. Diigo[36] is by far the most advanced of these tools, combining the power of social bookmarking tools with that of annotation and blogging tools. It allows the user to highlight multiple Information Fragments, clip them to the social bookmarking tools and append popup annotations for them (see Figure 39).

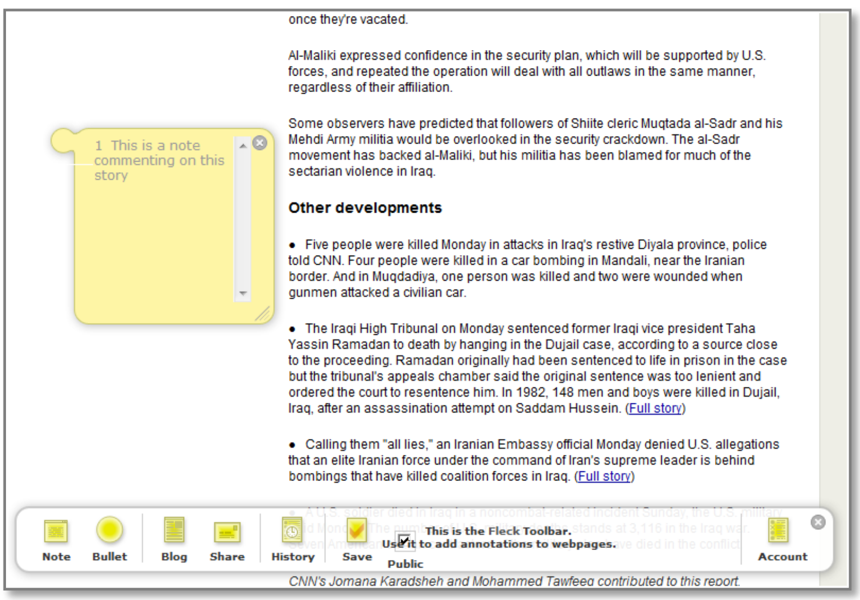


Figure 36 - Fleck Toolbar and Note[37]

---

[35] http://trailfire.com
[36] http://diigo.com
[37] From http://fleck.com

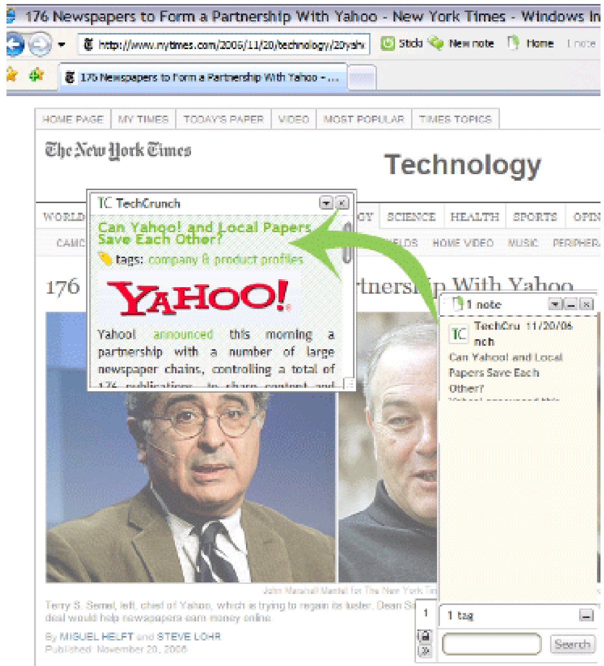


Figure 37 - Stickis[38]

Figure 38 – Trailfire[39]

[38] From http://www.stickis.com/faq/
[39] From http://trailfire.com/pferrel/marks/107508

Figure 39 - Diigo[40]

2.5 WEB SEARCHING

Both the relevance and the quality of the results retrieved by search engines have improved dramatically over the last few years. Algorithms available today have proven to produce fairly reliable results, at least in comparison with early search engine results. The “in-degree” approach has been the simplest one used. It ranks pages higher simply by considering the number of links coming in to them. Page and Brin’s PageRank[71], the algorithm used by Google considers the pages which link to a given page and how they rank in terms of importance. If a given page has links from pages considered to be of high quality, it receives a higher PageRank. Relevance is achieved through sophisticated text queries. The combination of relevant results with what is considered to be of higher quality results produces what made Google the number one search engine today. Kleinberg’s[72] algorithm determines which pages can be considered “authorities”

[40] From http://www.diigo.com/

and which pages can be considered “hubs”. He explains that “Hubs and authorities exhibit what could be called a mutually reinforcing relationship: a good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs”. Kleinberg’s techniques are further enhanced by Chakrabarti et al.[73] who introduce a weighted approach by assigning to each link a positive numerical weight. This weight increases with the amount of text around the link matching the searched topic.

These algorithms and techniques go a long way towards improving information retrieval. However, their scope is not to help produce more manageable information resources. They do the best with what is available to them, i.e. a vast number of interlinked pages. Search engine providers, with the exception of enhancement products such as Google Sitemaps[74], tend to focus on retrieving the information once it has been generated. This being their primary concern, they do attempt to take full advantage of every other format capable of providing them with more information, or information structured in a more manageable format. A very good example of this is the handling of RSS-based news feeds (a grassroots XML specification). Google’s “Google News”[75], Yahoo’s “Yahoo News”[31] and especially MSN’s “NewsBot”[76], based on the work of Microsoft Research and RSS feeds pioneer Moreover Technologies have taken full advantage of the immense popularity of RSS news feeds. Microsoft Research’s NewsJunkie project[77] introduced a system which among its several features has the capability of considering the news stories already reviewed by a user in order to determine the novelty of stories. Other approaches include the use of time-aware ranking algorithms [78]. A lot of these projects may appear interesting only as product improvements or academic exercises, but the application of this research can have enormous economic, social and political significance.

Just like in the examples mentioned above, the use of Information Fragment Association has the potential of being a major contributor to the improvement of ranking. It is conceivable to see the emergence of algorithms similar to that of Chakrabarti's mentioned above, which take advantage of these fragment association as they have been explicitly defined by users. Taking user-defined fragments into consideration would obviate the need of arbitrary selection of text surrounding a link, thus concentration on the fragment the user considers to be semantically important.

2.6 SOCIAL BOOKMARKING

Since the introduction of Internet browsers, the creation of bookmarks pointing to resources of interest has been one of the most favorite practices among users. We have since witnessed the evolution of bookmarking from a private practice of listing useful resources to a publicly shared activity. Social Bookmarking is an internet practice which has proven to be very successful and popular. The premier social bookmarking service, del.icio.us, has found broad acceptance among the Internet users, attracting more than just the usual enthusiasts. Evidence of this popularity is its financial success which led to its purchase by Yahoo, one of the premier internet portal providers. Social Bookmarking has extended the already popular practice of bookmarking to allow for sharing bookmarks and for assigning multiple tags to them which allows for a better categorization arrangement than the single-category hierarchical arrangement offered by web browsers.

The fact that a tool supporting Information Fragment Association provides the capability of creating entries about a specific URI, providing metadata information about this URI and sharing this entry with others may prompt its consideration as a Social Bookmarking system. A closer look, however, will reveal several fundamental differences alongside with the few similarities. Table 1 outlines some of the central features of the main Social Bookmarking systems and compares them to possible Information Fragment Association tools.

	CitULike	Connotea	del.icio.us	Furl	Yahoo MyWeb 2.0	Simpy	Spurl	Unalog	Windows Live Favorites	Tools Supporting Information Fragment Association
Online Bookmarks	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Tags and other meta – information	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sharing	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Focus	A[1]	A[1]	G[2]	G[2]	G[2]	G[2]	G[2]	G[2]	G[2]	G[2]
Caching	No[3]	No	No	Yes	Yes	No	Yes	No	No	Yes
Full-text searching of cached content	No	No	No	Yes	Yes	No	Yes	No	No	No
Rating	O[4]	No	No	Yes	No	No	O[5]	No	O[6]	O[7]
Importing browser bookmarks	No	Yes	Yes	Yes	Yes	Yes	Yes		Yes	No
Recognizing and retrieving bibliographic reference information	Yes	Yes	No	No	No	No	No		No	No
Exporting to browser bookmarks	No	No	Yes	Yes	Yes	Yes	Yes		Yes	No

	CitULike	Connotea	del.icio.us	Furl	Yahoo MyWeb 2.0	Simpy	Spurl	Unalog	Windows Live Favorites	Tools Supporting Information Fragment Association
Exporting to bibliographic citation tools	**Yes**	**Yes**	**No**	**Yes**	**No**	**No**	**No**		**No**	**No**
Public Private and Limited	**O[8]**	**PPL[9]**	**PP[10]**	**PP[10]**	**PPL[9]**	**PP[10]**	**PP[10]**	**PPL[9]**	**PP[10]**	**PP[10]**
RSS Syndication	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**No**	**Yes**
Copying a fragment in a field	**No**	**No**	**No**	**Yes**	**No**	**No**	**Yes**	**No**	**No**	**No**
Focusing on specific fragment	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**Yes**
Unique identifier for each fragment	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**Yes**
Associating a fragment of a resource with fragments in other resources	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**Yes**
User / Group trust and reputation mechanism	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**Yes**

1. A = Academic
2. G = General
3. However, it allows uploading of PDF files
4. No ranking, but it provides “priority” choices
5. No ranking, but it allows marking as explicit or not
6. No ranking, but it allows marking as “Top Favorite” or not
7. No explicit ranking of fragments themselves, but it allows for indirect ranking through its association with ranked Identities
8. It allows only restriction of Notes as Private
9. PPL = Public, Private and Limited
10. PP = Public and Private

Table 1 - Comparison of Social Bookmarking Systems

The comparison does not cover every piece of functionality offered by these systems. It focuses on functionality having some relevance to the issues of interest to this work. What this comparison reveals is that there is some common ground between tools supporting Information Fragment Association and Social Bookmarking systems, but there are some fundamental differences in the handling of Information Fragments. It is clear than none of the Social Bookmarking systems considers the fragment issue. All such systems are content to bookmark the Information Entity as it was composed by its creator, regardless of the multiplicity of sections with diverse content. Two of these systems, Furl and Spurl, encourage the copying of fragments to designated fields, but no attempt is being made to record their boundaries. They are being copied only for indexing purposes. On the other hand, these systems contain functionality that is considered to be beyond the scope of this Study. For example, CitULike and Connotea interact with bibliographic reference management utilities and FURL, Yahoo MyWeb 2.0 and Spurl provide the capability of full-text searching of cached content. Although these two examples of functionality are useful they do not have much to do with the fundamental goals of Information Fragment Association. Another notable difference is the fact that the tool supporting Information Fragment Association does not attempt to provide bookmark importing or exporting capabilities. Importing bookmarks would not benefit such a tool because traditional bookmarks point to entire Information Entities.

2.7 SUMMARY

Some of the issues that this work aspires to address have been identified and tackled in the past, but not in a satisfactory way. A variety of Semantic Web[79, 80] initiatives provide good solutions which to this date have been successful only within a limited scope of applications.

Link-base solutions[4, 81] have introduced some of the concepts on which this work is based, but failed to provide a comprehensive solution tacking all of the issues outlined above. Electronic mail and discussion board systems have introduced interfaces providing good arrangement by header entries, but they have so far failed to address the users' practice of referring to specific sections within each other's messages. Various studies have exposed users' tendency to annotate[76, 82], and systems have been developed to provide good interfaces for annotation[83, 84]. However these studies have not tackled sufficiently the management and accessibility of these associations especially as they are useful for retrieval purposes. Spatial Hypertext approaches[35, 85] do not really tackle the issues outlined above, but they can be useful in the design of appropriate interfaces.

3 STUDY DESIGN AND EXECUTION

3.1 GOALS

The study conducted in this dissertation is built on the simple premise that if a user identifies a relationship between an Information Fragment “a” residing within an Information Entity “A” and an Information Fragment “b” residing within an Information Entity “B” (sometimes after painstaking research), s/he will certainly benefit from the ability to easily record and be reminded of that relationship in the future. The goals of this Study are the following:

- Measure the Efficiency of Bookmarking and Retrieving Information Fragments
- Estimate the Accuracy of Association
- Evaluate User Experience
- Determine the Need for Enhancement of Search Engine Results
- Calculate Recall and Precision of Searches in Social Bookmarking Tools

3.1.1 Measure the Efficiency of Bookmarking and Retrieving Information Fragments

The first goal of the Study is to determine whether Information Fragment Association improves the efficiency of bookmarking and subsequently (re)finding inter-related Information Fragments residing within Information Entities. Specifically this study attempts to establish whether an environment allowing users to define the boundaries of Information Fragments and associate them with other Information Fragments would increase the efficiency of information retrieval as compared with an environment offering similar capabilities except the ability to establish an Information Fragment Association. (see Hypothesis #1 and Tasks 1 and 2 in Section 3.2)

3.1.2 Estimate the Accuracy of Association

The second goal of this Study is to determine whether Information Fragment Association improves the accuracy of an association by focusing on the specific content of the Information Fragment which precipitated this association as opposed to an association involving the entire content body within which the Information Fragment resides (see Hypothesis #2 and Task 3 in Section 3.2)

3.1.3 Evaluate User Experience

The third goal of the Study is to consider the effects that an efficient juxtaposition and collection of Information Fragments has on the user experience and to determine whether users feel they can more easily locate Information Fragments using an Information Fragment Association interface (see Hypothesis #2 and Task 3 in Section 3.2)

3.1.4 Determine the Need for Enhancement of Search Engine Results

The fourth goal of the Study is to determine whether the presence of Information Fragment Associations improves the efficiency of retrieving resources containing interrelated Information Fragments. The study will attempt to determine the extent to which search engine results can be enhanced by Information Fragment Associations. The Study examines how closely two resources containing user-associated Information Fragments are collocated in the results generated by standard search engines which do not have Information Fragment Association enhancements (see Hypothesis #4 and Task 4 in Section 3.2)

3.1.5 Calculate Recall and Precision of Searches in Mainstream Tools

The fifth goal of the Study is to determine whether better retrieval effectiveness (measured by Recall and Precision) of keyword searches can be achieved, by integrating Information Fragments and the associations between them rather than more traditional keyword searches on Information Fragments. (see Hypotheses #5, #6 and #7 and Task 5 in Section 3.2)

3.2 HYPOTHESES

The Study consists of seven research questions and their corresponding hypotheses.

3.2.1 Research Question #1

Would the total time needed for the entire process of bookmarking and then of retrieving two Information Fragments within two Information Entities be shorter if an Information Fragment Association interface rather than a mainstream social bookmarking tool like SPURL is used?

3.2.1.1 Hypothesis #1

H_{1-0} - The first null hypothesis:

The time needed for the entire process of bookmarking and then retrieving two Information Fragments from two Information Entities when using an Information Fragment Association interface will be equal to the time needed for this process when using SPURL

H_{1-1} - The first alternative hypothesis

The time needed for the entire process of bookmarking and then retrieving two Information Fragments from two Information Entities will be shorter when using an Information Fragment Association interface than when using SPURL

3.2.2 Research Question #2

Would the overall usefulness and usability of the process of bookmarking and then of retrieving two Information Fragments within two Information Entities be better when using an Information Fragment Association interface rather than using a mainstream social bookmarking tool like SPURL?

3.2.2.1 Hypothesis #2

$H_{2\text{-}0}$ - The second null hypothesis:

The usefulness and usability of the process of bookmarking and then retrieving two Information Fragments from two Information Entities will be the same between using an Information Fragment Association interface and using SPURL

$H_{2\text{-}1}$ - The second alternative hypothesis

The usefulness and usability of the process of bookmarking and then retrieving two Information Fragments from two Information Entities will be better when an Information Fragment Association interface is used rather than SPURL

3.2.3 Research Question #3

How useful would users find the ability to establish Information Fragment Associations, to see Information Fragments side-by-side and to navigate from one Information Fragment to the other?

3.2.3.1 Hypothesis #3

$H_{3\text{-}0}$ - The third null hypothesis

Users would find the abilities to establish Information Fragment Associations, to see Information Fragments side-by-side and to navigate from one Information Fragment to the other are not useful

$H_{3\text{-}1}$ - The third alternative hypothesis

Users would find useful the abilities to establish Information Fragment Associations, to see Information Fragments side-by-side and to navigate from one Information Fragment to the other are useful

3.2.4 Research Question #4

When a human determines that two Information Fragments contained in two different Web pages have a strong relationship, can s/he expect that if s/he searches for one of these two Information Fragments in a Web search engine these two pages will appear within a reasonable distance from each other in the result set? In detail, the research question is: if we have a set of such pairs of Information Fragments, will a major Web search engine (such as Google or Live Search) be able to return the two Web pages containing the two Information Fragments within the same result page (i.e., the difference between the ranks of the two pages is less than 10) in most (more than 75%) cases when the query issued to the search engine is designed to retrieve one of the Information Fragments, or would it be necessary to insert the related Information Fragment in the result set in order to accomplish that?

3.2.4.1 Hypothesis #4

$H_{4\text{-}0}$ - The fourth null hypothesis

More than or equal to 75% of a given set of pairs of related Information Fragments will have a rank difference less than 10 in the Web search results when a query designed to retrieve one of the Information Fragments is applied to a Web Search engine

$H_{4\text{-}1}$ - The fourth alternative hypothesis

Less than 75% of a given set of pairs of related Information Fragments will have a rank difference less than 10 in the Web search results when a query designed to retrieve one of the Information Fragments is applied to a Web Search engine

3.2.5 Research Question #5

Would better retrieval effectiveness (measured by Recall, Precision and F-measure) be obtained when keyword searches are performed on a dataset containing user-defined Information Fragment pairs in a tool supporting Information Fragment Association rather than performed on the same dataset in a tool not supporting Information Fragment Association?

3.2.5.1 Hypothesis #5

$H_{5\text{-}0}$ - The fifth null hypothesis

Recall obtained when keyword searches are performed on a dataset containing user-defined Information Fragment pairs would be the same in a tool supporting Information Fragment Association as in a tool not supporting Information Fragment Association

$H_{5\text{-}1}$ - The fifth alternative hypothesis

Recall obtained when keyword searches are performed on a dataset containing user-defined Information Fragment pairs would be better in a tool supporting Information Fragment Association as in a tool not supporting Information Fragment Association

3.2.5.2 Hypothesis #6

$H_{6\text{-}0}$ - The sixth null hypothesis

Precision obtained when keyword searches are performed on a dataset containing user-defined Information Fragment pairs would be the same in a tool supporting Information Fragment Association as in a tool not supporting Information Fragment Association

$H_{6\text{-}1}$ - The sixth alternative hypothesis

Precision obtained when keyword searches are performed on a dataset containing user-defined Information Fragment pairs would be better in a tool supporting Information Fragment Association as in a tool not supporting Information Fragment Association

3.2.5.3 Hypothesis #7

$H_{7\text{-}0}$ - The seventh null hypothesis

F-measure obtained when keyword searches are performed on a dataset containing user-defined Information Fragment pairs would be the same in a tool supporting Information Fragment Association as in a tool not supporting Information Fragment Association

$H_{7\text{-}1}$ - The seventh alternative hypothesis

F-measure obtained when keyword searches are performed on a dataset containing user-defined Information Fragment pairs would be better in a tool supporting Information Fragment Association as in a tool not supporting Information Fragment Association

3.3 RESEARCH METHODOLOGY

This Study employed primarily quantitative research techniques but some qualitative analysis was also performed. The choice of the quantitative approach as opposed to the qualitative was based on the consideration of the objectives of the study and the nature of the tests feasible within the timeframe of the study. Since the main objective of the study was the examination of the effect that the presence or absence of Information Fragment Association functionality has on everyday information management and retrieval operations, the most suitable approach was a quantitative one, using causal-comparative and experimental research techniques.

The quantitative-experimental approach provided the advantage of better control for testing how a given process was performed with or without Information Fragment Association capabilities, how efficiently it was carried out, what steps it involved and how subjects rated the process during each step. Having this control was made possible by designing the various tasks to involve procedures using tools offering a set of functionalities that is parallel and similar in most regards except in the use of Information Fragment Association capabilities.

Given the nature of study and the timeframe in which it was carried out, relying primarily on a qualitative approach would not have yielded results as strong as those yielded by the quantitative approach. Some qualitative analysis was performed using conversations with the subjects. However, it had to be used very cautiously, as subjects were not so extensively exposed to the Study as to have a good grasp of the concepts involved. Their responses may have been affected by various factors other than Information Fragment Association, and even when asked explicitly to comment on this functionality their comments may have been based on incomplete understanding. The short time spent with the subjects made it difficult to provide qualitative assessments and interpretations of the way they interacted with the tested tools. Such assessments would have been far less reliable than the quantitative measurements taken during the observations.

The disadvantage of the quantitative approach followed in this Study is that it offers vary little understanding as to why subjects carried out their tasks in a specific way. For example, much more observation will be needed to determine why a particular functionality that was obvious to most subjects was not obvious to others. This is this Study attempted to use qualitative analysis to complement the quantitative analysis whenever possible.

Three Subject Groups were used for the Study. The Study involved four tasks, referred to as Task 1-Task 4. Measurements were taken both while they were performing the assigned tasks as well as after the tasks as an analysis of the recorded information. The measurements were the following:

3.3.1 Time, Completion and Retrieval Measurements as Performance Indicator

Both experimental and causal-comparative measurements were performed. The subjects were timed as they were performing assigned tasks. The effect was the time taken to perform the tasks and the causal factor was the use of two different tools for the same task, and more specifically the presence or absence of Information Fragment Association functionality. In particular, the subjects were timed as they were performing the task of defining Information Fragments using two tools being compared (see Section 3.9.2.2) and then they were timed again as they were performing the task of retrieving the Information Fragments using the two tools (see Section 3.9.2.4). The data was analyzed further by drawing distinctions between specific groups of observations. For example, in Task 1 an analysis by topic was performed (see Section 4.1.5). In Task 2 another analysis was performed by dividing the observations on the basis of the original creator of the Information Fragment (see Sections 4.1.6.1 and 4.1.6.2) and eventually dividing them further by gender (see Sections 4.1.6.3, 4.1.6.4, 4.1.6.5, 4.1.6.6 and 4.1.6.7). In addition to timing, other measurements were also collected throughout these tasks. For example, in Task 2 the observations during which an Information Fragment was actually retrieved were counted(see Section 4.1.1). Other indicators involved measurements such a calculation of the difference of the ranks of Information Fragments in result sets of Web search engine queries (see Section 4.2.3) and the calculation of recall, precision and F-measure of queries performed on Information Fragment sets (see Section 4.5)

3.3.2 Questionnaires as an Indicator of User Experience

Throughout the first two tasks of the Study, subjects were asked to respond to questionnaires. Most questions were posed using Likert scales. The quantitative descriptive data were collected and analyzed, providing measurements of the degree to which the evaluated interfaces were easy to use, helpful, effective or enjoyable (see Sections 4.3.1, 4.3.2.1, 4.3.2.2 and 4.3.2.3).

3.3.3 Quantification of Free-Text Questions

Free-text questions were posed to subjects giving them the opportunity of pointing out what they liked and what they disliked about the two tools they had used. The responses were analyzed by categorizing the contents and by extracting measurements of positive or negative comments on specific functionalities (see Section 4.4). This quantification was very helpful in drawing conclusions, as it helped weed out factors influenced by functionalities unrelated to the object of the Study, which was the use of Information Fragment Association.

3.3.4 Qualitative Data

Some more data were collected through casual conversations with the subjects. Although to a certain degree some effort was made to quantify these responses as well, some qualitative conclusions regarding the usefulness of Information Fragment Association were derived from these conversations (see end of Section 4.4).

3.3.5 Experimental Validity

3.3.5.1 Selection-Treatment Interaction Threat to Internal Validity

The subjects were University of Pittsburgh students primarily in the Social Sciences. This preference was aimed at ensuring that the subjects were more motivated to carry out their tasks, since the material used was probably of more interest to them than to other students. This selection is believed to have yielded for the most part results that are characteristic of the average user.

3.3.5.2 Experimenter Effects

Since the selection of the stories was not the goal of this study, and in order to save some of the time the subjects were to spend in selecting the stories, the lists of story pairs were provided to the subjects. Special care was taken so as not to use any criteria in this random selection other than the simple judgment of their being suitable. Attention was also paid to avoiding any suggestions for their use or selection.

3.4 SUBJECTS

We totally recruited three groups of subjects, each consisting of 6 University of Pittsburgh students, whose majors are primarily in Social Sciences. The subjects were given an Entry Questionnaire in order to ensure that some basic information about them has been gathered. The profiles of the subjects are:

- **Gender Balanced**

  50% of the subjects were Male and 50% Female. A gender balance was being targeted, but the perfect 50/50 distribution was simply a lucky coincidence

- **Mostly Young**

  Most subjects were undergraduate students. The mean age of the subjects was 24.1 and the median age was 21.5, with the youngest being 18 and the oldest being 43.

- **University students mostly from Arts and Sciences**

  Twelve subjects were students in the School of Arts and Sciences, two in the Graduate School of Business, two in the School of Education, one in the Graduate School of Public and International Affairs and one in the School of Social Work. Twelve of them had only a High School Degree, three had a Bachelor's degree and three had a Master's degree

- **Frequent computer users**

  The subjects were asked how much time on average they spend per day using a computer. Six use a computer more than four hours a day, six use a computer between three to 4 fours, three use a computer between two and three hours and three use a computer between one and two hours.

- **Comfortable with Web navigation tools**

  The subjects were self-rated on degree to which they feel comfortable with web navigation tools, with 1 as less comfortable and 10 as the most comfortable. The mean rating was 7.47

- **Casual news readers**

  When asked how much time on average they spend per day reading, listening and watching news, one of the subjects stated that he/she spends two to three hours a day on news, eight of them spend one to two hours and nine of them spend less than an hour.

- **Mostly unfamiliar with Social Bookmarking tools**

  The twelve subjects of Subject Groups A and B were asked to specify the Social Bookmarking tools they had used. Three of them had used *Slashdot*, three had used *Yahoo My Web* and one had used *Digg*. No subjects had used *SPURL*, which was the Baseline of this study

- **Familiar with Web Search Engines**

  The six subjects of Subject Group C were asked to specify the Search Engines they use. All of them use *Yahoo* and *Google*. Two of them use *Altavista*, two use *Ask*, two use *Excite*, one uses *Baidu*, one uses *Vivismo*, one uses *Lycos* and one uses *Dogpile*

Two of these three Subject Groups were administered the test under a controlled environment being supervised by the test administrator. The third Subject Group was able to perform the assigned task within their usual environment and within a defined timeframe. Institutional Review Board permission was received prior to the test, specifying the details of interaction with the subjects.

Because of the fact that the sample size of 12 subjects was relatively small, testing has been performed on each measurement to ensure that the results possess statistical power. Since as Chapter 4 shows the tests were performed and they revealed statistical power, the sample size was proven adequate. The power of the results was to a large extent due to the fact that the test was based on the means of 12 actions of each one of the 12 subjects. The sample was carefully selected so as to be characteristic of the population being examined. Specifically, the subjects were undergraduate and graduate students in the Social Sciences who had indicated that they use on a regular basis online information resources and search tools for their information needs and who did not have expertise in information management. These were the characteristics indicative of the targeted population.

3.5 INDEPENDENT VARIABLES

The Study measured the effect of the presence or not of Information Fragment Association functionality in bookmarking tools and in the results of web searching. For this reason, two tools with a comparable feature set were selected, the one with Information Fragment Association capabilities and the other without. Specifically the Study examined the effect of the ability or not of defining multiple Information Fragments within the same Information Entity and ability or not of defining the boundaries of Information Fragments and displaying them within the context of their encompassing Information Entities. It also considered the ability or not of displaying inter-related Information Fragments side-by-side. In all of the tasks performed in this Study, the Independent Variable is the availability of Information Fragment Association capabilities. The presence or absence of these capabilities is the cause for the differing performance or usability measurements recorded using the tools employed.

3.6 DEPENDENT VARIABLES

Several variables were considered in an effort to assess the degree of improvement in social bookmarking and web searching tools. First, the Study measured the efficiency of retrieval of pairs of Information Fragments residing within different Information Entities and identified by users as bearing a certain relationship. Then, using questionnaires, it measured the user satisfaction in the process of retrieval of pairs of Information Fragments and the perceived reference accuracy of Information Fragments. It also measured the degree to which a mainstream web searching tool succeeded or failed to provide collocation of Information Entities encompassing inter-related Information Fragments. Finally, the study measured the Precision, Recall and F-measure of search queries performed in the two social bookmarking tools with the purpose of retrieving relevant Information Fragments.

3.7 CONCEPTS RELATING TO AN INFORMATION FRAGMENT ASSOCIATION INFRASTRUCTURE

3.7.1 Information Entity

For the purpose of our discussion, an Information Entity is any piece of recorded human communication, regardless of format, length, subject matter and scope. An Information Entity may consist of a single statement or a series of statements. It may be a piece of casual communication, such as an e-mail, or an elaborately composed scholarly publication. As such, an information entity may have some kind of relationship with a pre-existing Information Entity. It may be a response, confirmation, refutation, elaboration etc. Under this assumption, an

Information Entity would be a piece of communication issued as a literally or conceptually single item by an individual or a group. For example, a document written by a group of members of a committee constitutes an Information Entity, and each piece of communication issued by each one of the members of the committee in the process of composing this common document would be an information entity. The final document would not subsume any one of these information entities; it would simply inherit content from them. There would therefore exist a certain dependence between these Information Entities. The Information Entities issued during the process are dependent on the final document, and the final document is dependent on the earlier information entities as they constitute the history of its creation. Depending on an organization's information retention policies this history may or may not be considered important. Earlier communications are often discarded. Sometimes, however, they are invaluable.

A simple example of inter-related information entities is electronic mail, since information entities are essentially pieces of human communication. Electronic mail exchanges demonstrate the dependence of an Information Entity on another one, as they are often responses to statements expressed by the other party. In human communication, if Message B contains a reference to Message A, Message B is sequentially dependent on Message A. The degree of this dependence will vary based on the nature of this reference. The dependence can be absolute, such as in the case of a simple response:

e.g. Message A: Are you available for a conference call this afternoon at 2:00pm?

Message B: Yes

In this simple example, Message B has no real semantic value of its own. It can be a response to any question expecting a "yes" or "no" answer. A more in-depth demonstration of this type of absolute dependence can be found in Louwerse and Mitchell's[86] discussion of

"Discourse Markers". They define Discourse Markers "verbal and nonverbal devices that mark transition points in Communication". They are phrases instructing participants in an information exchange how to consider an upcoming utterance. Therefore, such phrases often do not have any meaning by themselves. Knott and Mellish[87] examine a large number of "cue phrases" which, when used, render a phrase dependent on further context in order to be interpreted. As an example, they offer the following two phrases:

- Bill is six feet tall. (requires no additional context to be understood)
- *But* Bill is six feet tall. (only makes sense as the follow-up to some previous statement)

This sequential dependence is not limited to oral communication, which is the concentration of the above studies. Typed electronic communication, either synchronous or asynchronous can exhibit the exact same sequential dependencies. Moreover, this kind of dependence expressed in the above studies can be further applied to the content of statements being exchanged. Even a free-standing, fully coherent statement may only make complete sense if viewed with a statement that prompted its creation.

In other cases, the reference may be passing and insignificant. Most cases, however, fall somewhere between these two extremes. Statement B above does have semantic content of its own, but the lack or modification of Statement A has some effect on the message it conveys.

To further complicate things, Information Entities often refer to knowledge which is assumed common. Humans have the ability of building mental information knowledge bases which are often called upon by embedded allusions in a given information entity. These knowledge bases are dependent on cultural background, age, education etc. Perhaps the most striking examples of concealed allusions occur in humorous contexts. An old television political satire may make sense only to audiences of a certain age. Scholars are still struggling to discover what the numerous humorous allusions in Aristophanes' comedies are. Any information entity saved for posterity exhibits some degree of lack of semantic integrity if the assumed “common” knowledge is not recorded along with the information entity. This latter inadequacy of recorded information entities will not be discussed by this study. It is only referred to here in order to demonstrate how context is important, especially when it can be assumed that some context will inevitably be lost.

3.7.2 Information Fragments

The sequential dependence of two Information Entities can be traced in interrelated fragments within these two Information Entities. In the case of literary works this may mean the detection of points of influence of the one work on the other. In the case of news stories this may mean the identification of points of explicit or implicit reference to a previous news story. These points being referred to are fragments of the entire body of an Information Entity. An *Information Fragment* is a content portion of an *Information Entity* the boundaries of which have been clearly defined by somebody. The long and multi-faceted content of some Information Entities makes it often very difficult to discern a useful piece of information buried in the content of such a large Information Entity. The definition of boundaries for Information Fragments allowing the user to focus on the content of interest and the potential of navigating from fragment to fragment is a much needed capability.

3.8 TOOLS EXAMINED

Since the objective of this study is to determine the desirability of an Information Fragment Association Framework by establishing whether the presence of Information Fragment

Association capabilities would improve information retrieval, it was necessary to make a comparison between a mainstream Social Bookmarking tool which does not support Information Fragment Association as a Baseline and a Test Tool (called FW) with similar functionality plus the Information Fragment Association capability.

More detailed discussion of these two tools and their functionality, including screen snapshots is available in Section 3.9 below, as part of the discussion of the tasks performed in this Study.

The focus of the study was on a few select features of these tools and not on their entire functionality. The tested features were those providing the basic capabilities of defining Information Fragments, submitting the information to the system as a bookmark and subsequently retrieving that bookmark.

3.8.1 Baseline

SPURL[41] was selected as the Baseline because it incorporates most functionalities found in typical Social Bookmarking tools, such as bookmarking, subject tagging and bookmark sharing, and in addition to these functionalities it supports features rendering it more suitable for comparison with a tool supporting Information Fragment Association. These additional functionalities are the ability to select and define Information Fragments (albeit inadequately) and the ability to cache the content of the encompassing Information Entity.

Therefore, SPURL possesses, to some degree, some of the features an envisioned Information Fragment Association tool should have, but it does not support Information

[41] http://spurl.net

Fragment Association. SPURL is a web resource accessible using a web browser. For this test the Internet Explorer web browser was used.

3.8.2 FW (Test Tool)

FW is the test tool developed for this Study. It provides the basic capability for Information Fragment Association with necessary data and organization structure. As a test tool it is not polished or full-featured like a real product. FW used Web Services[88, 89] and XQuery[19] for data transfer and manipulation. Despite its simplicity, FW was designed as envisioned for a Framework supporting Information Fragment Association at much grander scale. FW was a stand-alone application with built-in web capabilities.

3.8.3 Live Search

To test Research Question #4 and Hypothesis #4 a retrieval system built on top of Live Search with a custom designed interface was used. It connects with the Live Search engine via the Live Search API. The subjects used a web browser to interact with the retrieval system.

3.9 TASKS

The study involves four tasks. Task 1 and Task 2 test hypotheses #1, #2 and #3. These two tasks together were designed to draw a comparison between the benefits offered by two tested tools during the bookmarking process and the subsequent use of bookmarks to retrieve the collected resources, and specifically the user-defined Information Fragments. The three hypotheses tested involve performance, usability and feature usefulness. Measurements for all three have been built into the design of these tasks. Throughout the design, the focus is on the effect that the presence or absence of Information Fragment Association capabilities on these measurements. The tasks involve both objective and subjective measurements (such as time and questionnaire responses), so special care has been taken in the design to make sure that even if other factors influence the results, the main criterion is always the difference in functionality introduced by the use of Information Fragment Association. For example, answers to the questions in the questionnaires can be influenced by a variety of factors, such as aesthetics, but these are only peripheral factors. The main focus of the tasks is always on the effectiveness of accomplishing the objective of creating and retrieving Information Fragments.

Task 3 tests hypothesis #4 and it involves the use of a mainstream Web Search Engine. It attempts to assess the degree to which everyday information seeking activities can be enhanced through integration of Information Fragment Associations.

Task 4 tests hypotheses #5, #6 and #7 and it tries to find out how the presence or absence Information Fragment Association affects the performance (measured by recall, precision and F-measure) of queries performed against a dataset containing Information Fragments.

3.9.1 Task 1 (Fragment Definition)

Task 1 covers the first part of the traditional bookmarking process, that is, the identification and reference recording of resources.

3.9.1.1 Data Provided and Data Collected

12 pairs of stories were provided for each subject. The stories were selected from four topic categories:

- International other than Iraq (many dealing with the Iran nuclear issue)
- Events in Iraq
- U.S. Politics
- Miscellaneous stories

The pairs of stories were pre-selected, but the specific fragments were selected by the subjects themselves. The stories provided the subjects with ample opportunity of identifying Information Fragments bearing some relationship with each other. Very often the stories were reports from different news agencies on the same event, or different analyses on same issue. Additional stories were kept in the event that the subject failed to identify any related Information Fragments, but they were never used for that purpose. However, backup resources proved to be useful in a few occasions in which a particular resource was not responding or it was presenting some other technical malfunction.

Two groups of 6 subjects (User Group A and User Group B) participated in Task 1, and they totally selected 144 pairs of Information Fragments. Their subjective opinions of the two environments were also collected using questionnaires.

3.9.1.2 Subject Group A

The subjects were first asked to fill an Entry Questionnaire and then participated in a training session that included reading the instructions and playing with the two tools.

The news stories were provided to the subjects in the form of a simple web page with a series of links as shown in Figure 40. In all, 13 pairs of stories were provided with the first one as a practice. This practice story pair was the same for all subjects and its results were used in the result analysis. During that practice pair creation, subjects were encouraged to ask questions in case the process was unclear to them.

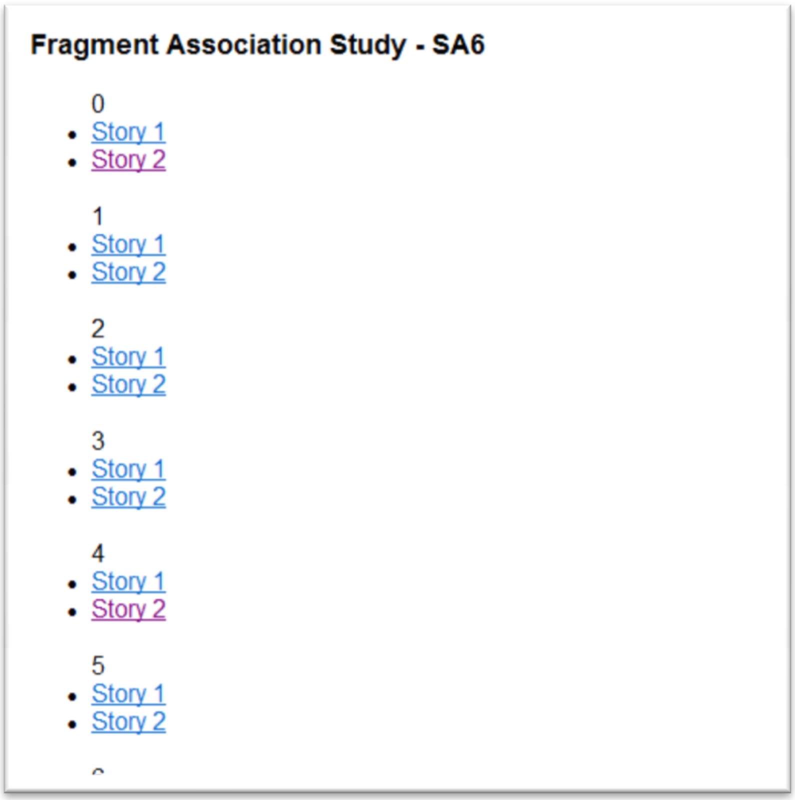


Figure 40 - Task 1 List of News Stories Links

The subjects had the following open on their machines:

- An Internet Explorer session for Baseline
- An FW session for Information Fragment Association creation

The subjects were asked to read the content of these 12 pairs of news stories in a fashion similar to their normal everyday news consumption and to identify a pair of Information Fragments bearing some relationship with each other from each of these 12 pairs of stories. The nature of the relationship was left entirely up to the subjects to decide. It could be similarity of content, a different slant or bias in reporting the same story, emphasis on different details, refutation, or any other positive or negative relationship.

The following process was followed by Subject Group A in Task 1. We took the within subject design so that each subject tackled a pair of stories in an iteration which included the use of both tools. Special care has been taken so as to avoid learning effect. Each subject performed 6 selections of Information Fragments, with the one sequence of the two tools (e.g., using Baseline first then FW) and then 6 remaining selections with the switched sequence (e.g., using FW first, then Baseline).

In summary, each subject submitted 6 pairs of Information Fragments using this sequence:

- Within one submission, the subject
  - displayed the first story in the first tab of the Baseline IE session
  - displayed the second story the second tab of the Baseline IE session
  - displayed the first story in the first tab of FW
  - displayed the second story in the first tab of FW
  - read the two stories
  - identified the first fragment
  - identified the second fragment

  - started timing 🕑 Baseline
  - bookmarked the first story using Baseline, highlighting the fragment of interest and keeping it as a “snip”
  - bookmarked the second story using Baseline, highlighting the fragment of interest and keeping it as a “snip”
  - ended timing 🕑 Baseline

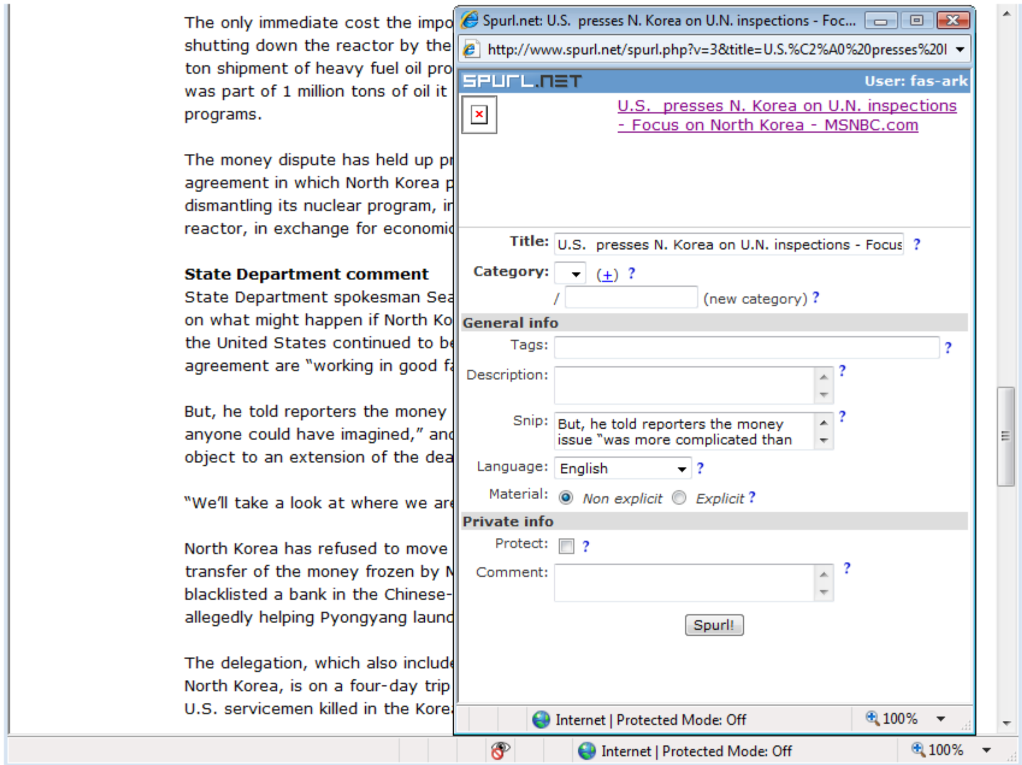


Figure 41 - Baseline Information Fragment Definition

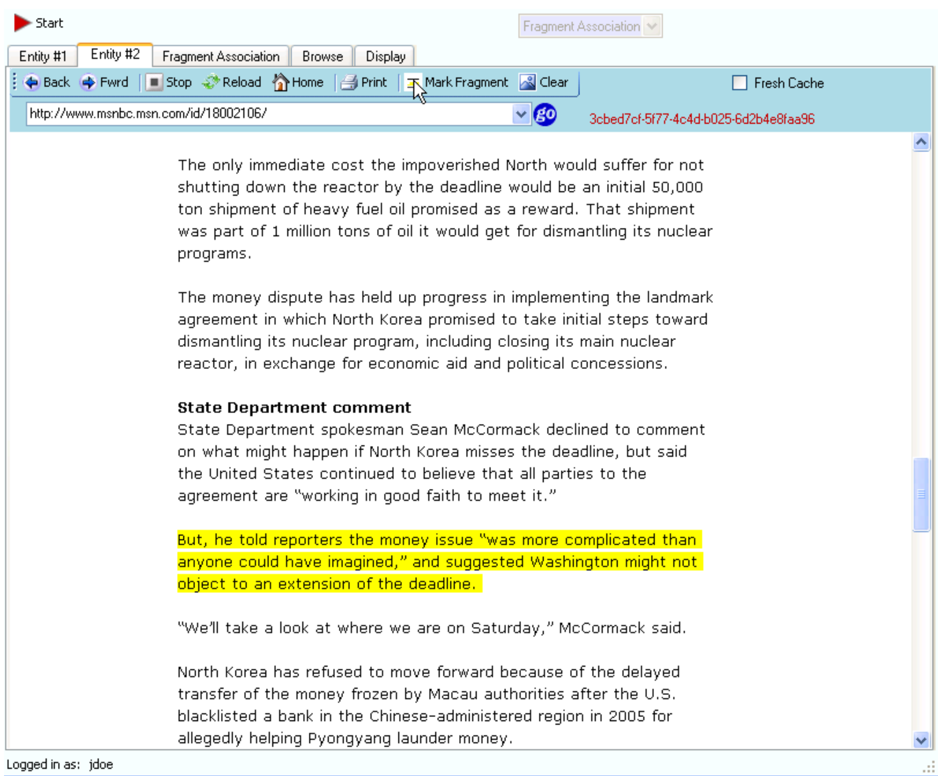


Figure 42 – FW (Test Tool) Information Fragment Definition

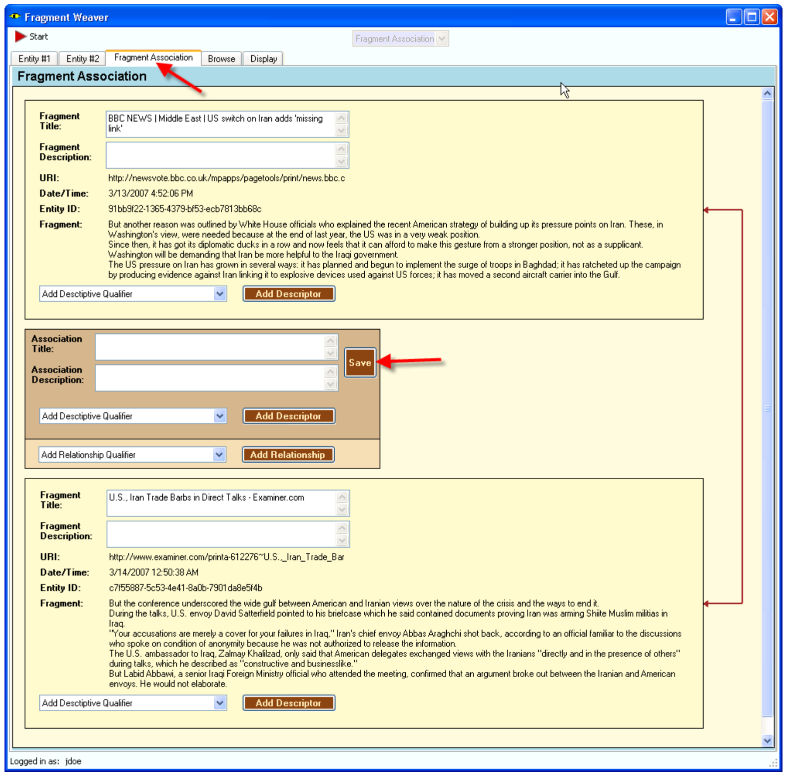


Figure 43 - Information Fragment Association Submission

- started timing 🕑 FW
- created Information Fragment Association for the same Information Fragments using FW (Figure 42 and Figure 43)
- ended timing 🕑 FW

- Then the subjects were asked to submit 6 more Information Fragments using a slightly different sequence, using FW first and then Baseline
- After finishing the 12 selections, the subjects filled out a questionnaire with the following questions:
  - How difficult was it to define a fragment in SPURL?
  - How difficult was it to define a fragment in FW?
  - How useful is the FW feature allowing the connection of two fragments?
  - How helpful was SPURL in accomplishing the task of defining fragments?
  - How helpful was FW in accomplishing the task of defining fragments?

- How enjoyable was using SPURL for accomplishing this task?
- How enjoyable was using FW for accomplishing this task?

- At the end two sets of fragments were collected by each subject. The two sets were labeled A1 and A2. Set A1 contained all first story Information Fragments (12 items) and A2 contained all second story Information Fragments (12 items), i.e. the Information Fragments with which the members of set A1 were paired (see Figure 44).

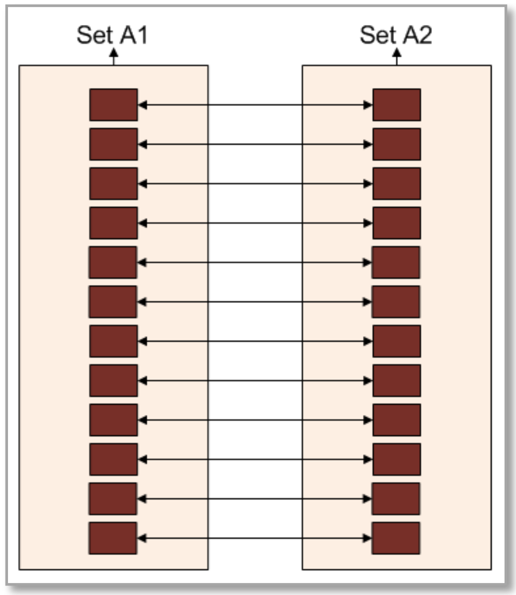


Figure 44 - Information Fragment Sets

Task 1 by itself is not able to fully exhibit all of the benefits of Information Fragment Association. Yet the fact that the two Information Fragments are being associated so effortlessly adds to the overall final assessment of the value of Information Fragment Association. Had this process been time-consuming or convoluted, it would have cast doubt on the benefits offered by Information Fragment Association.

3.9.1.3 Subject Group B

Subject Group B performed the exactly same test as Subject Group A. The two sets of Information Fragments were labeled B1 and B2.

3.9.2 Task 2 (Matching Fragment Retrieval)

In Task 2, subjects were presented with a listing of references to the first story and used these references to retrieve the matching Information Fragment in the second story (see Figure 45). The process was divided in such a way that subjects had to use the one tool for half of these retrievals and the other tool for the other half. These references were also divided into two groups. Half of them corresponded to the Information Fragments that the subject had defined herself or himself before, and the other half of the references corresponded to the Information Fragments defined by somebody else.

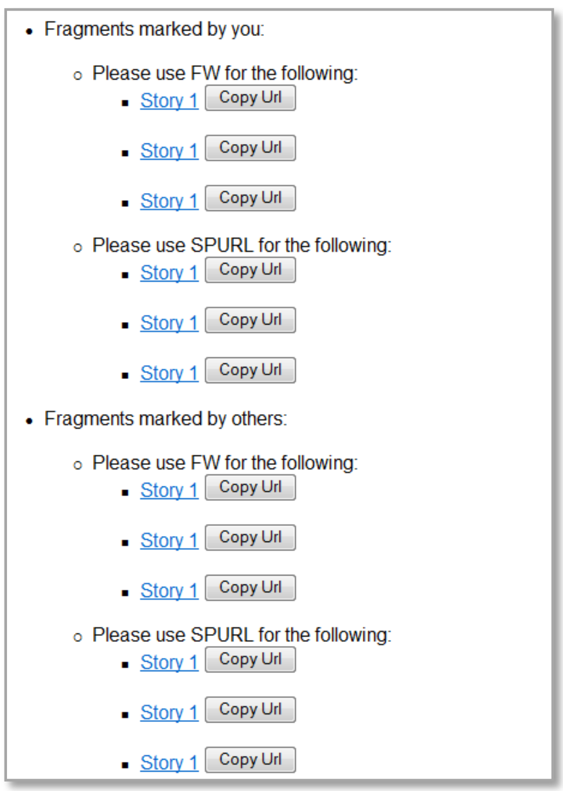


Figure 45 - Task 2

In spite of the complexity of its design, this task is very simple in its objective and execution. It provides the opportunity to observe how users may benefit in their information seeking activities from the existence of pre-defined Information Fragment Associations.

Conversely, it helps estimate the extent to which these activities may be more onerous if either the user cannot take advantage of relationships between Information Fragments or there is no such association to be used.

3.9.2.1 Subdivision of Information Fragment Sets for Task 2

Figure 46 represents the two Information Fragment Sets containing the first Information Fragment of each pair. These two sets are subdivided for Task 2 and they are used as outlined below and aligned in the sequence in which they were used by each Subject Subgroup.

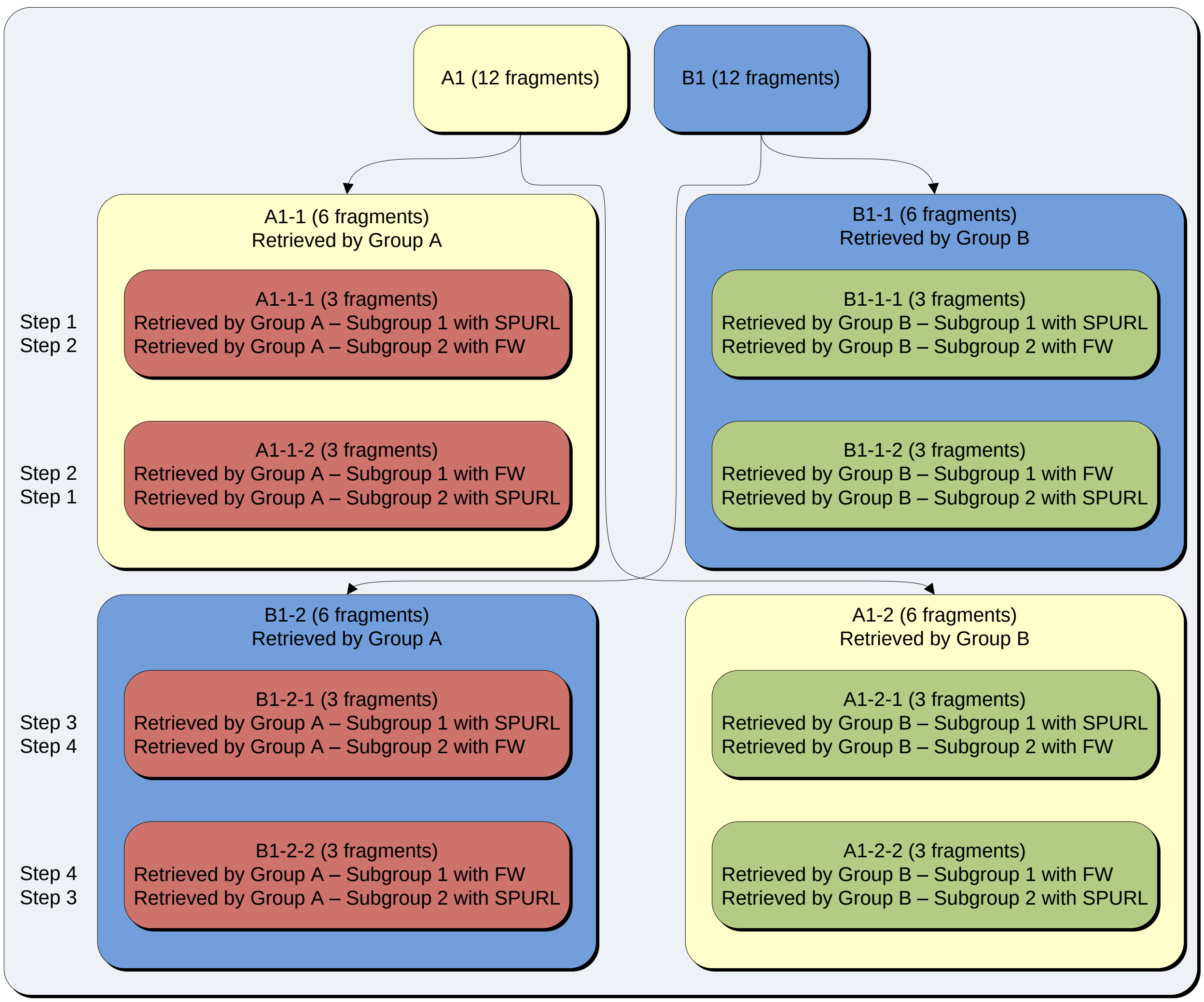


Figure 46 - Subdivision of Information Fragment Sets for Task 2

3.9.2.2 Subject Group A

The subjects of Subject Group A were given the resources of set A1-1 (the first half of the first set of Information Fragments created by Subject Group A – i.e. 6 items) and set B1-2 (the second half of the first set of Information Fragments created by Subject Group B – i.e. 6 items). They had the following open on their machines:

- An IE session containing the A1-1 and B1-2 URIs
- An IE session for SPURL (Baseline) for the A1-1 and B1-2 stories
- An IE session for SPURL (Baseline) for the A2-1 and B2-2 stories
- An FW (Test Tool) session for fragment association creation

A *within-subject design* was employed to remove the difference between subjects using the two tools. Subject Group A was divided into two subgroups each consisting of 3 subjects. Subgroup 1 used SPURL first and then FW, while Subgroup 2 used FW first and then SPURL:

Subject Group A – Subgroup 1

As Figure 46 shows, the first subgroup of Subject Group A performed the following Information Fragment retrievals. The steps are presented in detail in Section 3.9.2.4 below.

- Own Information Fragments
  - Subset A1-1-1 (3 fragments) retrieved with SPURL (Baseline) – Step 1
  - Subset A1-1-2 (3 fragments) retrieved with FW (Test Tool) – Step 2
- Others' Information Fragments
  - Subset B1-2-1 (3 fragments) retrieved with SPURL (Baseline) – Step 3
  - Subset B1-2-2 (3 fragments retrieved with FW (Test Tool) – Step 4

Subject Group A – Subgroup 2

As Figure 46 shows, the second subgroup of Subject Group A performed the following Information Fragment retrievals. The steps are presented in detail in Section 3.9.2.4 below.

- Own Information Fragments
  - Subset A1-1-1 (3 fragments) retrieved with FW (Test Tool) – Step 2
  - Subset A1-1-2 (3 fragments) retrieved with SPURL (Baseline) – Step 1
- Others' Information Fragments
  - Subset B1-2-1 (3 fragments retrieved with FW (Test Tool) – Step 4
  - Subset B1-2-2 (3 fragments) retrieved with SPURL (Baseline) – Step 3

3.9.2.3 Subject Group B

The subjects of Subject Group B were given the resources of set B1-1 (the first half of the first set of Information Fragments created by Subject Group B – i.e. 6 items) and set A1-2 (the second half of the first set of Information Fragments created by Subject Group A – i.e. 6 items)

Exactly the same test performed by Subject Group A was performed by Subject Group B, using sets B1-1 and A1-2.

Subject Group B – Subgroup 1

As Figure 46 shows, the first subgroup of Subject Group B performed the following Information Fragment retrievals. The steps are presented in detail in Section 3.9.2.4 below.

- Own Information Fragments
  - Subset B1-1-1 (3 fragments) retrieved with SPURL (Baseline) – Step 1
  - Subset B1-1-2 (3 fragments) retrieved with FW (Test Tool) – Step 2
- Others' Information Fragments

- Subset A1-2-1 (3 fragments) retrieved with SPURL (Baseline) – Step 3
- Subset A1-2-2 (3 fragments retrieved with FW (Test Tool) – Step 4

Subject Group B – Subgroup 2

As Figure 46 shows, the second subgroup of Subject Group B performed the following Information Fragment retrievals. The steps are presented in detail in Section 3.9.2.4 below.

- Own Information Fragments
  - Subset B1-1-1 (3 fragments) retrieved with FW (Test Tool) – Step 2
  - Subset B1-1-2 (3 fragments) retrieved with SPURL (Baseline) – Step 1
- Others' Information Fragments
  - Subset A1-2-1 (3 fragments retrieved with FW (Test Tool) – Step 4
  - Subset A1-2-2 (3 fragments) retrieved with SPURL (Baseline) – Step 3

3.9.2.4 Task 2 Steps

Step 1 – Retrieving Matching Fragment for One's Own Fragment – SPURL (Baseline)

Each Subgroup used a different subset for this step:

- The Subgroup 1 of Subject Group A used A1-1-1/A2-1-1 pairs
- The Subgroup 2 of Subject Group A used A1-1-2/A2-1-2 pairs
- The Subgroup 1 of Subject Group B used B1-1-1/B2-1-1 pairs
- The Subgroup 2 of Subject Group B used B1-1-2/B2-1-2 pairs

Each subject was first asked to perform 3 matching fragment retrievals using the above subset of pairs:

- used Baseline to find the first story

- brought up Baseline in the first tab
- pasted the URL in the Baseline “Search” box and submitted search
- clicked on the plus “+” icon to display the fragment (Figure 47)

- used Baseline to find the corresponding second story
  - brought up Baseline in the second tab
  - used information from the first fragment to find the second fragment either by searching (Figure 48) or by browsing (Figure 49)
  - displayed the second fragment

After completing their 3 matching fragment retrievals, the subjects were asked to answer the following questions (on 7-point Likert scales):

- Using SPURL, how easy was it to find the fragments you had previously defined?
- How effective was SPURL in accomplishing the task of retrieving the fragments?
- How enjoyable was using SPURL for accomplishing this task?

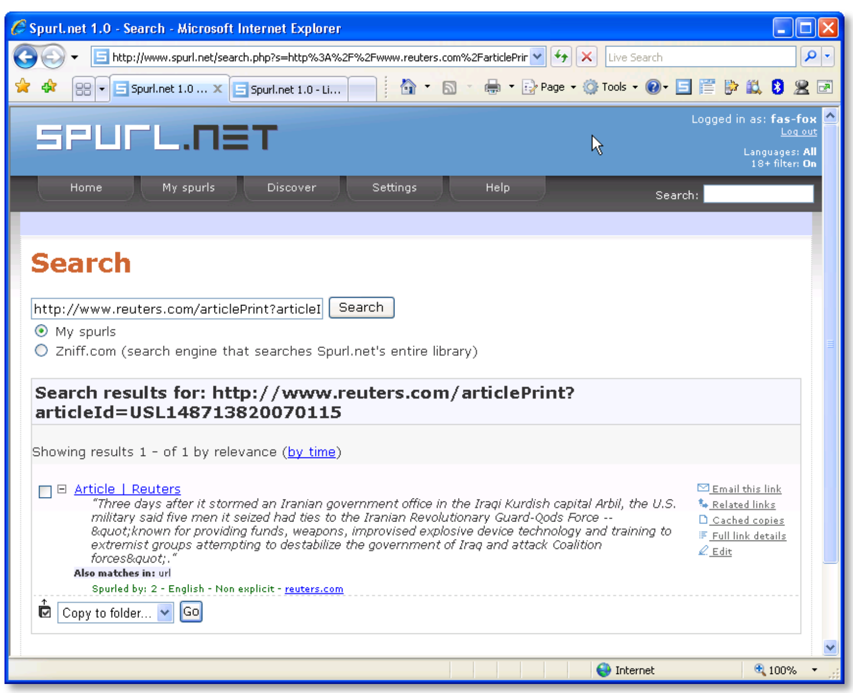


Figure 47 - Task 2 - Getting the First Fragment in Baseline

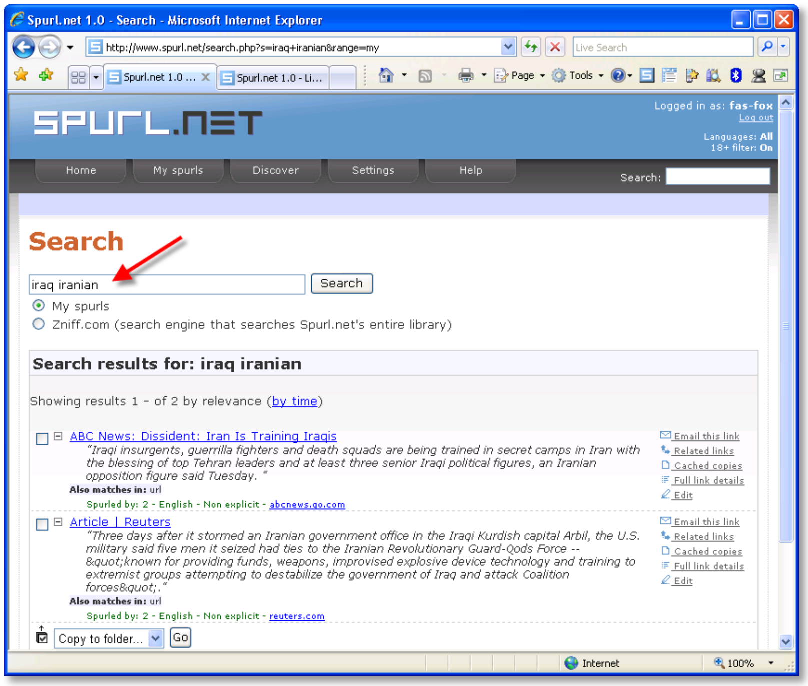


Figure 48 - Task 2 - Searching in Baseline for the Matching Fragment

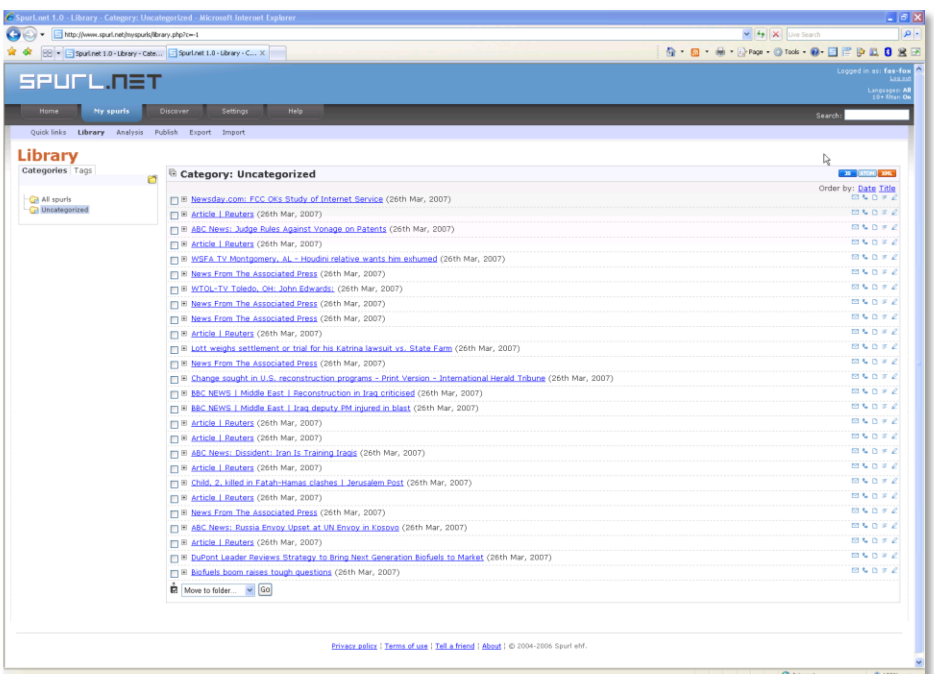


Figure 49 - Task 2 - Browsing in Baseline for the Matching Fragment

Step 2 – Retrieving Matching Fragment for One's Own Fragment – FW (Test Tool)

Each Subgroup used a different subset for this step:

- The Subgroup 1 of Subject Group A used A1-1-2/A2-1-2 pairs
- The Subgroup 2 of Subject Group A used A1-1-1/A2-1-1 pairs
- The Subgroup 1 of Subject Group B used B1-1-2/B2-1-2 pairs
- The Subgroup 2 of Subject Group B used B1-1-1/B2-1-1 pairs

Each subject was first asked to perform 3 matching fragment retrievals using the above subset of pairs:

- used FW to find the first story
    - went to the "Browse" tab
    - pasted the URL in the "Search by URI' box and submitted search
    - identified the two fragments (Figure 50)

After completing their 3 matching fragment retrievals, the subjects were asked to answer the following questions (on 7-point Likert scales):

- Using FW, how easy was it to find the fragments you had previously defined?
- How effective was FW in accomplishing the task of retrieving the fragments?
- How enjoyable was using FW for accomplishing this task?
- How useful is it to you to be able to see two related fragments side by side?
- How useful is it to you to have the ability to navigate from one fragment to the other?

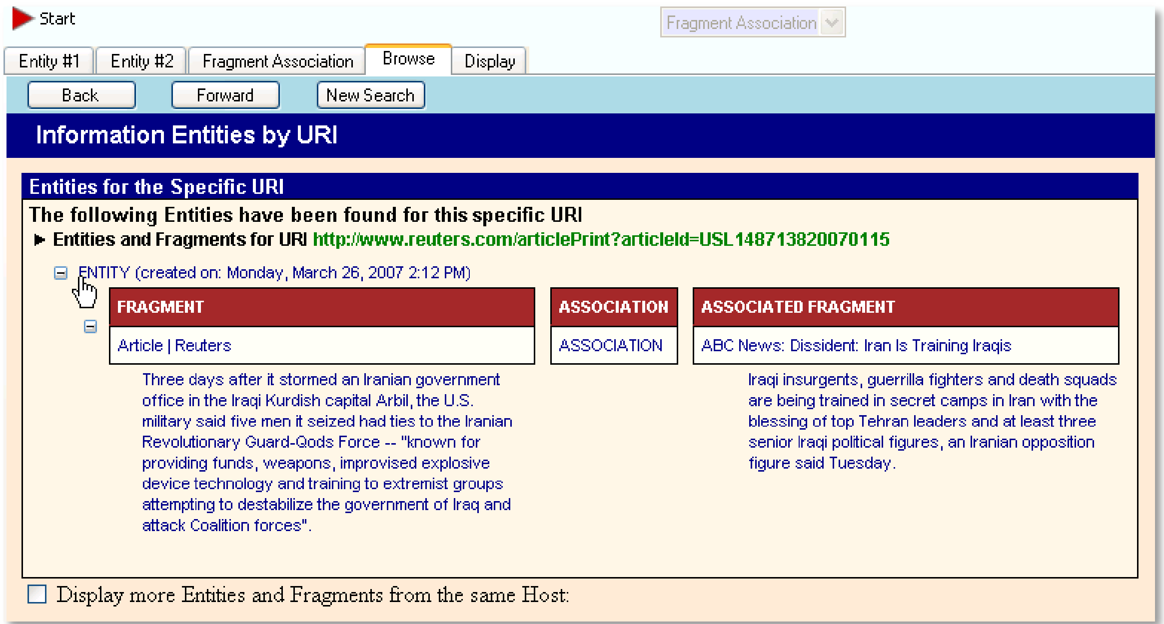


Figure 50 - Task 2 - Retrieving the two Fragments Side-by-Side in FW

Step 3 – Retrieving Matching Fragment for Others' Fragment – SPURL (Baseline)

Each Subgroup used a different subset for this step:

- The Subgroup 1 of Subject Group A used B1-2-1/B2-2-1 pairs
- The Subgroup 2 of Subject Group A used B1-2-2/B2-2-2 pairs
- The Subgroup 1 of Subject Group B used A1-2-1/A2-2-1 pairs
- The Subgroup 2 of Subject Group B used A1-2-2/A2-2-2 pairs

In Step 3 the subjects were asked to perform 3 matching fragment retrievals using a subset created by members of the other Subject Group (i.e. members of Subject Group A used a subset created by members of Subject Group B and vice versa). This means that in this step the subjects performing the retrieval encountered these stories and the Information Fragments defined by other subjects for the first time. This step, as well as Step 4, have been introduced in the design in an effort to provide a way of eliminate the possibility of content recollection.

Moreover, it was meant to ensure that any conclusions derived from this process are also applicable to a collaborative environment.

- used Baseline to find the first story
  - brought up Baseline in the first tab
  - pasted the URL in the Baseline “Search” box and submitted search
  - clicked on the plus “+” icon to display the fragment (Figure 47)
- used Baseline to find the second story
  - brought up Baseline in the second tab
  - used information from the first fragment to find the second fragment either by searching (Figure 48) or by browsing (Figure 49)
  - displayed the second fragment

After completing their 3 matching fragment retrievals, the subjects were asked to answer the following questions (on 7-point Likert scales):

- Using SPURL, how easy was it to find the fragments others had defined?
- How effective was SPURL in accomplishing the task of retrieving the fragments?
- How enjoyable was using SPURL for accomplishing this task?

Step 4 – Retrieving Matching Fragment for Others’ Fragment – FW (Test Tool)

Each Subgroup used a different subset for this step:

- The Subgroup 1 of Subject Group A used B1-2-2/B2-2-2 pairs
- The Subgroup 2 of Subject Group A used B1-2-1/B2-2-1 pairs
- The Subgroup 1 of Subject Group B used A1-2-2/A2-2-2 pairs
- The Subgroup 2 of Subject Group B used A1-2-1/A2-2-1 pairs

In Step 4 the subjects were asked to perform 3 matching fragment retrievals using the above subset of pairs

- used FW to find the B1-2 story
  - went to the “Browse” tab

  - pasted the URL in the “Search by URI’ box and submit search
  - identified the two fragments (Figure 50)

After completing their 3 matching fragment retrievals, the subjects were asked to answer the following questions (on 7-point Likert scales):

- Using FW, how easy was it to find the fragments others had defined?
- How effective was FW in accomplishing the task of retrieving the fragments?
- How enjoyable was using FW for accomplishing this task?
- How useful is it to you to be able to see two related fragments side by side?
- How useful is it to you to have the ability to navigate from one fragment to the other?

3.9.3 Task 3 (Web Searching Test)

Task 3 was designed to test Hypothesis #4, with the consideration that rejection of the null hypothesis would suggest that traditional Web search engines such as Google and Live Search fail to bring together two Information Entities containing two strongly related Information Fragments and that an Information Fragment Association registry is needed to record and supply association information. In addition to testing Hypothesis #4, we also observed the subjects’ searching behaviors in the process, to obtain further insights. The design of this task involved the following activities:

- Interface Selection and Configuration
- Query Construction and Submission (Subject Group C)
- Gathering of Saved Information

3.9.3.1 Interface selection and configuration

Two mainstream Web Search Engines, Google and Live Search, were considered for this task. Because of Google’s popularity, it was originally considered to be a better choice. However, on December 2006 Google decided to limit its support for the standards-based SOAP API because it "was an experimental free program, so the resources available to support the program are limited"[42]. As Google's pre-existing API accounts were still functional, it was evaluated through preliminary testing along with Live Search. A single test interface (Figure 51) was developed allowing testers to submit the same query to the two search engines. The goal of the preliminary testing was to determine the functionality and suitability of the searching interface to our Task 3.

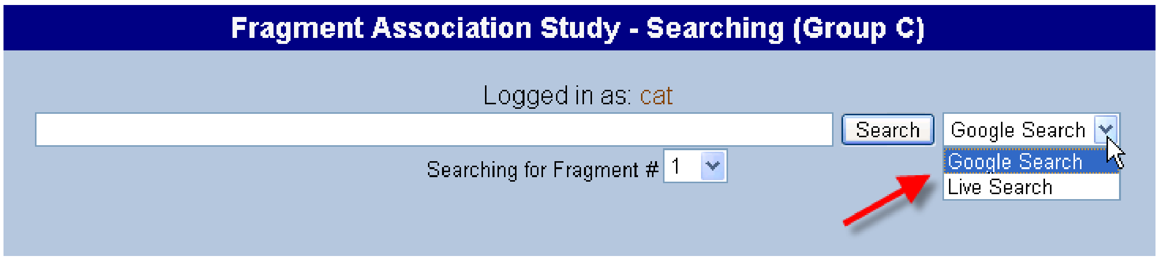


Figure 51 - Preliminary Testing for Search Interface

The preliminary testing concluded that since Google's response was occasionally slower than subjects' tolerance and at the same time Live Search API performed consistently better, we chose Live Search as the search engine for Task 3.

We then enhanced the searching interface with other capabilities. It can record the subjects' browsing activity while they are constructing the optimal search query for retrieving Information Fragment X and it can also record any sites visited by clicking the links in the search results. The Page Visits Interface (Figure 52) and the Browsing History Interface (Figure 53) were designed for the purposes of subsequent result analysis. They were not accessible to the subjects.

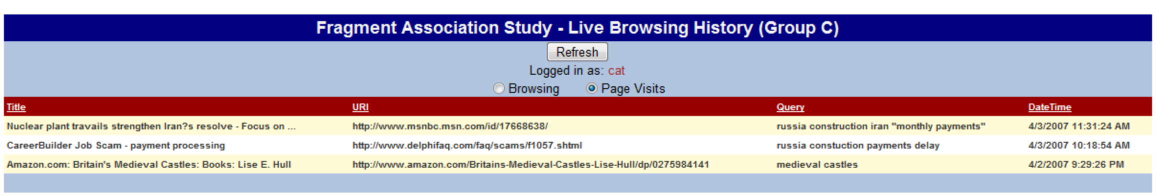


Figure 52 - Page Visits Interface

---

[42] http://code.google.com/apis/soapsearch/api_faq.html

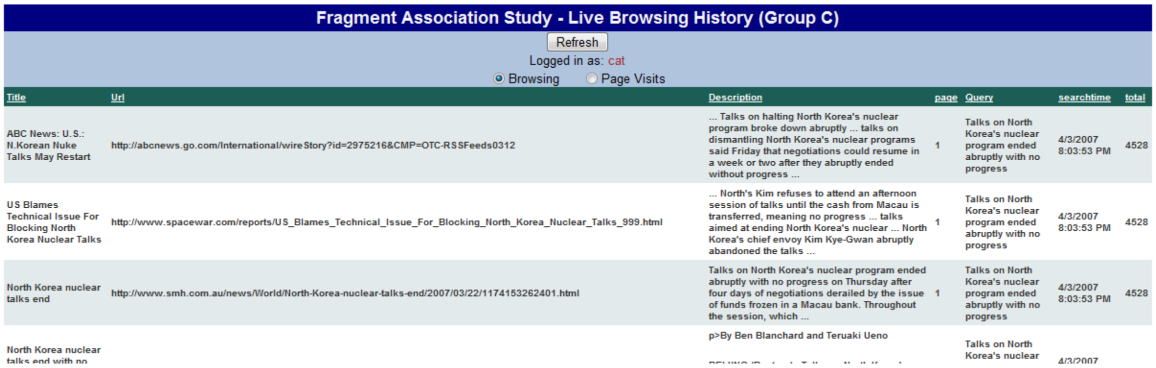


Figure 53 - Browsing History Interface

Figure 54 – Web Searching Interface (Subject Group C)

Figure 54 shows the Web Searching Interface for subjects. The subjects were asked to specify the number of the Information Fragment for which they were searching. This number was being recorded along with every activity in order to assist with the final data processing.

Each result set entry consisted of:

- The title of the story along with a hyperlink to it
- A snippet from the story containing the terms used in the query
- The URI of the story
- The link to the cached page

Subjects used the snippet information to determine whether the retrieved resource contained the targeted Information Fragment. Each subject in Subject Group C was given 24 Information Fragments out of the Information Fragments defined by User Group A and User Group B.

3.9.3.2 Query Construction and Submission (Subject Group C)

As stated, each subject was given 24 fragments from sets A1 and B1, and was asked to devise a query to retrieve each Information Fragment. The subject was told that the query should consist of the terms most likely to retrieve the Information Fragment at hand or an Information Fragment of similar content.

The subjects were asked to try several searches in the web searching interface until they feel comfortable that they have devised a good query. They them submitted the query to the Query Submission Interface immediately. Every search activity and every resource visited were automatically recorded for later analysis.

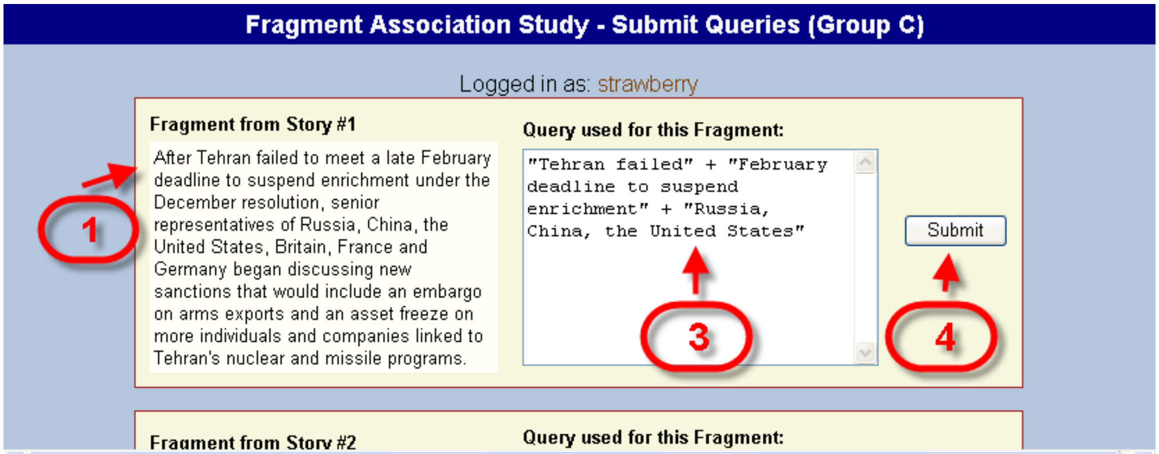


Figure 55 - Query Submission Interface - Typing and submitting query

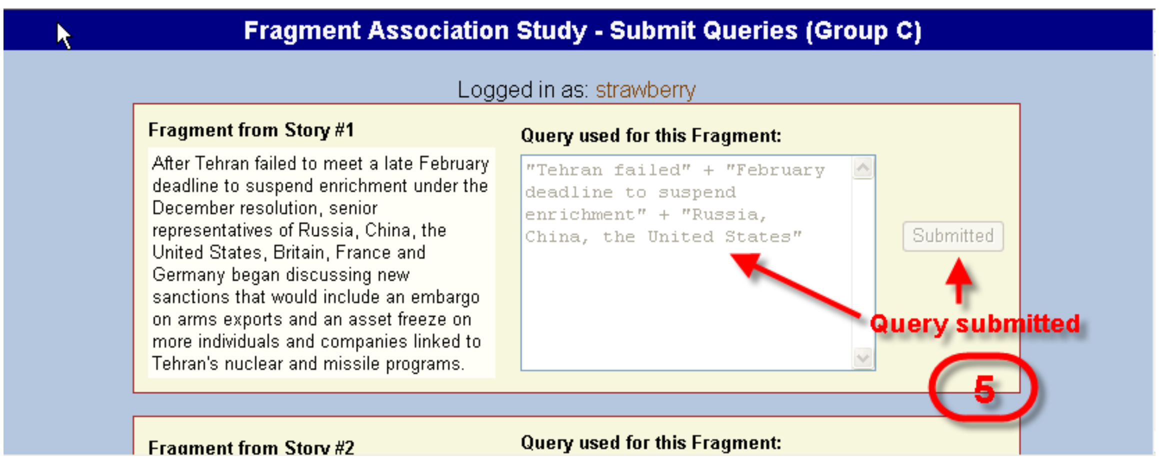


Figure 56 - Query Submission Interface - Query submitted

The entire process of Task 3 is depicted in Figure 55 and Figure 56 and involved the following steps:

1. The subject logged in the Query Submission Interface and read the Information Fragment carefully
2. Logged in the Web Searching Interface and constructed a good query by trying out several word and phrase combinations
3. Went back to Query Submission Interface and typed the query in the allocated box
4. Submitted the query
5. The query was marked as submitted

6. An automated background process started, submitting the query to the Live Search API and retrieving and storing the first 1000 results

Subjects in Group C did not have the benefit of looking at the associated Information Fragments (in sets A2 and B2) so as not to be in any way influenced by the content of those fragments in their selection of keywords for retrieving fragments in sets A and B. The stories used in this task were from all topic categories involved in the study.

The subjects were also not timed or physically observed, besides mechanisms in the interfaces. They were asked to finish the entire task within 48 hours, but extensions were liberally given. The task was designed in such a way that it provided the subjects with the flexibility of choosing whether to perform all of the query submissions at once or gradually according to their schedules.

3.9.3.3 Gathering of Saved Information

Using the saved data sets resulting from the automatic submission of the queries devised by the Subject Group C subjects (step 6 above) the test administrator extract some information. As part of this process, the following items were associated with each query:

- the web page containing the Information Fragment for which the query was devised by a Group C subject – for reference we call this web page P1
- the web page containing the Information Fragment for which an Information Association had been established by a Groups A or Group B subject with the Information Fragment in P1 – for reference we call this web page P2

For each one of the searches the following figures were calculated:

- the total number of web pages retrieved
- the position of P1
- the position P2 (if this web page is at all retrieved by the search)

- The degree of collocation

Task 3 was designed to allow subjects to perform an activity with which they were already very familiar. For this reason, neither direct observation nor detailed instructions were needed. Valuable conclusions were drawn only during the subsequent analysis of the results. This analysis was meant to shed light on the difference between the way Information Entities and Information Fragments are handled by Web search engines. For the purpose of this comparison, we accept that two highly related Information Entities have a short rank difference within the result set of a Web search engine, if this result set is yielded by a query designed to retrieve the one of the two Information Entities. The question this task is trying to find an answer for is: does the same short rank difference occur when we are dealing with Information Fragments instead of Information Entities? Will a query designed to retrieve a specific Information Fragment residing within an Information Entity yield a result set with short rank difference between this Information Entity and one encompassing a highly related Information Fragment? If this is not the case, we can conclude that Web search engines do not handle relationships between Information Fragments very well, so there is justification for the use of an infrastructure for recording and retrieving Information Fragment Associations which can be used to complement a Web search engine's results.

A conceivable Web search engine enhancement would involve the attachment of associated Information Fragments to Web search engine results. A relevance distinction can be made in these attachments between Information Fragment Associations surrounding the search engine result snippet and the rest of the Information Fragment Associations belonging to that Information Entity.

3.9.4 Task 4 (Recall, Precision and F-measure)

This task was undertaken in order to compare the recall and precision of a search query against a dataset containing Information Fragment Associations with that of a search against the same dataset but without Information Fragment Associations.

The task utilized the queries provided by the User Group C subjects. The queries were submitted by the Test Administrator to the Baseline SPURL collection built by all subjects of User Group A and User Group B and to the FW collection built by the same groups of subjects. The total of queries submitted were 6 x 24 - 1= 143 (a fragment pair was not submitted in task 2 because of a broken link issue). The Precision, Recall and F-measure of the searches submitted to the two systems were calculated and compared, revealing the impact of the presence of Information Fragment Association.

This task helps determine whether Information Fragment Association offers any benefits when search queries are performed directly on the content of the Information Fragments. If such benefit exists, more retrieval options and configurations are available in addition to those considered by Task 3. Independent searching in subject specific collections may be desirable. Another possibility may be to provide yet another type of enhancement to web search engine results, this time by submitting the query in parallel to universal Information Fragment Association collections and returning the results alongside the search engine results.

3.10 TIMETABLE

Each Task was administered to each subject by appointment. Two appointments were made with each subject on separate days for each one of the two tasks.

Task 1 and Task 2 (Groups A and B) took place in the last two weeks of March 2007 and Task 3 (Group C) took place in the first two weeks of April 2007. Following that, the data were tabulated, processed and analyzed, and Task 4 was carried out by the Test Administrator. The results are reported in the next chapter.

4 STUDY FINDINGS

The tasks assigned to the subjects and the collected data were organized in such a way as to test the hypotheses for this study, whose headings are presented below again for reference.

- Performance (Hypothesis #1)
- Subject Experience (Hypothesis #2)
- Subject Assessment of the Degree of Usefulness of Fragment Association (Hypothesis #3)
- Fragment Separation within Search Engine Result Sets (Hypothesis #4)
- Precision of Searches (Hypotheses #5-7)

4.1 PERFORMANCE

4.1.1 Success Rate of Identification of Matching Fragment

The subjects were able to successfully identify most of the matching fragments using the two tools. They identified 81.69% of the matching fragments with Baseline and 100% of the matching fragments with FW. In 15.49% of the observations the subjects failed to identify the matching fragment with Baseline. In 2.82% of the observations the subjects gave up. Subjects were told that they could give up after 5 minutes. This option was exercised only in Baseline and only in 2 occasions.

	Observations in which the matching fragment was found		Observations in which the WRONG fragment was found		Observations in which the subject gave up	
	Count	Percentage	Count	Percentage	Count	Percentage
SPURL (Baseline)	58	81.69	11	15.49	2	2.82
FW (Test Tool)	72	100.00	0	0.00	0	0.00

Table 2 - Matching Fragment Identification Success

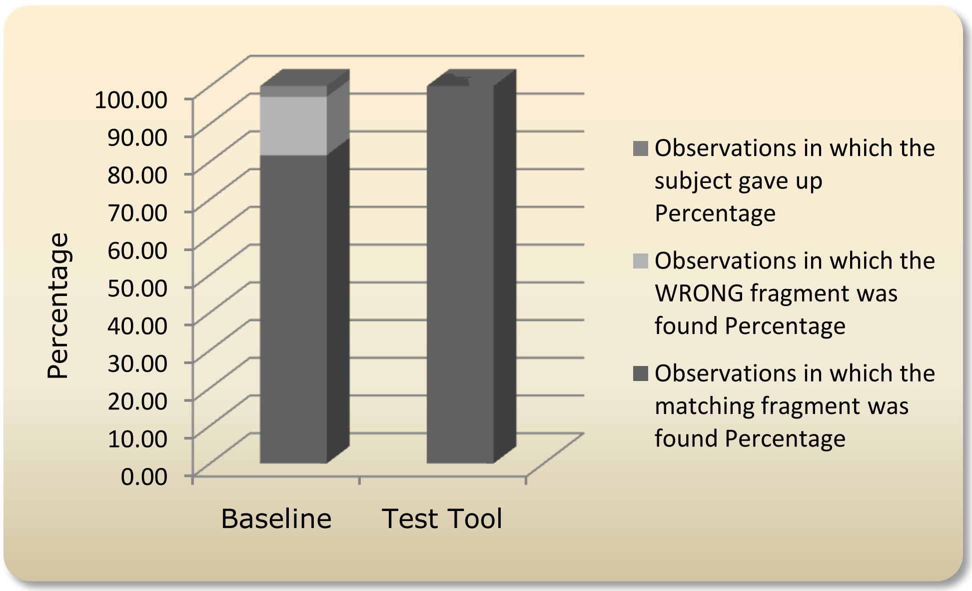


Figure 57 - Matching Fragment Identification Success

A t-test performed on the observations in which the matching fragment was found, revealing that there is a probability of 0.000094373 that the two tools will perform equally. Since we used $\alpha=0.05$ and this figure is much lower, we conclude that the degree of failure with Baseline is significant enough to support the alternative hypothesis, i.e. that Information Fragment Association would improve the retrieval process. The rate of failure with Baseline will be higher in non-laboratory setting because a larger number of bookmarked resources will have more similar but not relevant items, which increases the possibility of selecting the wrong item.

4.1.2 Actions Performed Prior to Identifying the Matching Fragment in Baseline

Because of the lack of Fragment Association capabilities in Baseline, subjects had to use other techniques available to them in order to compensate. The two techniques readily available to them were:

- *Keyword Searching* (Performed in 73.24% of the cases)
- *Browsing* (Performed in 26.76% of the cases)

4.1.2.1 Keyword Searching in Baseline

We use the number of keyword searches performed in Baseline by subjects prior to identifying the matching fragment as one of the indicators of the retrieval difficulty. User productivity suffers when one is forced to perform several searches before finding the matching fragment. Most subjects (73.24%) used keyword searches in Baseline to find the matching fragment. Their queries were based on the first fragment and in most of the cases the full text of the first story.

The number of bookmarked resources used in this Study was lower than the number of bookmarks an average user is expected to have accumulated in a real life scenario. This means that with fewer resources to sift through in pursuit of the matching fragment, the identification of the matching fragment in our Study should have been relatively easy. However, as the results of the count of keyword searches performed in each observation demonstrate, this was not a straightforward process. This puts Baseline in an even more significant disadvantage.

With the reasonable expectation that the process of identifying the matching fragment would have been even more arduous had the number of bookmarked resources been large, the results support the alternative hypothesis that Fragment Association would improve the retrieval process, as it would obviate the need of performing a series of unnecessary keyword searches in order to navigate between two interrelated Information Fragments.

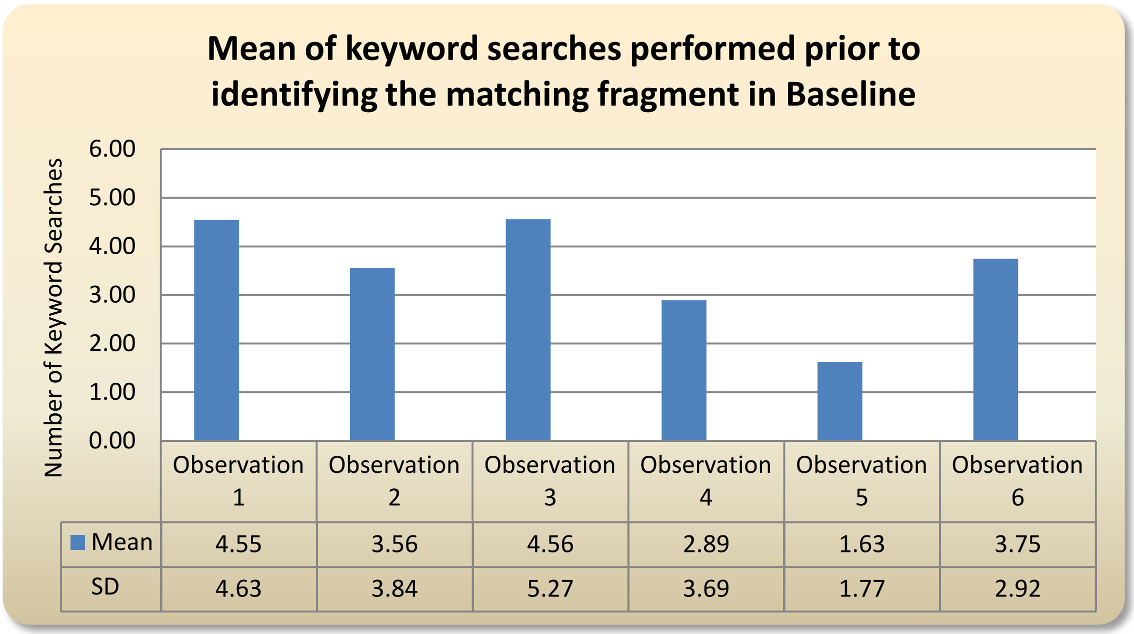


Figure 58 - Keyword Searches Performed

4.1.2.2 Browsing in Baseline

Some subjects opted for using browsing in order to identify matching fragments in Baseline (in 26.76% of the cases). Although this would have been a very inefficient technique in real life because of a large number of bookmarks, the relatively small set of bookmarks in this case allowed browsing to be a viable option. The number of Information Fragments viewed during this process prior to identifying the matching fragment has been recorded and the mean of each observation is presented in Figure 59. Baseline allowed the subjects to scan a list of fragment titles, and selectively open an Information Fragment to view its content prior to making the determination of whether it was the matching fragment. The results indicate that subjects were forced to view the content of several Information Fragments prior to finding the matching fragment, because browsing the titles alone was not sufficient. The more Information Fragments the subjects had already opened and viewed, the fewer they had to view again. This accounts for the fact that Observation 3 (the third observation using their own Information Fragments) and Observation 6 (the third observation using Information Fragments defined by others) required them to open less Information Fragments for viewing.

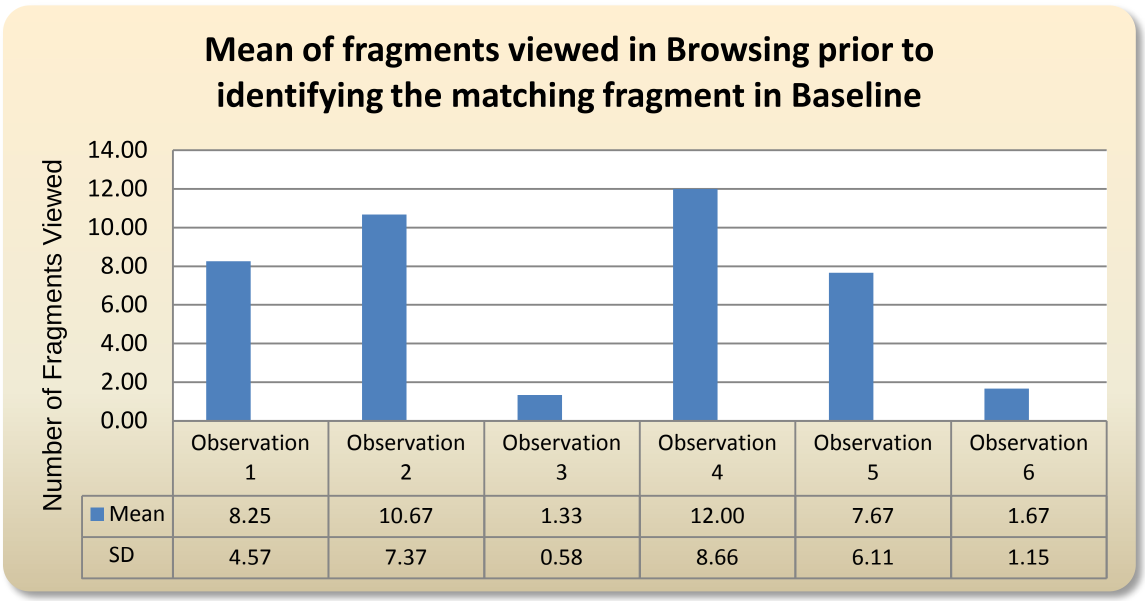


Figure 59 - Fragments Viewed in Browsing

4.1.3 Matching Fragment Viewing

This section reports on data collected merely because they offer an interesting picture of the use of the two tools in this test. The data do not support or refute any of the hypotheses. The task of the subjects was to move from one Information Fragment to the other. The techniques to be employed were left up to them.

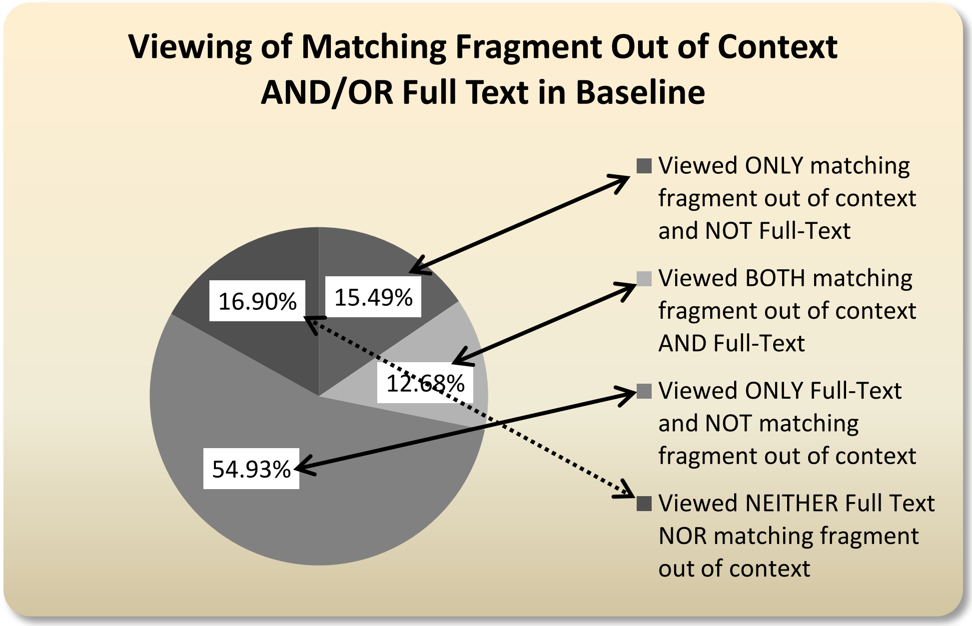


Figure 60 - Matching Fragment Viewing in SPURL (Baseline)

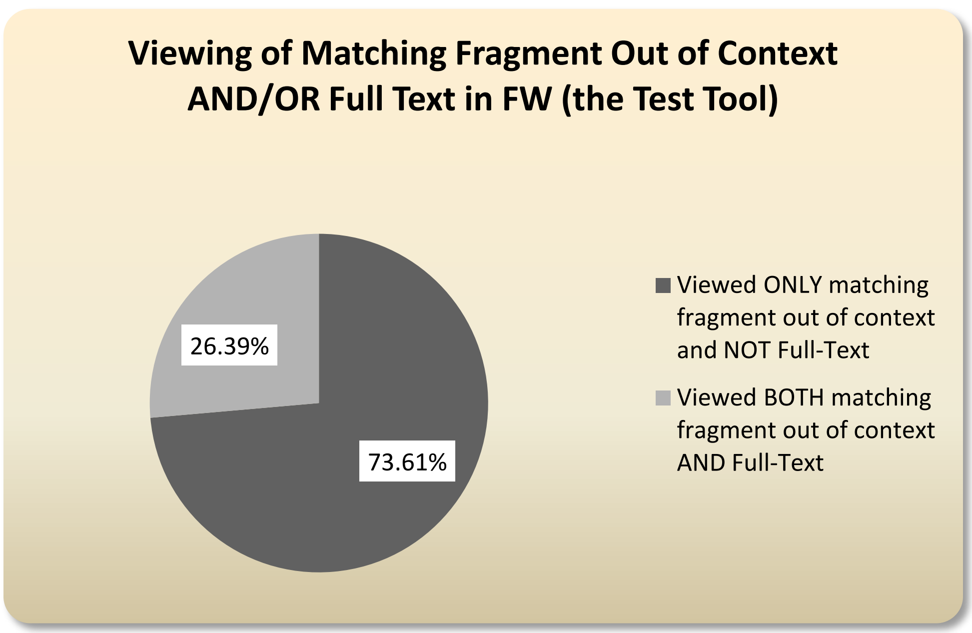


Figure 61 - Matching Fragment Viewing in FW (Test Tool)

During the process of finding the matching fragment, the subjects had the option of looking only at the Information Fragment or bringing up and viewing the entire body of the Information Entity (news story). Some of them viewed both the Information Fragment and

Information Entity. Their actions were counted and the analysis speaks more of the interfaces than of the Fragment Association functionality.

As shown in Figure 60, in order to identify the matching fragment subjects who used Baseline brought up the Full-Text of the story in 67.61% of the observations. In 12.68% of these observations the subjects viewed both the matching fragment out of context and the Full-Text, and in 54.93% of these observations the subjects viewed only the Full-Text without looking at the matching fragment out of context.

Figure 61shows that the Information Fragment was viewed in FW in 100% of the observations. Although the presence of Information Fragment Association makes it not necessary to view the Full-Text for finding the matching fragment, in 26.39% observations, the subjects decided to view the Full-Text of the story. In 13.89% of the cases, they went one step further and clicked on a button allowing them to view the Information Fragment in context.

The results are consistent with what one would expect given the respective interfaces. Since Baseline did not offer Information Fragment Association and subjects needed to find a way of retrieving the matching fragment, they were forced to open the Full-Text, often several times. This was not necessary with FW, because it conveniently presented the two Information Fragments side-by-side. This explains the performance differences between the two tools. Viewing the matching fragment in context was not an option in Baseline. This may have been a factor for those few subjects in their expression of the degree to which they enjoyed the process.

4.1.4 Task 1 Completion Times

In Task 1 the subjects were given two related Information Entities and asked to identify and submit related Information Fragments using Baseline and FW. In the case of FW, they were at the same time establishing a Fragment Association between the two Information Fragments.

By Fragment Pair		
Fragment Pair	**Baseline Mean**	**Test Tool Mean**
1	33.55	25.00
2	29.62	27.11
3	36.26	24.94
4	52.52	21.02
5	31.11	21.57
6	30.75	21.60
7	24.39	18.49
8	29.50	19.12
9	26.26	20.14
10	34.91	20.99
11	23.99	20.81
12	24.24	19.09

.

By Subject		
Subject	**Baseline Mean**	**Test Tool Mean**
Group A - 1st Subject	19.77	13.31
Group A - 2nd Subject	37.93	27.88
Group A - 3rd Subject	25.93	12.60
Group A - 4th Subject	40.79	28.34
Group A - 5th Subject	40.25	22.25
Group A - 6th Subject	27.93	18.91
Group B - 1st Subject	24.81	27.02
Group B - 2nd Subject	15.04	11.19
Group B - 3rd Subject	23.31	17.38
Group B - 4th Subject	38.14	23.86
Group B - 5th Subject	51.99	30.95
Group B - 6th Subject	31.21	26.18

.

t	5.220767
P(T<=t) one-tail	**0.000143**

Table 3 - Task 1 Completion Times

It was expected that there would be no difference between the completion times of the two tools because the process was fairly similar. However, the results are consistently displaying better completion times with FW, in spite of the fact that FW performed an additional function, that of Fragment Association.

A Paired t-Test for the means of each observation was used to test the hypothesis, as shown in Table 3.

The null hypothesis in this case is that the time needed for defining and submitting the two Information Fragments would not be different using a Fragment Association tool than using Baseline. In other words, the null hypothesis is that the difference of the means would be zero. After calculating t and the one-tail value of p for $\alpha=0.05$, we find out that p is 0.000143 which is much smaller than 0.05, so the null hypothesis is rejected.

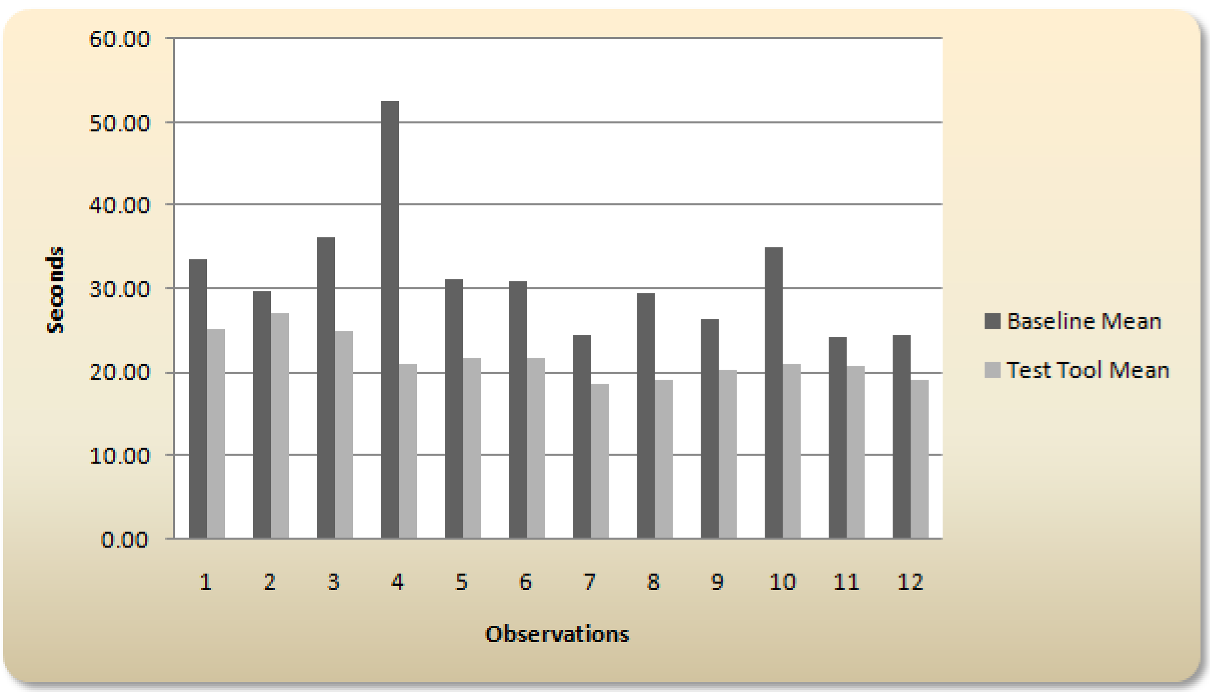


Figure 62 - Task 1 Completion Times

4.1.5 Task 1 Completion Times – By Topic

The stories given to the subjects were in 4 basic topic areas:

- International (mostly covering the Iran nuclear issue)
- Events in Iraq
- U.S. Politics
- Miscellaneous News

The analysis in this section is performed in terms of these topics. Because the time spent reading the stories (i.e. the time prior to the determination of which two Information Fragments to use) was not being counted, what we see here is the extent to which the complexity, or more generally the content of the story, has affected the mechanics of fragment definition and submission.

By Topic		
Topic	SPURL (Baseline) Mean	FW (Test Tool) Mean
International	33.14	25.68
Iraq	38.12	21.40
U.S. Politics	26.72	19.25
Misc. News	27.71	20.29

Table 4 - Task 1 Completion Times - By Topic

Moreover, we observe that in the case of FW, times for all topics were approximately the same with the exception of International. The reason might be because the International stories were the first assigned to each subject, when the subject was still in the process of familiarizing himself or herself with the process. The same longer time for International stories in Baseline might be because of the same reason too. When it comes to the rest of the topics in Baseline however, we cannot fail to notice that the figure for stories about events taking place in Iraq is significantly higher. One possible interpretation is that the stories about events in Iraq were more repetitive, more unpleasant and more complex on account of the dense narration of war events and mention of multitude of unfamiliar places. We can therefore surmise that this kind of complexity has affected the process of Information Fragment definition and submission and that this happened only in Baseline.

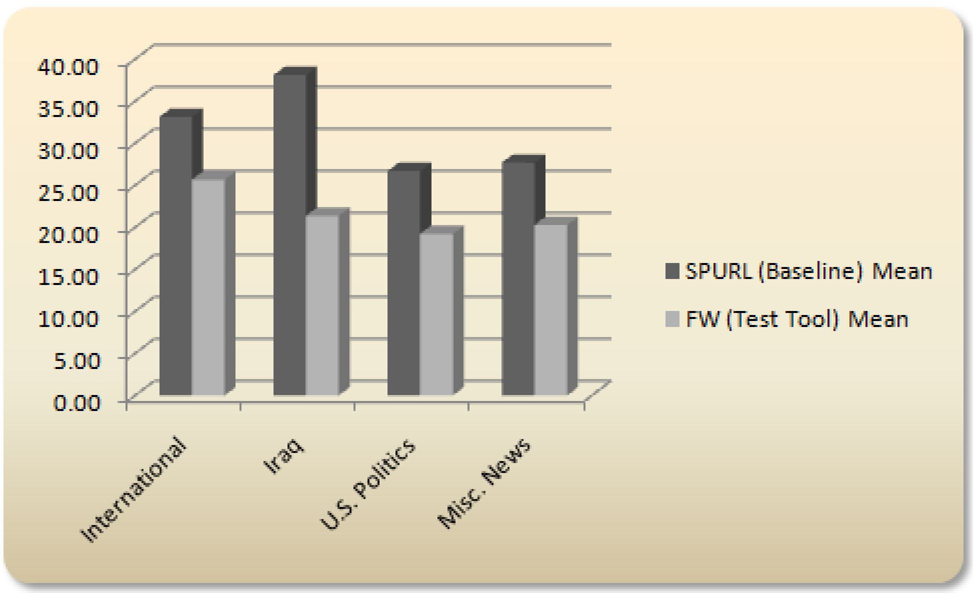


Figure 63 - Task 1 Completion Times - By Topic

4.1.6 Task 2 Completion Times

In Task 2, the subjects were given Information Fragments along with their encompassing Information Entities and were asked to find the matching Information Fragment and Information Entity. They had to match 12 pairs altogether, 6 using Baseline and 6 using FW. Of those 6 pairs, 3 pairs were the Information Fragments previously defined by the same subject, and the other 3 pairs were defined by somebody else. The data analyzed here is related to Hypothesis #1.

4.1.6.1 All Subjects - Own Fragments

The subjects were first given the Information Fragments they had defined themselves

Fragment	SPURL (Baseline) (seconds)	FW (Test Tool) (seconds)	t	P(T<=t) one-tail
First Own Fragment	273.45	110.08		
Second Own Fragment	152.75	35.37		
Third Own Fragment	126.85	19.03		
All Own Fragments	184.35	54.83	3.63415765	**0.00196399**

Table 5 - Task 2 Completion Times - All Subjects

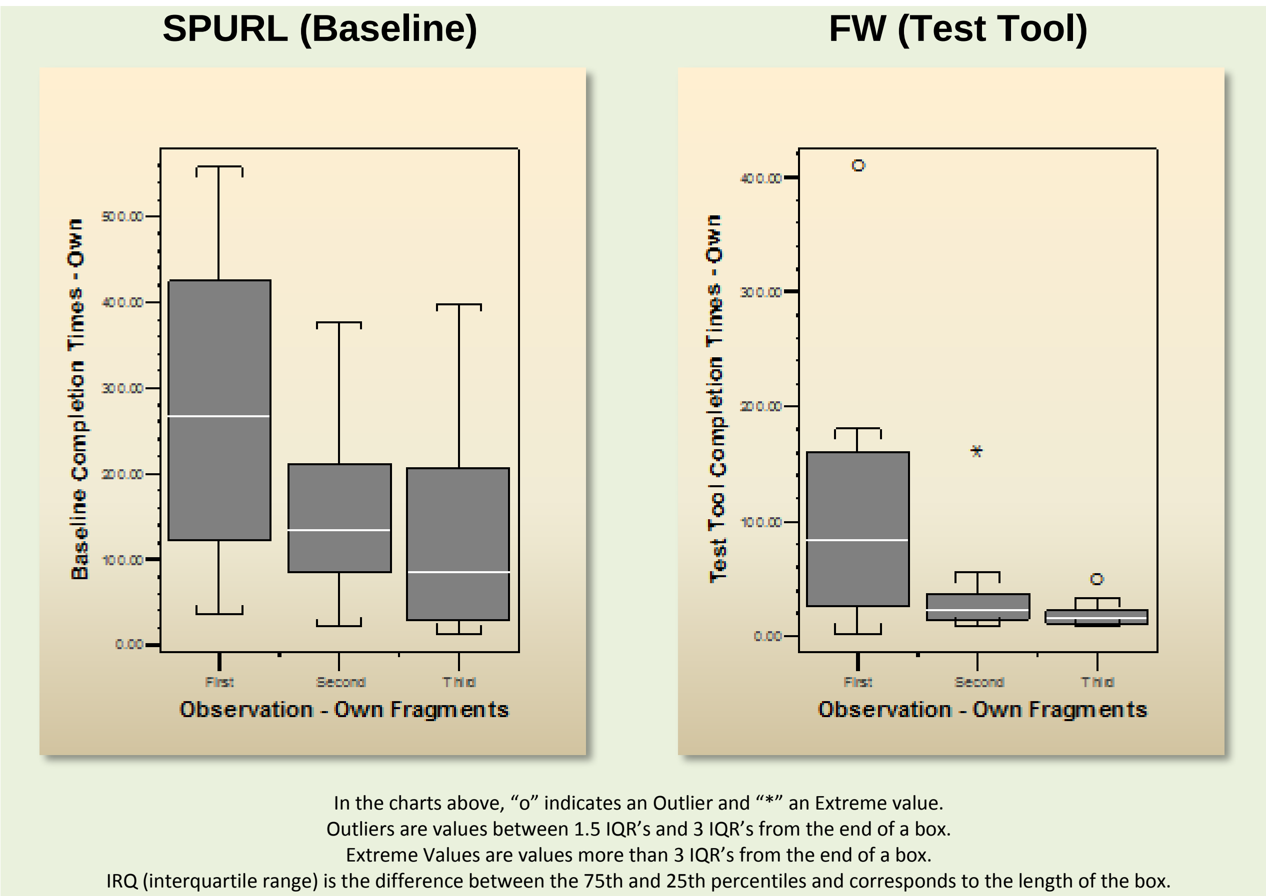


Figure 64 - Task 2 Completion Times - All Subjects - Own Fragments

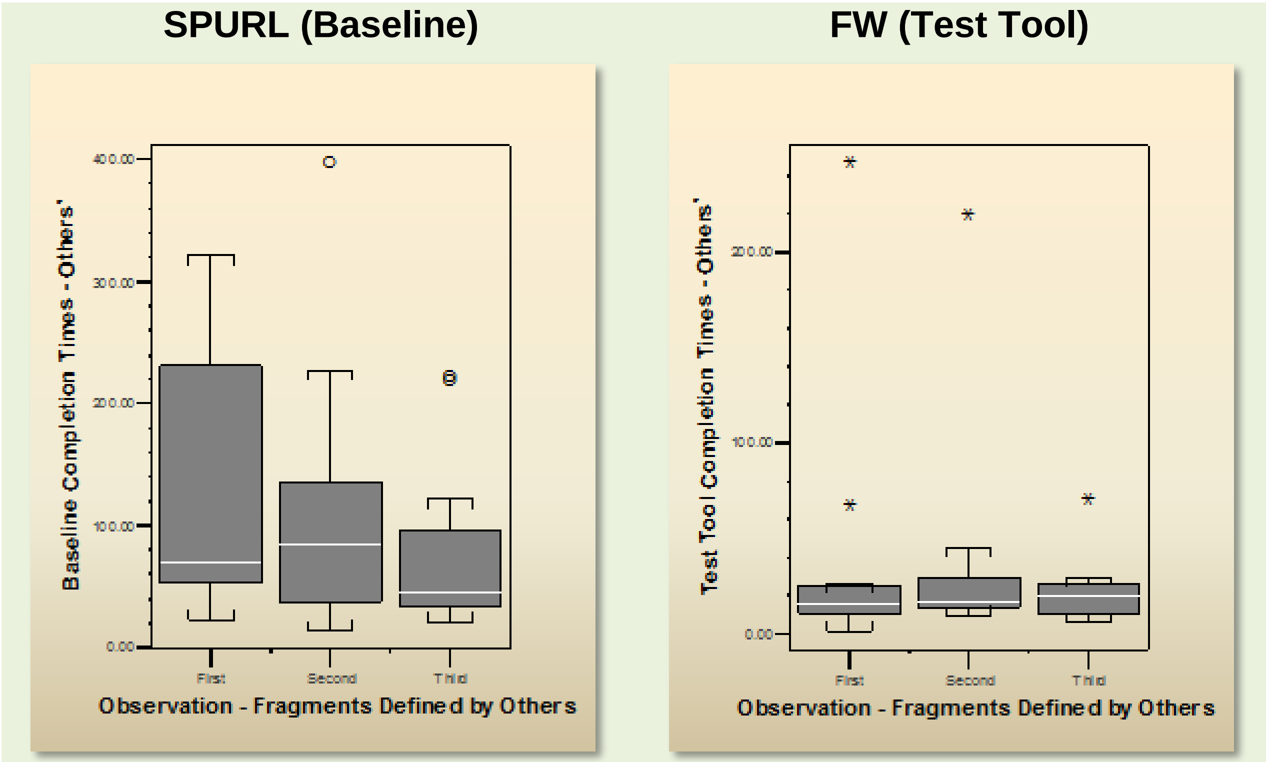
The results, displayed in Table 5, indicate that the subjects were able to perform their assigned task significantly faster in FW than in Baseline. A Paired t-Test for the means of each observation was used to test the hypothesis. The null hypothesis in this case was that the time needed for retrieving an associated Information Fragment would not be shorter using a fragment association tool than using Baseline. In other words, the null hypothesis was that the difference of the means is zero. After calculating t and the one-tail value of p for α=0.05, we find out that p is smaller than 0.05, so the null hypothesis is rejected. This performance difference can also be seen in Figure 64 which presents the value ranges for the completion times of the process of retrieval of the Information Fragments subjects had previously defined themselves. The two plots show that completion times in both tools decrease as the subject moves from the first Information Fragment to the second and then the third. We also see that for each observation, the completion time with Baseline (SPURL) is longer than with the test tool (FW). In the case of FW, we also observe two outliers (marked with a "o" on the plot) and one extreme value (marked with an "x" on the plot), which may suggest that in those particular instances the subject encountered some difficulties which were not characteristic of the performance of the tool in general. Content difficulties and the subject's careless handling of the timer contributed to these outliers.

4.1.6.2 All Subjects - Fragments Defined by Others

The subjects were then given the Information Fragments others had defined.

Fragment	SPURL (Baseline) Mean	FW (Test Tool) Mean	t	P(T<=t) one-tail
First Others' Fragment	123.82	38.72		
Second Others' Fragment	114.62	38.67		
Third Others' Fragment	78.51	21.90		
All Others' Fragments	105.39	32.94	3.7496907	**0.00160582**

Table 6 - Task 2 Completion Times - All Subjects – Fragments Defined by Others



In the charts above, "o" indicates an Outlier and "*" an Extreme value.
Outliers are values between 1.5 IQR's and 3 IQR's from the end of a box.
Extreme Values are values more than 3 IQR's from the end of a box.
IRQ (interquartile range) is the difference between the 75th and 25th percentiles and corresponds to the length of the box.

Figure 65 - Task 2 Completion Times - All Subjects - Fragments Defined by Others

This part was meant to simulate the real-life use of shared social bookmarking. As expected, the first observation took longer than the other observations, as subjects used some time to familiarize themselves with the retrieval functionalities of the two tools. The rest of the observations, as Figure 65 reveals, are more in line with each other and they demonstrate a consistently significant difference between Baseline and FW. This is the case in spite of the fact that the number of bookmarks used in this test was deliberately kept very low so as to ensure that subjects might not face insurmountable difficulties using Baseline. The difference demonstrated here would have been substantially larger had a larger number of resources been bookmarked.

The test attempted to determine whether a subject would be able to retrieve Information Fragments s/he had previously defined more easily than Information Fragments defined by somebody else. Contrary to what was expected, completion times for one's own Information Fragments were longer than completion times for others' Information Fragment. Three factors seem to have contributed to this outcome. First, there was a time lag (usually two days) between Task 1 (the definition of the Information Fragments) and Task 2 (the retrieval). Although the time lag sounds small, subjects probably still did not recollect very well the content details of the stories. Second, the content was not distinctive enough so as to be particularly memorable. Third, any potential benefits of recollection may have been counteracted by the fact that the subjects were ask to retrieve their own Information Fragments before retrieving Information Fragments defined by others. This means that they were increasingly becoming more familiar with the process by the time they got to retrieving others' Information Fragments.

We had three outliers (marked with a "o" on the plot) with Baseline and four extreme values (marked with an "x" on the plot) in FW, which may suggest that in those particular instances the subject encountered some difficulties which were not characteristic of the performance of the tools in general. Content difficulties and the subject's careless handling of the timer contributed to these outliers.

4.1.6.3 Female Subjects – Own Fragments

Since female subjects seemed to approach the test differently than male subjects, a separate data analysis (see Table 7 and Figure 66) was performed in order to shed some light on these differences. We observed through our Study that female subjects generally approached the test with more maturity, spending more time in the beginning to familiarize themselves with the functionality of the two tools and with the contents of the resulting screens. We therefore see a more smooth progression in each observation, which reveals performance improvement in both tools. However, performance was consistently better with FW. A t-test performed on the all observations revealed that there is a probability of 0.01916362 that the two tools will perform equally. Since we used $\alpha=0.05$ and this figure is lower, we conclude that there is significant performance difference between the two tools when used by female subjects in retrieving their own Information Fragments.

Female Subjects				
Fragment	**SPURL (Baseline) Mean**	**FW (Test Tool) Mean**	**t**	**P(T<=t) one-tail**
First Own Fragment	325.88	127.60		
Second Own Fragment	169.86	30.95		
Third Own Fragment	61.93	17.55		
All Own Fragments	185.89	58.70	2.79259454	**0.01916362**

Table 7 - Task 2 Completion Times - Female Subjects

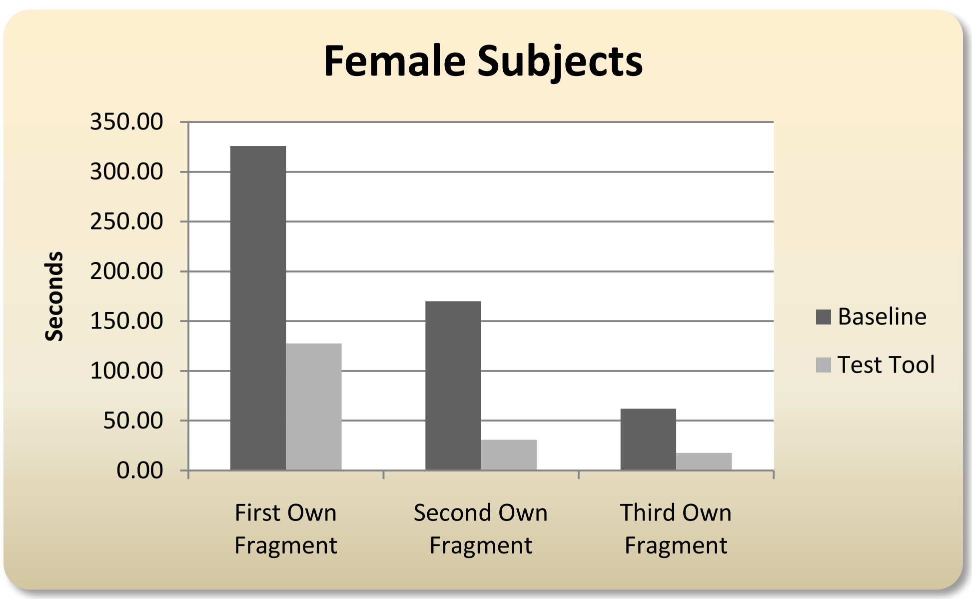


Figure 66 - Task 2 Times - Female Subjects – Own Fragments

As Figure 66 demonstrates, female subjects exhibited that they were able to improve their performance as they were becoming more familiar with the tools and to retain the memory of their own fragments.

4.1.6.4 Female Subjects – Fragments Defined by Others

A similar picture was observed in the case of Information Fragments defined by others. Times were progressively improving with FW having a significant advantage over Baseline throughout the test. A t-test performed on the all observations revealed that there is a probability of 0.0251785 that the two tools will perform equally. Since we used α=0.05 and this figure is lower, we conclude that there is significant performance difference between the two tools when used by female subjects in retrieving Information Fragments defined by others.

Female Subjects				
Fragment	SPURL (Baseline) Mean	FW (Test Tool) Mean	t	P(T<=t) one-tail
First Others' Fragment	160.50	16.56		
Second Others' Fragment	87.60	22.97		
Third Others' Fragment	47.87	19.61		
All Others' Fragments	98.66	19.71	2.56472097	**0.0251785**

Table 8 - Task 2 Completion Times - Female Subjects

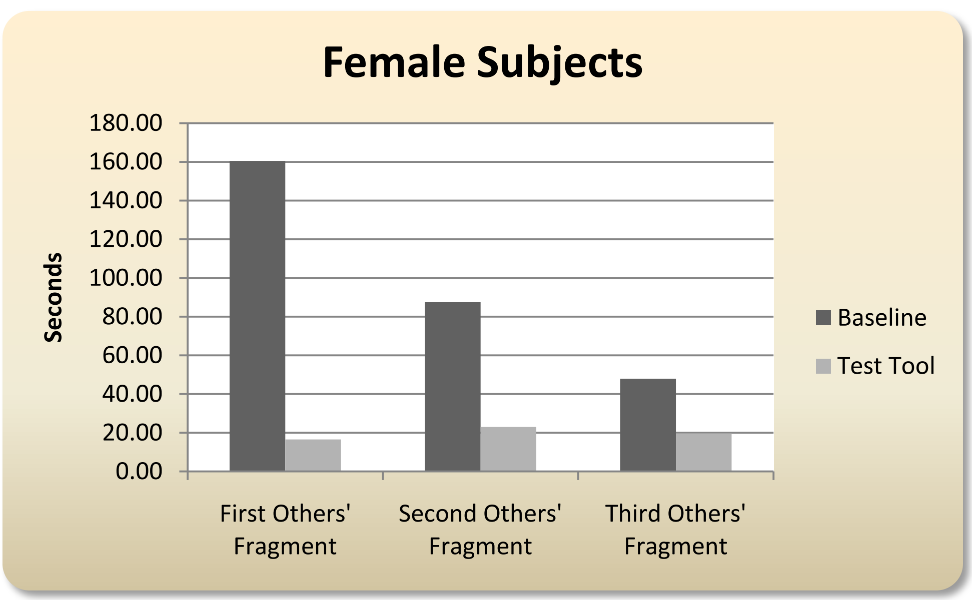


Figure 67 - Task 2 Times - Female Subjects - Fragments Defined by Others

4.1.6.5 Male Subjects – Own Fragments

The results from male subjects were more erratic. We do not see the same kind of smooth progression in each observation that we have seen in the case of the female subjects. A t-test performed on the all observations revealed that there is a probability of 0.03833123 that the two tools will perform equally. Since we used α=0.05 and this figure is lower, we conclude that there is significant performance difference between the two tools when used by male subjects in retrieving their own Information Fragments.

Male Subjects				
Fragment	**SPURL (Baseline) Mean**	**FW (Test Tool) Mean**	**t**	**P(T<=t) one-tail**
First Own Fragment	221.02	92.55		
Second Own Fragment	135.65	39.80		
Third Own Fragment	191.77	20.51		
All Own Fragments	182.81	50.95	2.22481901	**0.03833123**

Table 9 - Task 2 Completion Times - Male Subjects

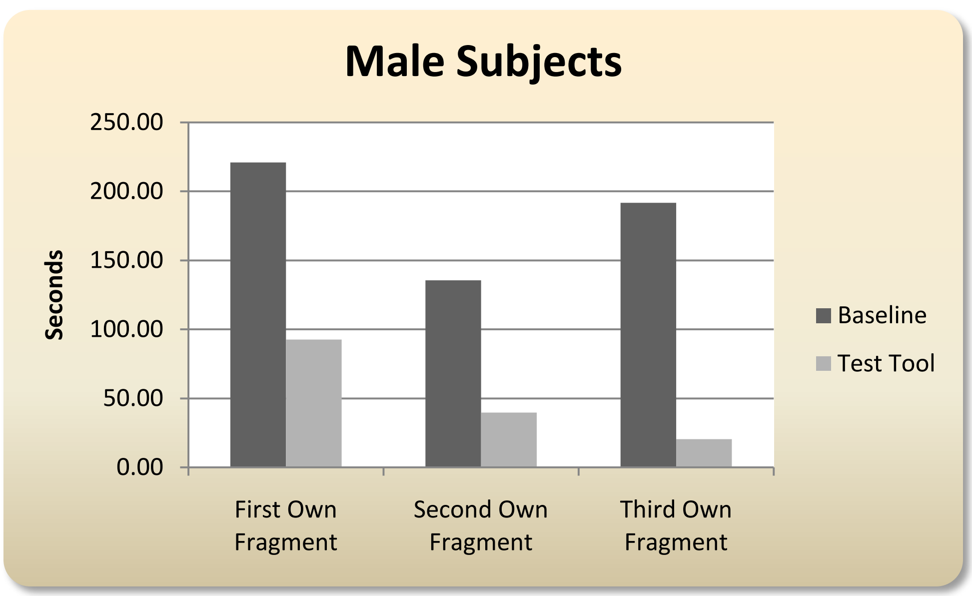


Figure 68 - Task 2 Completion Times - Male Subjects – Own Fragments

4.1.6.6 Male Subjects – Fragments Defined by Others

Male Subjects				
Fragment	**SPURL (Baseline) Mean**	**FW (Test Tool) Mean**	**t**	**P(T<=t) one-tail**
First Others' Fragment	87.14	60.89		
Second Others' Fragment	147.05	57.52		
Third Others' Fragment	109.15	24.19		
All Others' Fragments	112.53	46.94	2.55223257	**0.02556352**

Table 10 - Task 2 Completion Times - Male Subjects

Only one fact remains consistently true in the data gathered from male subjects: the FW times are significantly lower that the corresponding Baseline times. A t-test performed on the all observations revealed that there is a probability of 0.02556352 that the two tools will perform equally. Since we used α=0.05 and this figure is lower, we conclude that there is significant performance difference between the two tools when used by male subjects in retrieving Information Fragments defined by others

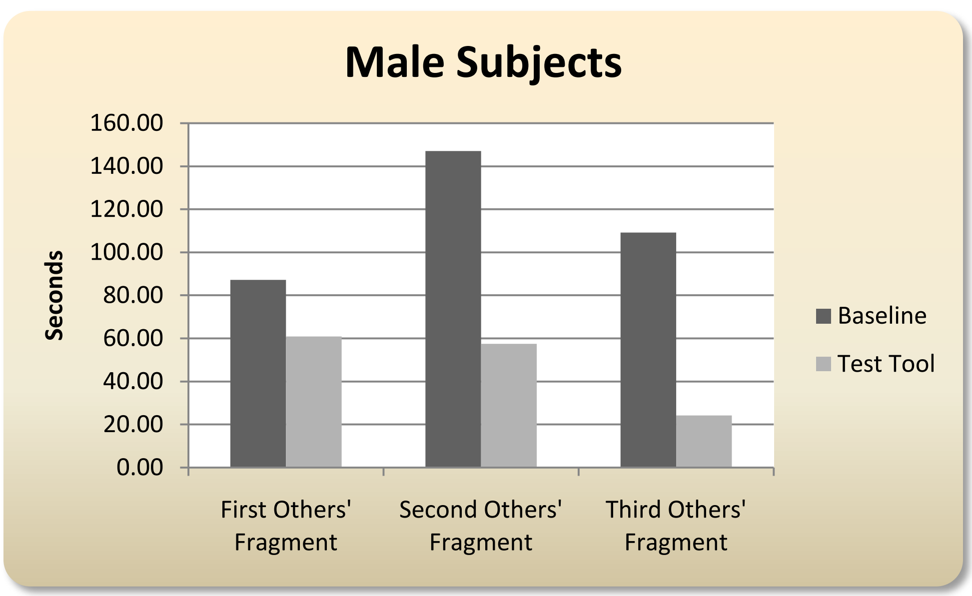


Figure 69 - Task 2 Completion Times - Male Subjects - Fragments Defined by Others

4.1.6.7 Male – Female Subject Comparison

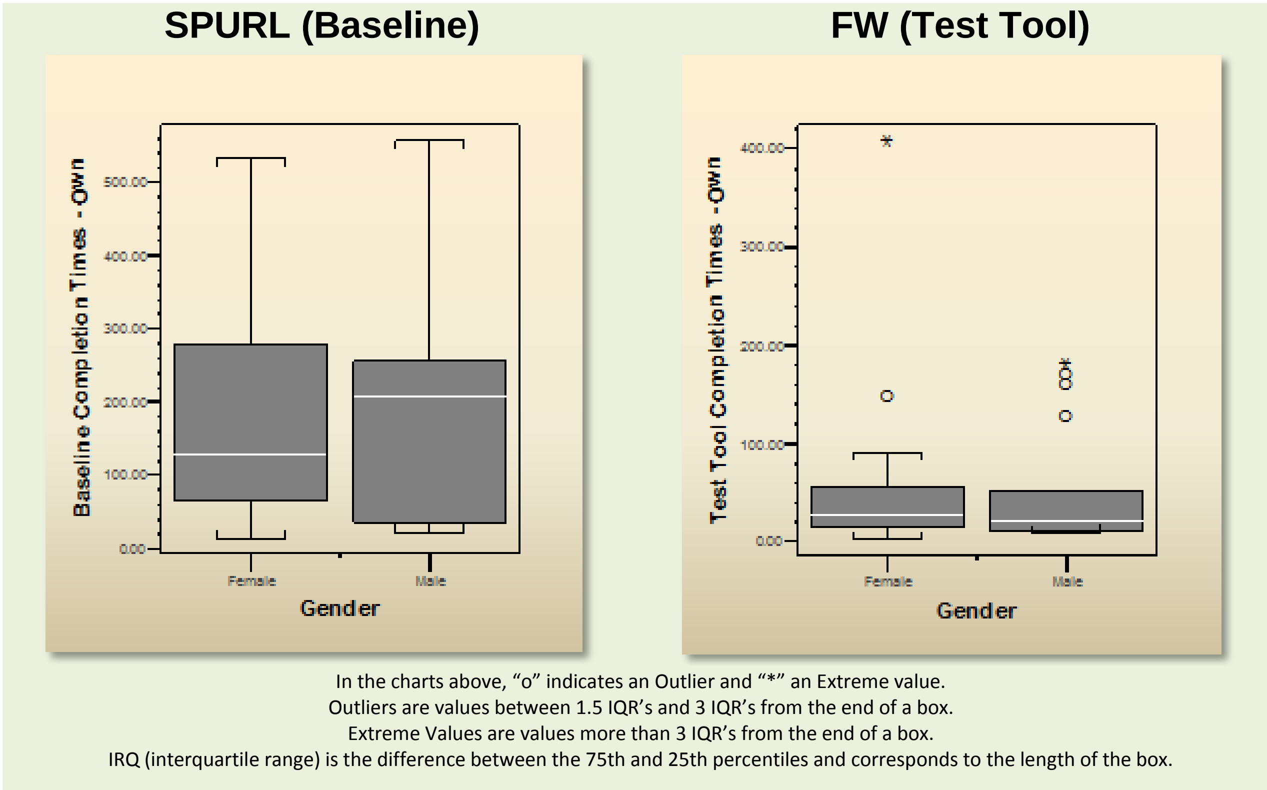
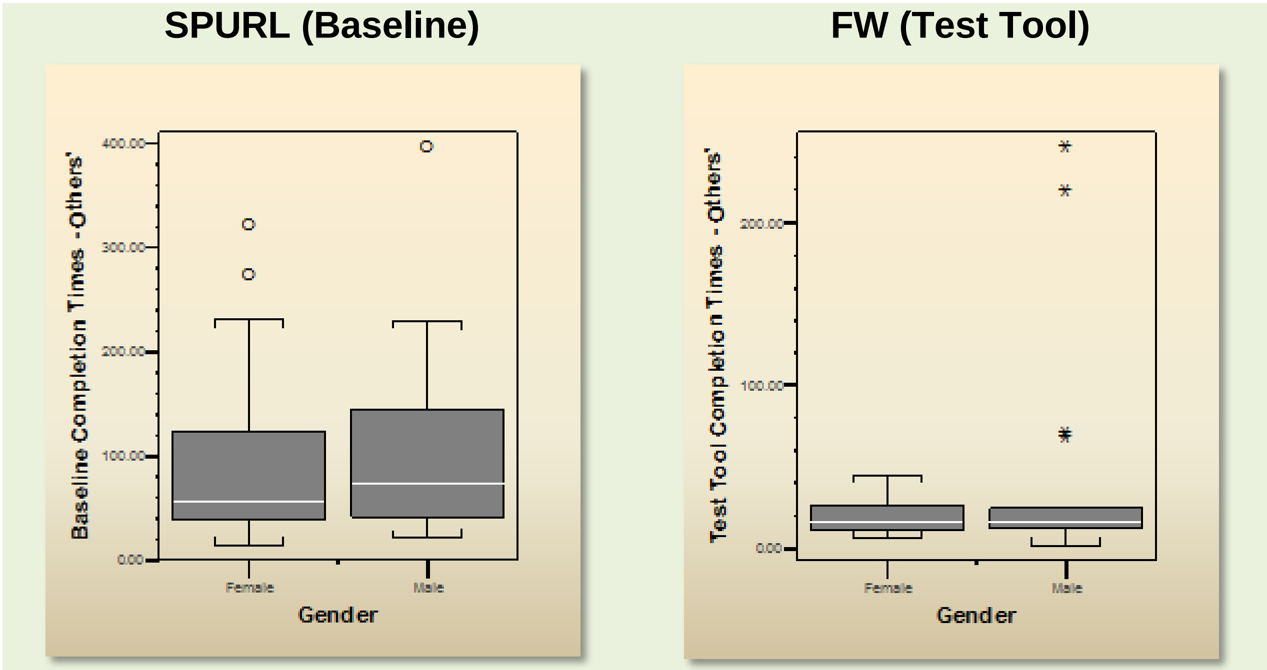


Figure 70 - Task 2 Completion Times - All Subjects - Own Fragments

There were more outliers and extreme readings among male subjects than among female. The outliers and extreme readings in the case of FW (the Test Tool) belong to a single male subject who was failing to observe all of the contents of the result pages and was moving quickly around the interface until he was finally successful in viewing the matching Information Fragment.



In the charts above, “o” indicates an Outlier and “*” an Extreme value.
Outliers are values between 1.5 IQR’s and 3 IQR’s from the end of a box.
Extreme Values are values more than 3 IQR’s from the end of a box.
IRQ (interquartile range) is the difference between the 75th and 25th percentiles and corresponds to the length of the box.

Figure 71 - Task 2 Completion Times - All Subjects – Fragments Defined by Others

4.2 FRAGMENT SEPARATION WITHIN SEARCH ENGINE RESULT SETS (TASK 3)

In Task 3, a group of 6 subjects (Group C) were each presented with 24 Information Fragments and asked to construct, test and submit queries retrieving the Information Entities containing those Information Fragments (one of the subjects was given only 23 fragments). Each one of these Information Fragments was the first part of a pair of Information Fragments previously selected and associated by Groups A and B. The objective of this test was to determine:

- whether the queries devised by the members of Group C yielded the Information Entity containing the matching Information Fragment (the second part of the pair)
- in the cases in which the matching Information Fragment was retrieved, the distance between the Information Entities containing the two Information Fragments within the result set

4.2.1 Retrieval and 'Specific' Queries

Among the 143 queries submitted by Group C, only 25, or 17.5% yielded results containing the second Information Fragment. A possible interpretation of this outcome is that the 118 queries whose result sets did not contain the matching fragment were primarily 'specific' queries, i.e. queries yielding smaller result sets therefore they give much less room to contain the matching fragment than more generic queries. Further analysis of the data supports this interpretation.

Number of Queries with Result Length between 1-50	58
Number of Queries with Result Length between 50-100	13
Number of Queries with Result Length between 100-150	7
Number of Queries with Result Length between 150-200	6
Number of Queries with Result Length between 200-250	4
Number of Queries with Result Length between 250-500	9
Number of Queries with Result Length between 500-1000	7
Number of Queries with Result Length > 1000	14

Table 11 - Result Set Length – Matching Fragment NOT Retrieved

As Table 11 and Figure 72 demonstrate, 71 of queries yield short result sets (less than 100). Smaller result sets may explain the absence of the matching resource. However, there are still 47 queries with result sets larger than 100, which we cannot consider to be very 'specific'. The search engine simply fails to retrieve the matching fragment which supports the fourth alternative hypothesis.

Whether the presence of matching fragments within search engine results is useful to users or not is something to be determined by future studies. The current study clearly demonstrates the inadequacies of search engines regarding Information Fragments, which provides justification for such future studies. Of course, such studies will be possible only if some solution is introduced. If Information Fragment Associations are stored in a public registry, that information can be used to easily insert the matching fragment within a search engine's results.

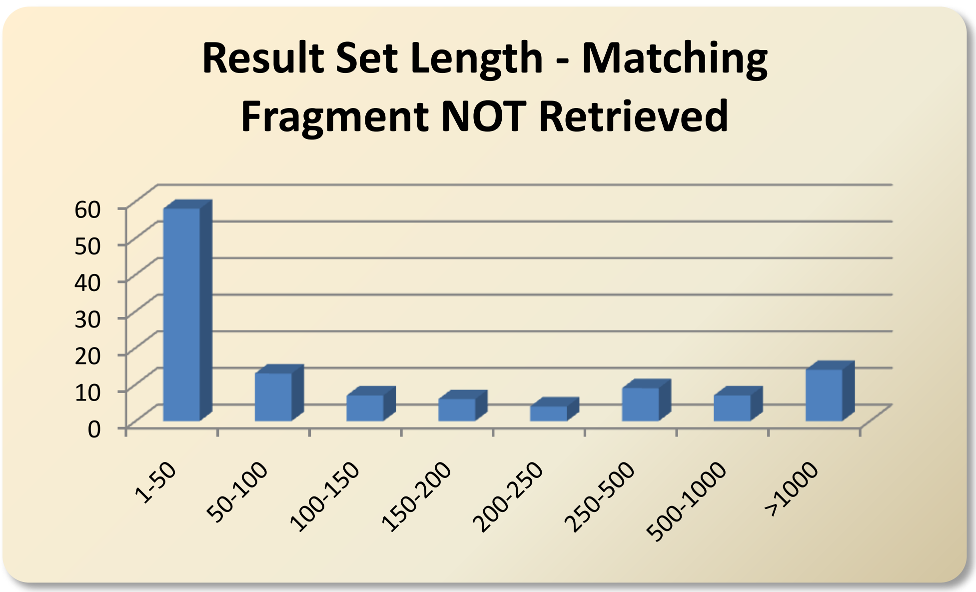


Figure 72 - Result Set Length – Matching Fragment NOT Retrieved

It should be noted that in the case of queries yielding result sets larger than 1000, the above results are inconclusive because for practical purposes only the first 1000 items were retrieved and examined.

Also as expected, 20 out of the 25 queries which yielded result sets containing the matching Information Entity have result sets larger than 100, as Table 12 and Figure 73 clearly demonstrate.

Number of Queries with Result Length between 1-50	4
Number of Queries with Result Length between 50-100	1
Number of Queries with Result Length between 100-150	1
Number of Queries with Result Length between 150-200	0
Number of Queries with Result Length between 200-250	2
Number of Queries with Result Length between 250-500	1
Number of Queries with Result Length between 500-1000	4
Number of Queries with Result Length > 1000	12

Table 12 - Result Set Length – Matching Fragment Retrieved

Again it should be noted that all 25 of these result sets contained the matching Information Entity within the first 1000 items because that was the extent to which result sets were examined.

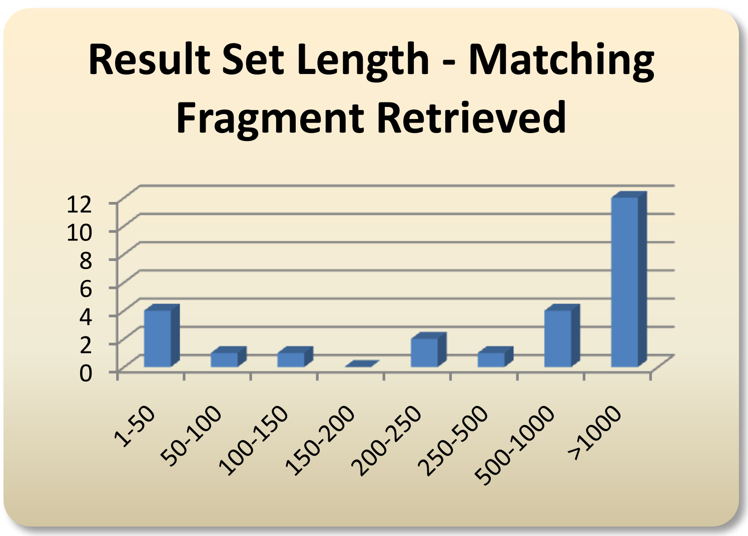


Figure 73 - Set Length – Matching Fragment Retrieved

4.2.2 Nature of queries

In an effort to identify possible ways of measuring the specificity of queries it was deemed useful to consider the nature of the query submitted by the subject. The "nature" of the query is defined by the types of its components, i.e. simple keywords or phrases. Some subjects used only words, others used a phrase or multiple phrases and others used a combination of phrases and words. A phrase in the query is a string of words surrounded by quotes, indicating that these words are being searched in that exact order within a document. Table 13 provides a count of queries falling in these three categories.

Number of Queries using Both Phrase(s) and Words	19
Number of Queries using Phrase(s) only	74
Number of Queries using Words only	50

Table 13 - Types of Queries

The availability of this information prompted for measuring the correlation between the number of words or number of phrases used in the queries and the reported size of the result set. More detailed data used in this section are included in 0.

Correlation	Pearson Correlation Coefficient
Between the Number of Words in Query and Total of Result Set	-0.25633
Between the Number of Phrases in Query and Total of Result Set	-0.089
Between the Number of Words in Query and Total of Result Set (applied only against subset of queries returning the second fragment)	-0.10626
Between the Character Length of the Largest Phrase in Query and Total of Result Set	-0.14686
Between the Sum of Character Lengths of Phrases in Query and Total of Result Set	-0.18045
Between the Number of Words in Query and the Fragment Character Length	-0.15778
Between the Maximum Phrase Character Length in Query and the Fragment Character Length	-0.25334
Between the Total Phrase Character Length in Query and the Fragment Character Length	-0.19535

Table 14 - Query Correlations

The calculations, as appearing in Table 14, indicate that the correlation of the number of words or phrases in the query and the total of the result set is weak. Even correlations between the character length of the largest phrase in the query and the total of the result set and between and the sum of the lengths of phrases in a query and the total of the result set are weak.

Another correlation examined, not necessarily related to the outcome of the query, was the one between the character length of the phrases and the character length of the Information Fragment from which the query had been constructed. It appears that this correlation is also weak, suggesting that the determination of the queries was not affected by the size of the passage the Group C subjects had available to them but rather the content of passage.

4.2.3 Distance between Associated Information Fragments

Query Number	Distance between matched items	Truncated Total	Actual Total
18	69	1000	68
19	719	1000	19
53	85	1000	14
54	502	526	52
55	210	1000	19
56	3	1000	19
60	26	1000	42
61	2	643	64
62	2	1000	52
72	94	683	68
73	587	1000	25
79	27	47	47
82	149	1000	53
84	14	93	93

Query Number	Distance between matched items	Truncated Total	Actual Total
90	1	130	13
93	45	212	21
99	145	228	22
109	2	24	24
112	8	44	44
116	2	36	36
122	927	1000	10
123	25	657	65
131	216	1000	56
133	71	1000	50
136	7	265	26

Table 15 - Distance between Matched Items

Table 15 lists the 25 queries which retrieved the matching Information Fragment. The "Query Number" column refers to the numbers of the queries as listed in 0. The "Truncated Total" column lists the values of the "Actual Total" truncated to 1000 where the value exceeds 1000. Among those 25, the mean of the distances between associated items was 157.52, which for all practical purposes is too long a distance to allow ordinary users easy retrieval of items they had considered as having some kind of relationship.

Figure 74, using the data in Table 39 presents the count of pairs of Information Fragments by the number of pages separating them within the result set. Two Information Fragments are arbitrarily defined as being in the same page if their distance within the result set is 10 or less, where 10 is the usual number of results in most Web search engine result pages. The number of pages between two Information Fragments is calculated as the rounded down result of the division of the distance by 10. We see that only 8 of these pairs (5.6% of the total number of pairs and 32% of the pairs for which the second Information Fragment was found) are found to be in the same page which is lower than 75% and strongly supports the fourth alternative hypothesis.

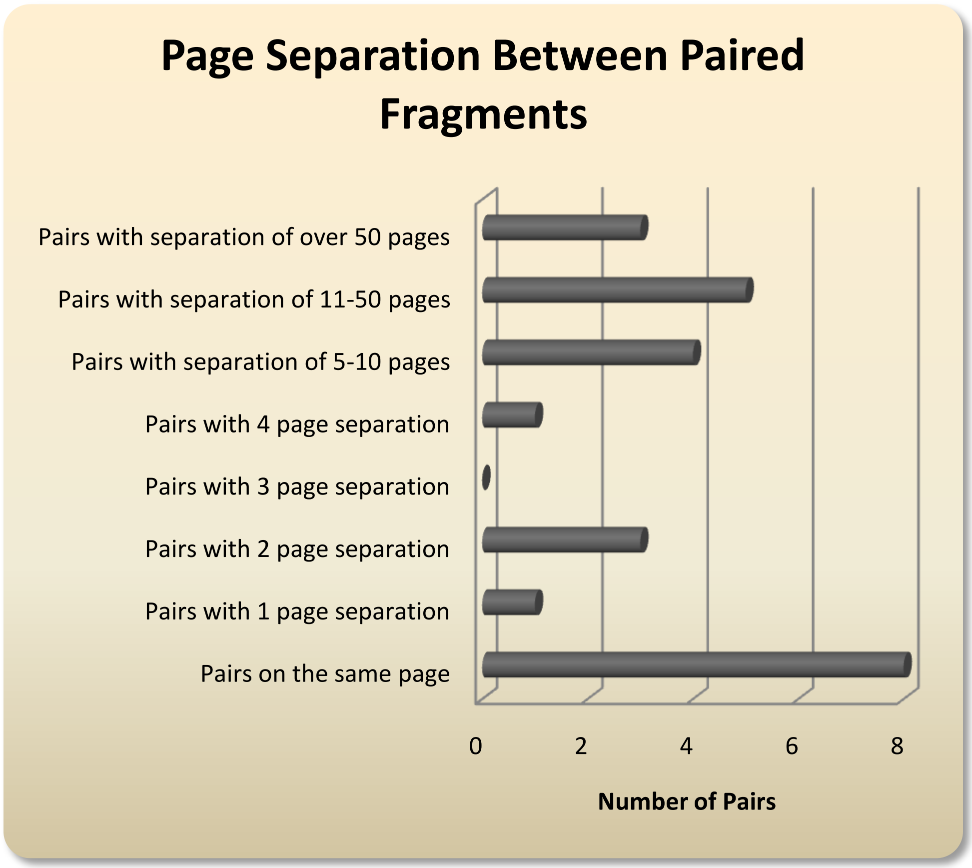


Figure 74 - Page Separation between Paired Fragments

Other than the effect of the query specificity noted earlier, no other pattern has been found in the data available here. No attempt to explain the differences between these distances is made here. The only conclusion derived here is that two Information Entities containing Information Fragments determined to be related by actual users will often either not appear together at all or be fairly far apart in result sets of traditional search engines.

4.2.4 Query Count before Query Submission

Another measurement taken during this task was the activity of the subjects as they were browsing in order to construct their search queries. Although the information collected was not relevant to the hypotheses at hand, it seems fitting to report here.

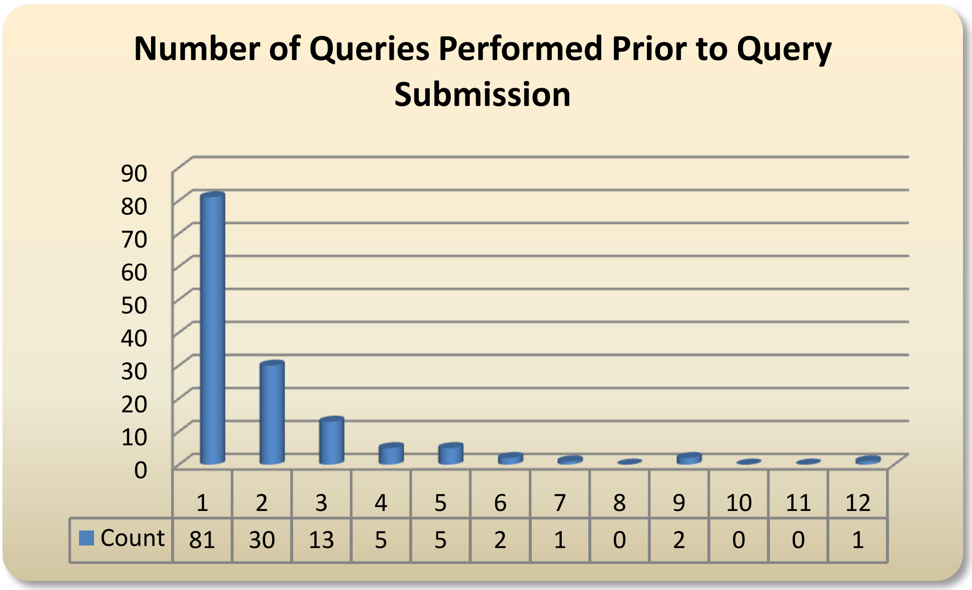


Figure 75 - Number of Queries Performed Prior to Query Submission

The results in Figure 75 indicate that in most of the cases subjects used only one query in preparation for their final query submission. This suggests that they tended to be as specific as possible and they were successful in retrieving their targeted resource close to the top of the result set.

4.2.5 Visits Count

Another measurement taken was the number of visits to specific web pages (i.e. views of the full-text) during the process of query construction for a given fragment. Throughout this process, the subjects' visits to specific web pages were recorded. A 'visit' is defined as the event of clicking on a link in the result set of a search query for the purpose of viewing the full content of the resource (instead of just the snippet offered in the result set). The results of this measurement are presented here:

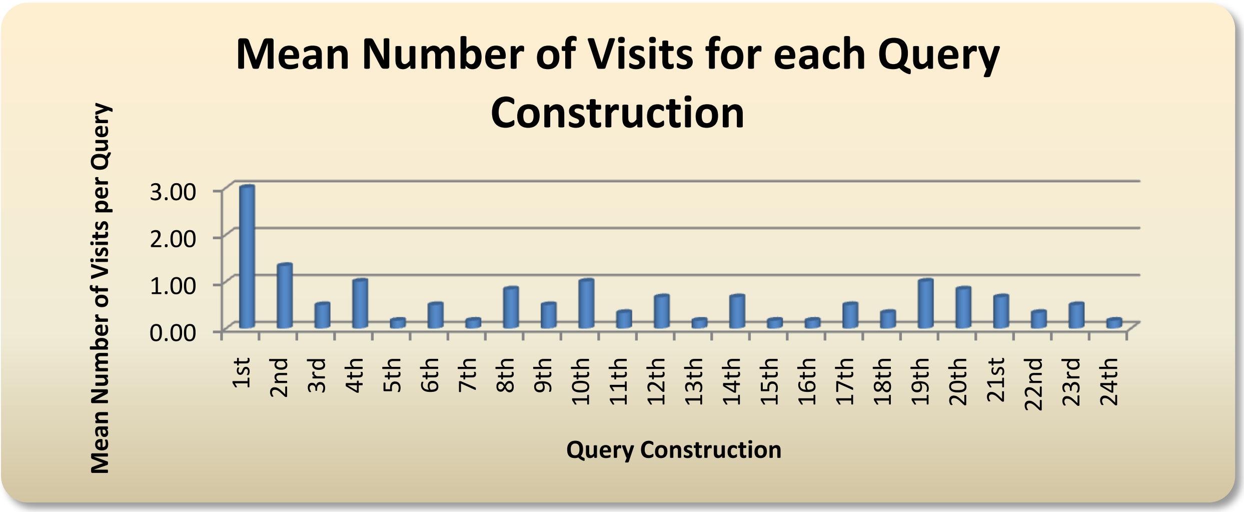


Figure 76 - Visits Count

Query	Visits Mean	Visits SD
1st Query	3.00	2.37
2nd Query	1.33	1.03
3rd Query	0.50	0.84
4th Query	1.00	1.55
5th Query	0.17	0.41
6th Query	0.50	1.22
7th Query	0.17	0.41
8th Query	0.83	0.75
9th Query	0.50	0.84
10th Query	1.00	2.00
11th Query	0.33	0.52
12th Query	0.67	0.82
13th Query	0.17	0.41
14th Query	0.67	1.21
15th Query	0.17	0.41
16th Query	0.17	0.41
17th Query	0.50	0.84
18th Query	0.33	0.52
19th Query	1.00	1.55
20th Query	0.83	1.60
21st Query	0.67	1.21
22nd Query	0.33	0.52

Query	Visits Mean	Visits SD
23rd Query	0.50	0.55
24th Query	0.17	0.41

Table 16 - Visits Count

The results indicate a relatively low number of visits to the full content of the resources. This means that in most cases subjects felt comfortable with the information they were receiving in the snippet presented to them as part of the result set. The disproportionably higher number of visits during the first query construction can be attributed to the subjects' desire to familiarize themselves with the information available to them. Once they started feeling comfortable using the information provided to them in the snippets, they resorted to visiting the resource only when necessary.

In order to better understand the reasons behind a subject's choice to visit the full content of a resource during the query construction process, the count of visits has been broken by Topic. Figure 77 demonstrates the number of full content visits for each one of the four Topic categories used in this study.

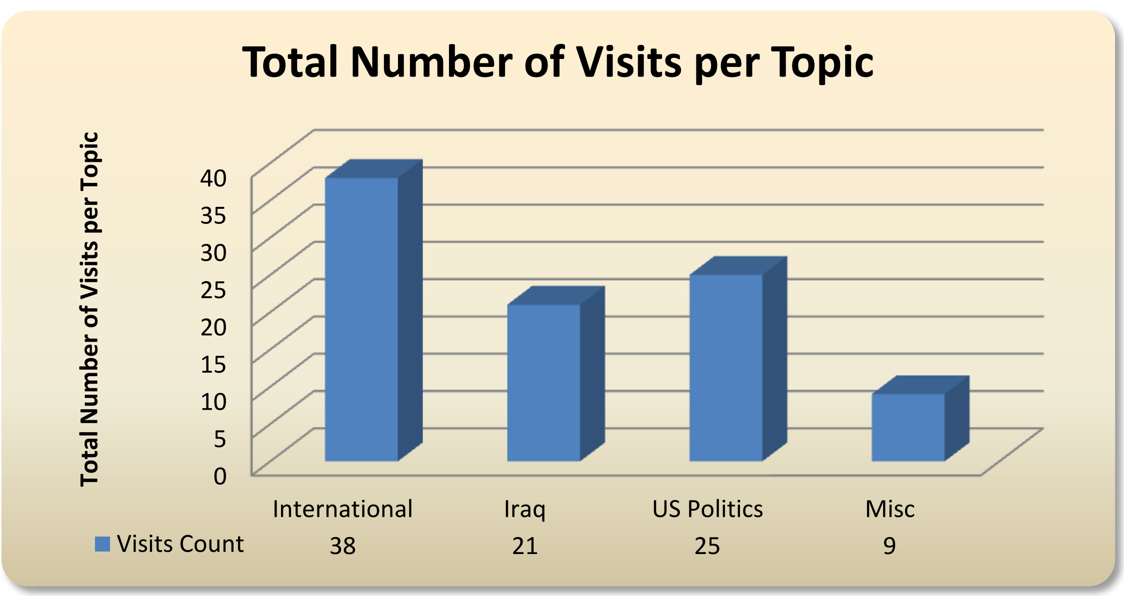


Figure 77 - Visits per Topic

The fact that the visits count for International stories is high is no surprise, because International stories were presented to the subjects first, and as noted above, their tendency to view the full content may have been the result of their desire to explore.

4.3 SUBJECT EXPERIENCE

In addition to performance measures, the test has been measuring the reactions of subjects at several points using brief questionnaires in order to test the truth or not of Hypothesis #2.

4.3.1 Task 1 Questionnaire Responses – Subject Experience

The questions were presented with 7-point Likert scales:

- How difficult was it to define a fragment in ____?
- How helpful was ____ in accomplishing the task of defining fragments?
- How enjoyable was using ____ for accomplishing this task?

The responses were comparable between the two Tools, slightly favoring FW. The means of the responses are:

	Baseline Mean	FW Mean	Baseline SD	FW SD	Mann-Whitney U	Z	2-tailed Asymptotic Significance (p)
Easy	4.45	4.73	1.21356	1.190874	51.500	-0.618	0.536
Helpful	4.09	4.18	1.30035	1.250454	57.500	-0.203	0.839
Enjoyable	3.73	4.18	1.272078	1.167748	47.000	-0.927	0.354

Table 17 - Task 1 Questionnaire Responses – Subject Experience

A non-parametric Mann-Whitney test was applied against these responses:

For the subject assessment of how *easy to use* the tools were, the test produced a 2-tailed

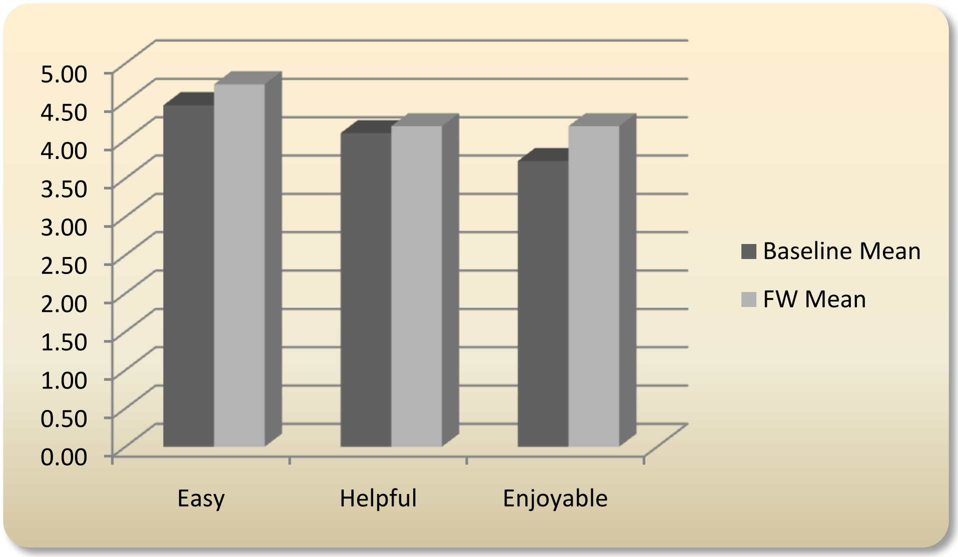


Figure 78 - Task 1 Questionnaire Responses – Subject Experience

asymptotic significance level of p=0.536 which means that there is no significant difference between the results of the responses for the two tools. Similarly, the test of the assessment of how *helpful* the tools were, produced a significance level of 0.839 and the test of assessment of how *enjoyable* the tools were, produced a significance level of 0.354. Because of this, the null hypothesis H_{1-0} cannot be rejected. Therefore the subjects' responses do not prove that subjects believe that either of the two tools is more easy, helpful or enjoyable than the other in the process of defining Information Fragments

4.3.2 Task 2 Questionnaire Responses –Subject Experience

Subjects were asked several questions throughout Task 2. They had to answer the same questions every time they finished one part of the task. The following questions were presented as 7-point Likert scales:

- Using ____, how easy was it to find the fragments you had previously defined?

- How effective was ____ in accomplishing the task of retrieving the fragments?
- How enjoyable was using ____ for accomplishing this task?

The following charts and tables present the results of tallying the responses of the subjects. Task 2, was designed to show whether bothering to create Fragment Associations has any benefits. The results demonstrate that there is a clear benefit.

4.3.2.1 Easy to Find

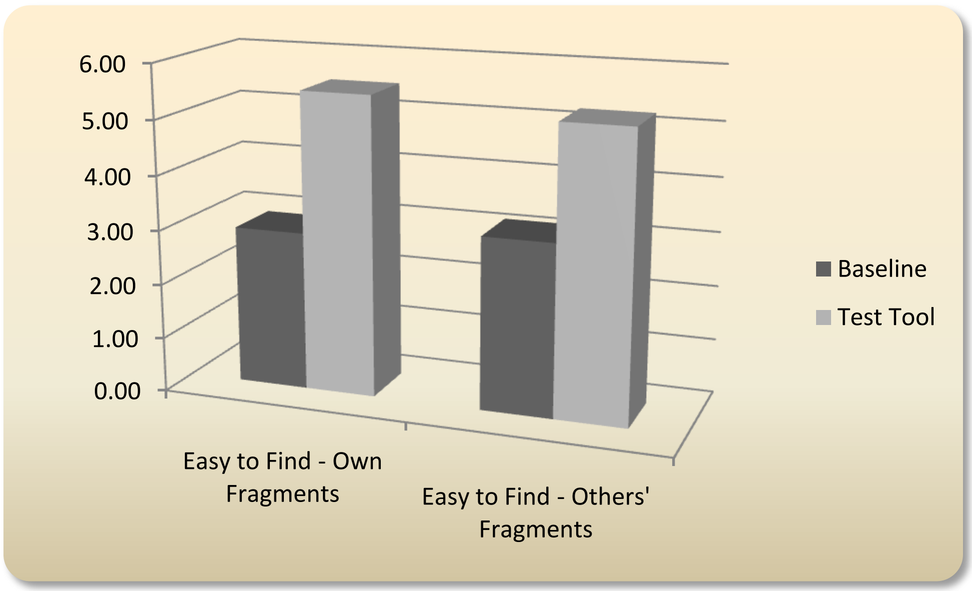


Figure 79 - Easy to Find

	Easy to Find - Own Fragments	Easy to Find - Others' Fragments
SPURL (Baseline) Mean	2.92	3.17
SPURL (Baseline) STD	1.88	1.53
FW (Test Tool) Mean	5.50	5.25
FW (Test Tool) STD	0.67	1.22
Mann-Whitney U	18	18
Z	-3.217141588	-3.194301178
2-tailed Asymptotic Significance (p)	0.001294747	0.001401698

Figure 80 - Easy to Find

A non-parametric Mann-Whitney U test was applied against these responses. The test produces a 2-tailed asymptotic significance level of 0.001294747 which means that the probability that the responses to the two tools are statistically significant at level $\alpha < 0.05$. Because of this, the null hypothesis $H_{2\text{-}0}$ has to be rejected. Therefore the subjects' responses prove that subjects believe that with FW it was significantly easier for them to retrieve the matching fragment among their own Information Fragments. When it comes to the subjects' responses about the task of finding the matching fragment created by others, the subjects' responses again show the significant difference between their views of FW and Baseline ($p = 0.001401698$ in Mann-Whitney test), so the null hypothesis $H_{2\text{-}0}$ has to be rejected again.

4.3.2.2 Effective

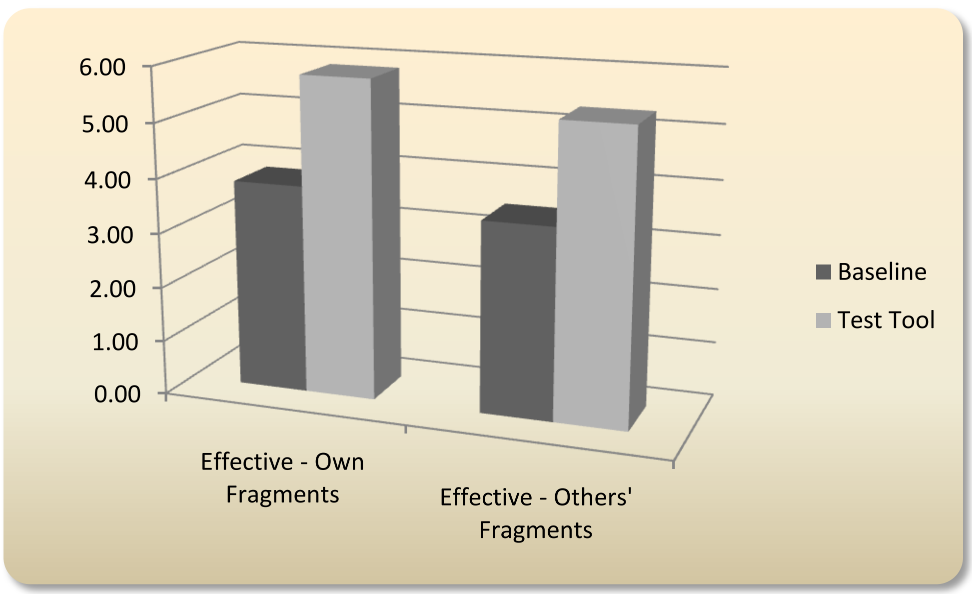


Figure 81 - Effective

	Effective - Own Fragments	Effective - Others' Fragments
SPURL (Baseline) Mean	3.83	3.50
SPURL (Baseline) STD	1.850470866	1.314257481
FW (Test Tool) Mean	5.83	5.33
FW (Test Tool) STD	0.389249472	0.887625365
Mann-Whitney U	23	18
Z	-3.093924621	-3.212682624
2-tailed Asymptotic Significance (p)	0.001975277	0.001315015

Figure 82 - Effective

A non-parametric Mann-Whitney U test was applied against these responses. The test produces a 2-tailed asymptotic significance level of 0.001975277 which means that that the responses to the two tools are statistically significant at level $\alpha < 0.05$. Because of this, the null hypothesis $H_{2\text{-}0}$ has to be rejected. Therefore the subjects' responses prove that subjects believe that FW was significantly more effective for retrieving the matching fragment among their own Information Fragments. When it comes to the subjects' responses about the task of finding the matching fragment created by others, the subjects' responses again show the significant difference between their views of FW and Baseline ($p = 0.001315015$ in Mann-Whitney test), so the null hypothesis $H_{2\text{-}0}$ has to be rejected again.

4.3.2.3 Enjoyable

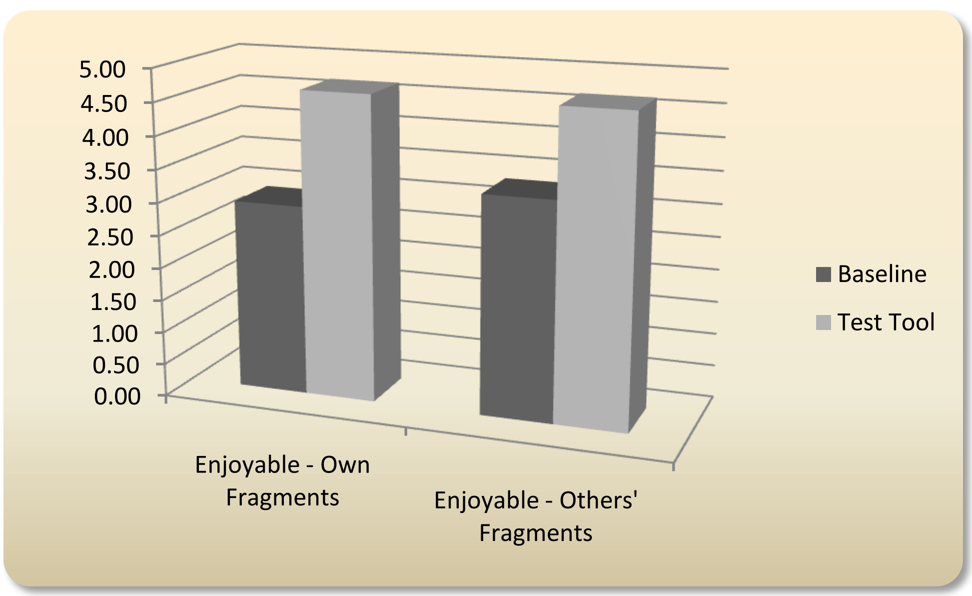


Table 18 – Enjoyable

	Enjoyable - Own Fragments	Enjoyable - Others' Fragments
SPURL (Baseline) Mean	2.92	3.33
SPURL (Baseline) STD	1.62	1.61
FW (Test Tool) Mean	4.67	4.67
FW (Test Tool) STD	1.15	0.89
Mann-Whitney U	26.5	34.5
Z	-2.686594199	-2.248221676
2-tailed Asymptotic Significance (p)	0.007218459	0.024562058

Table 19 - Enjoyable

A non-parametric Mann-Whitney U test was applied against these responses. The test produces a 2-tailed asymptotic significance level of 0.007218459 which means that that the responses to the two tools are statistically significant at level $\alpha < 0.05$. Because of this, the null hypothesis $H_{2\text{-}0}$ has to be rejected. Therefore the subjects' responses prove that subjects believe that using FW was significantly more enjoyable for retrieving the matching fragment among their own Information Fragments. When it comes to the subjects' responses about the task of finding the matching fragment created by others, the subjects' responses again show the significant difference between their views of FW and Baseline ($p = 0.024562058$ in Mann-Whitney test), so the null hypothesis $H_{2\text{-}0}$ has to be rejected again.

The responses to all three questions strongly support the alternative hypothesis that Information Fragment Association would improve information retrieval. It can be argued that the responses are not exclusively based on the Information Fragment Association capability. They may be partially dependent on interface design. This may indeed bear some truth, but the bottom line is that any interface design advantages an Information Fragment Association capable tool has are primarily due to its employment of Information Fragment Association and far less on other factors.

4.4 SUBJECT ASSESSMENT OF THE DEGREE OF USEFULNESS OF FRAGMENT ASSOCIATION

The degree of usefulness of Fragment Associations to the subjects was assessed with simple questions using 7-point Likert scales.

In Task 1 the following question was asked:

- How useful is the feature allowing the connection of two fragments?
    - The mean of the responses was 4.55

In Task 2 the following questions were asked:

- How useful is it to you to be able to see two related fragments side by side?
- How useful is it to you to have the ability to navigate from one fragment to the other?

The results were counted separately for the responses after retrieving one's own fragments and after retrieving fragments associated by others. The figures are roughly similar. It seems that the subjects responded very positively about the usefulness of the fragment association provided in FW.

Useful to see side-by-side - Own Fragments	5.50
Useful to see side-by-side - Others' Fragments	5.42
Useful to navigate - Own Fragments	5.17
Useful to navigate - Others' Fragments	5.17

Table 20 - Usefulness - Subject Responses to Questions

A question was posed regarding the capability of having a unique identifier for each Information Fragment for direct reference, which is an envisioned capability for a framework supporting Information Fragment Association. As shown in Table 21, the responses were very positive: 4.8 in a scale from 0 to 6.

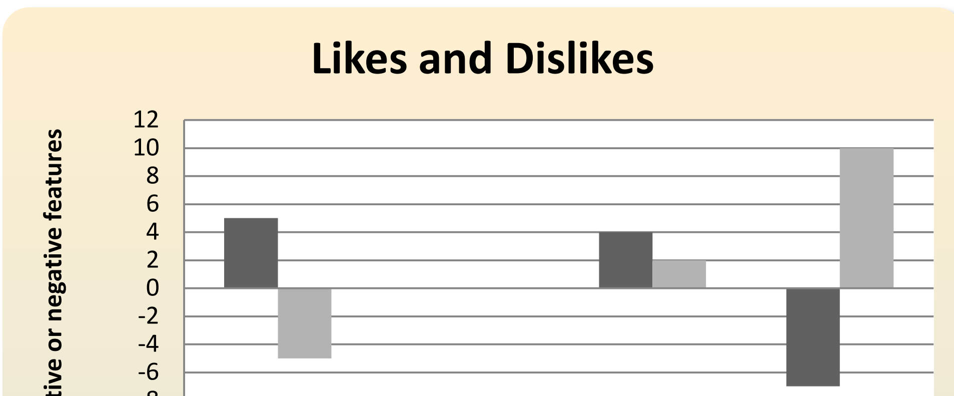
Subject	How useful is to you the capability of having a unique identifier for each fragment for direct reference?
1	6
2	No response
3	6
4	5
5	5
6	3
7	4
8	5
9	2
10	5
11	6
12	6
Average	**4.818181818**

Table 21 - Usefulness of Unique Identifier

The usefulness of Information Fragment Association was also determined through free-text responses the subjects provided regarding their likes and dislikes of the two tools. The only evaluation that matters as far as the Hypotheses are concerned is that of Retrieval Functionality, but it was interesting to gather some more information. Obviously the appearance of the hastily developed Test Tool (FW) was a negative factor, yet in spite of its unpolished façade the incorporation of Information Fragment Association provided a significantly higher rating of the Retrieval functionality. The responses are presented here along with their interpretation in four categories:

- Attractive
- Intuitive
- Display Functionality
- Retrieval Functionality

The results are demonstrated by Figure 83 and the data are listed in Table 41 and Table 42 in 0. Figure 83 portrays the four categories of characteristics on which the subjects provided comments in their free-text responses. This treatment is helpful in understanding how these distinct aspects of the two tools were viewed by the subjects. All four of these categories affected the overall assessment of usability, but the objective of this Study concentrates only on the fourth category. The Study is trying to assess the how the presence of Information Fragment Association affects retrieval functionality. By presenting the responses in this fashion, we get a clear picture of how valuable Information Fragment Association (present only in the test tool FW) was for the subjects. We also derive information about these other categories of characteristics. Since these characteristics may have affected some of the other responses given by the subjects, this presentation is helpful for viewing those responses in a new light.



	Attractive	Intuitive	Display Functionality	RetrievalFunc tionality
Baseline	5	0	4	-7
Test Tool	-5	0	2	10

Figure 83 - Likes and Dislikes

The values in Figure 83 are calculated based on a scheme using 1 for a positive comment and -1 for a negative comment. The determination of what constituted a positive or negative comment and what the comment was about was based on several distinct words used by the subjects in their comments. For example, when a subject mentions liking or disliking the colors a value of 1 is added or subtracted from the “Attractive” count. When the retrieval functionality is called “limited” a value of 1 is subtracted and when called “dynamic” a value of 1 is subtracted. When a subjects mentions liking or disliking one aspect of the data display a value of 1 is added or subtracted, and so forth.

Usefulness was also assessed by taking into consideration brief casual conversations with the subjects upon completion of both tasks. Samples from recordings of these conversations can be found in 0. Although extracting information out of casual comments is an imprecise process, an attempt is made here to present this information focusing on three areas:

- the subjects’ assessment as to whether Information Fragment Association is useful
- whether the subjects feel that they can use Information Fragment Association now or in the future
- the positive and negative comments made by the subjects about the two tools they used in this test

All subjects provided their opinion regarding the usefulness of Information Fragment Association (Figure 84). 8 subjects found Information Fragment Association useful, one did not, and two found it useful but their comments were lukewarm.

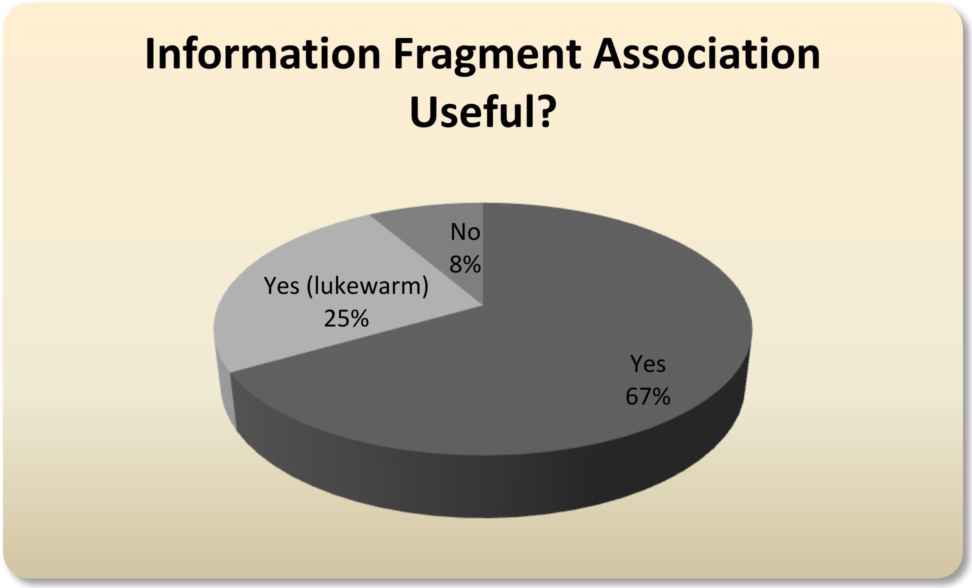


Figure 84 - Information Fragment Association Useful?

Six of the subjects expressed some opinion regarding the possibility that they might use this functionality in their own work now or in the future (Figure 85). Of the six, three of them said they could use it, one said it would not be useful, and two others expressed doubt that they would use it but they suggested that others might use it.

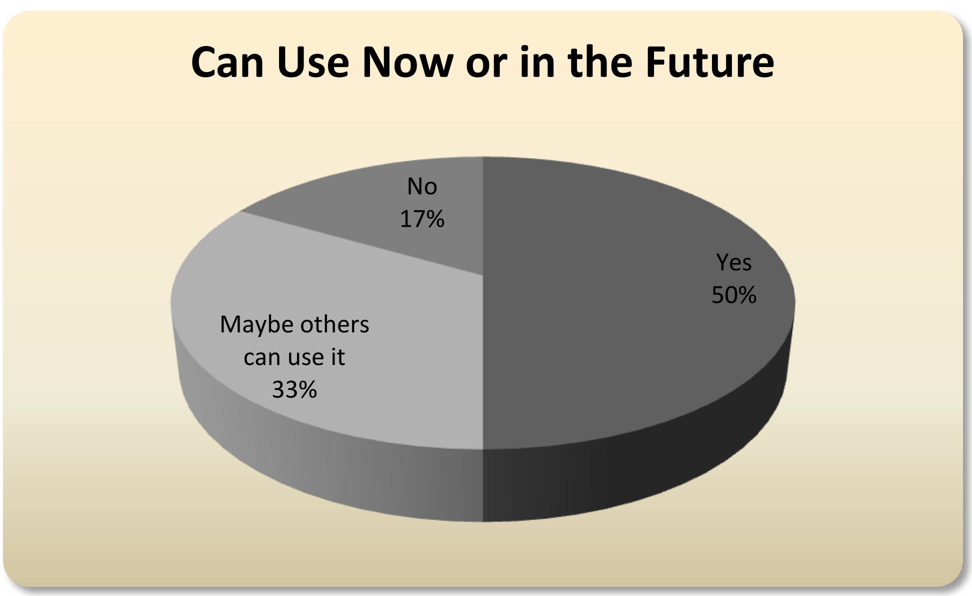


Figure 85 - Can Use Now or in the Future

Some of the subjects made positive or negative comments regarding the two systems (Figure 86). Some of the comments were made in the form of a comparison between the two tools; some of them were focusing on one tool. More positive than negative comments were made about both tools, but FW (the Test Tool) had a clear advantage.

The general feeling gathered from these comments is that the majority of the subjects feel that Information Fragment Association functionality would be useful in general. It is unclear whether some of them would choose to use Information Fragment Association if it was available today. Some of them expressed the opinion that it would be useful to them in the future. In general, most subjects viewed Information Fragment Association as a capability more useful for research than an alternative to casual bookmarking.

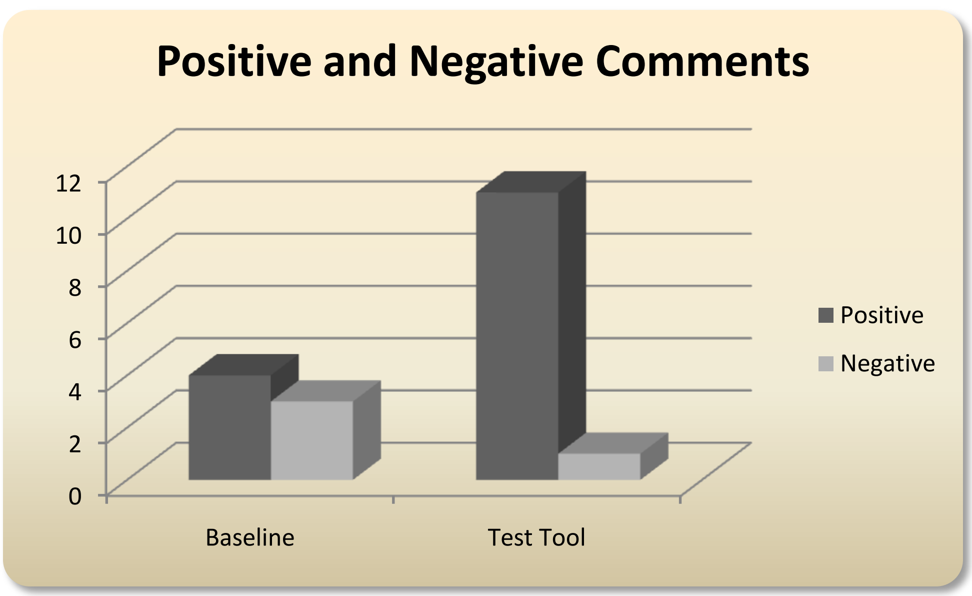


Figure 86 - Positive and Negative Comments

4.5 PRECISION, RECALL AND F-MEASURE OF SEARCHES – TASK 4

The objective of Task 4 was to determine the degree to which the presence of Information Fragment Association improved the Precision, Recall and F-Measure of queries within a set of information resources. This task utilized the queries submitted by the subjects of Group C as well as the full set of Information Fragments and Information Fragment Associations defined by Group A and Group B.

In Task 3, the queries were constructed against Live Search. In Task 4 these queries were submitted to the two tools under examination. Two minor modifications were made to the queries, namely the removal of the "+" operator and commas after words in order to make the queries compatible with the tools. Upon submission, the results were recorded as listed in Table 43 and Table 44 in 0. In 19 cases out of the 143 Baseline failed to retrieve anything. Upon examination, it was discovered that in 16 of these cases the failure was due to its inability to support single quotes within the query. The desired fragment was actually retrieved with a modified query in order to verify its existence, and the results of this modified query were used for the test. The other 3 cases were due to similar small technical inadequacies, and after several attempts using modified searches the one of the three queries was rendered functional and counted. FW (the Test Tool) retrieved all 143 Information Fragments.

Because of the specificity of the queries, in most of the cases only one Information Fragment was retrieved by Baseline and only two Information Fragments (one pair) was retrieved by FW. This gave FW a big advantage on Recall, but both tools had less than optimal Recall performance.

The outcome strongly supports the alternative Hypotheses H_{5-1}, H_{6-1} and H_{7-1}.

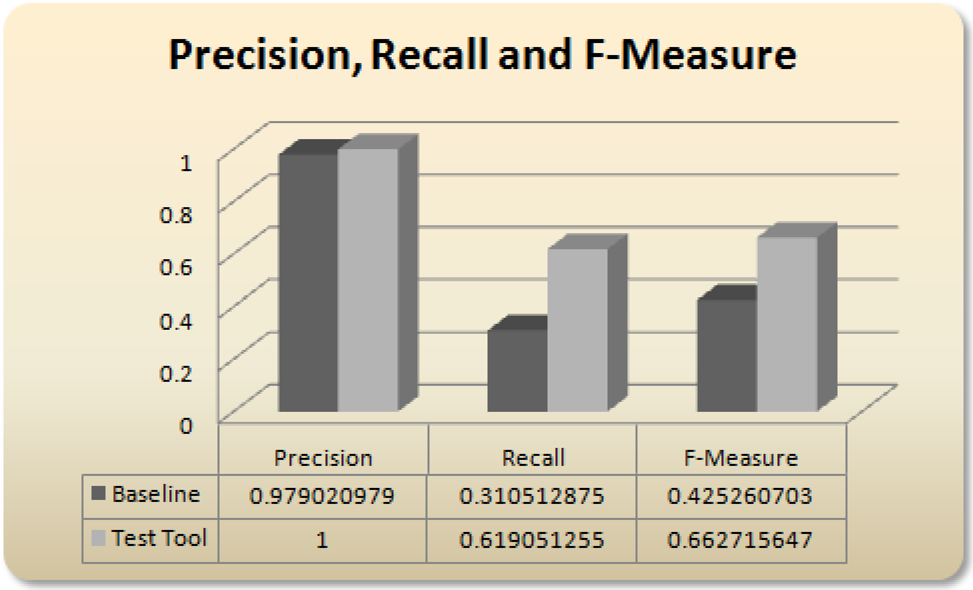


Figure 87 - Precision, Recall and F-Measure

A paired t–test was performed against these measurements shown in Table 22.

When applied against the Precision measurements, the test yielded a p of 0.0821749116 which means that the Precision is not statistically significant at level $\alpha < 0.05$. Therefore, Precision achieved by FW is not significantly better than that of Baseline and the null hypothesis $H_{6\text{-}0}$ cannot be rejected.

However, when applied to the Recall results the test yielded a p of 0.000000000001 and when applied to the F-measure results the test yielded 0.000000000008, which means that both the null hypothesis $H_{5\text{-}0}$ and the null hypothesis $H_{7\text{-}0}$ have to be rejected.

	P(T<=t) two-tail
Precision	0.082174911583
Recall	0.000000000001
F-measure	0.000000000008

Table 22 - Parametric Testing for Precision, Recall and F-Measure

4.6 STRENGTHS AND WEAKNESSES OF THE TWO TOOLS

4.6.1 Baseline Strengths

SPURL, the tool selected as Baseline, is a good social bookmarking service with useful features. Although it is not the most popular or trendy among its peers, it is a very representative one. This study has highlighted some of its strengths and has increased my awareness of particular design elements which can be considered to be of value for future system design. The two outstanding strengths that Baseline exhibited in this study were its familiar interface and its compact results view.

4.6.1.1 Familiar Interface

Baseline offered the subjects with a very familiar interface. All of its components were similar to just about any other web application they had previously used. This was a tremendous advantage, as subjects felt more confident in their attempts to discover and take advantage of the functionality of the interface.

4.6.1.2 Compact Results View

Baseline displayed the result set in a compact view. This means that each item in the result set was presented in a single line consisting of the title of the Information Fragment, but not the Information Fragment itself. The Information Fragment could easily be displayed with a click. This allowed for better examination of the search results and alleviated Baseline's significant retrieval shortcomings.

4.6.2 FW Strengths

FW (the Test Tool) was developed specifically for this study in order to provide a platform supporting Information Fragment Association. It lacked the polished design of the commercial tool to which it was being compared, but merely by merit of its incorporation of Information Fragment Association functionality it has fared significantly better both in the performance and usability tests.

4.6.2.1 Well-designed Information Fragment Definition Capability

A feature which made FW attractive was how the Information Fragments were defined in it. Unlike Baseline which accomplished the task with pop-up windows disappearing after the fragment is submitted, FW presented the user with a visible result of their Information Fragment definition (see Figure 42). Moreover, it presented the two associated Information Fragments clearly on a different screen for submission (Figure 43). This contributed to subject satisfaction in spite of the fact that the process involved an additional action.

4.6.2.2 Solid Information Fragment Definition Performance

Because of the similarity of the process, the two tools were expected to fare equally in Information Fragment Definition performance tests. However, a paired t-test performed on the Information Fragment definition completion time results for the two tools revealed that FW performed significantly better. (see Section 4.1.4).

4.6.2.3 Strong Matching Fragment Retrieval Performance

This is where the significant advantage of Information Fragment Association manifested itself. The fact that relationships between specific Information Fragments had been recorded by the subjects as part of the submission process made possible their immediate association and retrieval. This built-in capability obviated the need for undergoing the tedious retrieval process the subjects had to experience with Baseline in order to find related information. As a result, both the performance (see Section 4.1.6) and the user satisfaction (see Section 4.3.2) were significantly higher.

4.6.3 Baseline Weaknesses

Baseline fared much worse than expected in this Study. Its two main areas of weakness were in the poor Information Fragment definition technique and mostly in the lack of Information Fragment Association capability.

4.6.3.1 Less than Optimal Information Fragment Definition Process

The use of popup windows (Figure 41) for Information Fragment Definition was a negative factor. Subjects sometimes ended up attempting to resubmit their fragment in order to make sure it got submitted correctly. One possible way with which this tool can be improved is by incorporating the submission functionality in internal panels as opposed to using separate popup windows.

4.6.3.2 Poor Matching Fragment Retrieval Performance

In spite of the fact that Baseline performs every traditional social bookmarking operation very well, it lacks the ability allowing the user to define an explicit and specific relationship between Information Fragments. The study revealed that the impact of the lack of this capability is far greater than even imagined prior to the beginning of the study. The subjects had considerable difficulties finding an Information Fragment containing similar information to the one at hand even though their choice was among only 23 other Information Fragments. In a real life situation, with real life numbers of bookmarks, finding the matching fragment would have been like looking for a needle in a haystack, even if the bookmarks were categorized using the capabilities provided by most of today's social bookmarking tools.

4.6.4 FW Weaknesses

FW exhibited some shortcomings in spite of its significantly better performance. These shortcomings were the relative unfamiliarity of the interface, the use of non-hyperlinked URIs in some displays and the not very obvious result set browsing capability. Moreover, the use of an Information Fragment Association Interface necessitated the use of one additional screen (see Figure 43). Although the results of Task 1 indicate that this was not a problem for the subjects, it is still an additional step. Since, as we see above in Section 4.6.2.3, FW fared very well in performance and user satisfaction tests in spite of these shortcomings, we can appreciate the contribution of Information Fragment Association even more.

4.6.4.1 Unfamiliar Interface

The design of FW consisted of browser sessions embedded into a tabbed interface. Although it bore a fairly good similarity to modern tabbed browsers, the different feel and look caused the subjects to pause a little in the beginning. The use of buttons outside the web page display area to control the display of information and events within the web page display area was a novel approach and it was not very intuitive to the subjects.

The fact that FW was combining web applications with thick client applications, offered an interface less familiar albeit more powerful.

4.6.4.2 Use of Non-hyperlinked URIs

The Information Fragment detail page contained a reference to the URI of the Information Entity, but it did not provide a hyperlink. This was not a serious issue but it appeared odd to two adventurous subjects who wanted to display the Information Entity. The subjects eventually found the right button performing that action, but it was not immediately obvious to them.

4.6.4.3 Not Very Obvious Result Set Browsing Capability

FW provided the capability of browsing through specific results or through the entire set of defined bookmarks, just like Baseline did. However, this capability was not very obvious and was not used by any subjects. One of the reasons why it was not used was that it was not really needed, but it probably would have been used by intrigued subjects had it been more visible.

5 SUMMARY AND DISCUSSION

The goal of this study was to determine whether Information Fragment Association offers sufficient benefits to warrant further discussion regarding the feasibility of the deployment of an infrastructure such as the envisioned framework to be presented in Chapter 0 . As a means of making this determination two tools were selected for comparison: a traditional social bookmarking tool without Information Fragment Association functionality (e.g. SPURL) and a test tool, named FW, offering this functionality.

Since our objective has been to study the effect of Information Fragment Association, a lot of attention has been paid on deemphasizing any other factors. Using two tools in a study, even two carefully selected tools, inevitably invites comparison between features of these two tools other than the feature the study focuses on. This certainly did occur in the case of this study, however, the presence or not of Information Fragment Association functionality overshadowed all of the other differences.

5.1 ACCOMPLISHMENTS AND OMISSIONS OF THIS STUDY

This Study set out to meet the following goals, first outlined in Section 3.1:

- Measure the Efficiency of Bookmarking and Retrieving Information Fragments

- Estimate the Accuracy of Association
- Evaluate User Experience
- Determine the Need for Enhancement of Search Engine Results
- Calculate Recall and Precision of Searches in Social Bookmarking Tools

After the presentation of the study design, the task execution and the exhibit of the results, we are ready to revisit these goals in order to ascertain whether they have been achieved.

5.1.1 Measurement of the Efficiency of Bookmarking and Retrieving Information Fragments

Hypothesis #1 was designed in order to provide an answer to the question whether the total time needed for the entire process of bookmarking and then of retrieving two Information Fragments within two Information Entities would be shorter if an Information Fragment Association interface rather than a mainstream social bookmarking tool like SPURL is used. The completion times for the two tasks assigned to the subjects were used as indicators of performance.

The Hypothesis was successfully tested by analyzing the results of two tasks performed by subjects. These tasks produced measurements of the efficiency of bookmarking and retrieving Information Fragment with and without the Information Fragment Association enhancements. The null hypothesis was rejected both in the analysis of the Information Fragment definition results (see Section 4.1.4) and in the analysis of the Information Fragment retrieval results (see Section 4.1.6), suggesting that the tool supporting Information Fragment Association (i.e. FW) performed significantly better. An encouraging outcome was that there was no performance penalty in the process of association of the Information Fragments although a small penalty would have been reasonable.

The results were analyzed in a variety of ways, not because there was need for further testing the hypothesis, but rather in order to derive additional useful conclusions. The results of Information Fragment definition (Task 1) were analyzed by topic. The stories used in the Study belonged to four topic categories. After performing paired t-tests on the results of all four topics the null hypothesis was rejected for all of them, with the interesting observation that in the case of Miscellaneous News the t-test yielded a figure which was lower than the significance level of 0.05 but not as low as the figures for the other topics.

The results of Information Fragment retrieval (Task 2) were analyzed by the original creator of the Information Fragment pair (i.e. whether the subject was performing the retrieval on his/her own Information Fragments or somebody else's). They were also analyzed by male and female subjects. In all cases the null hypothesis was rejected. An interesting observation was that female subjects performed a little better with Baseline, but with FW both genders performed similarly.

Related to performance was the success rate of identification of the matching fragment (see Section 4.1.1). Counts of the observations in which a correct matching fragment was found, a wrong fragment was found or the subject gave up. Subjects were able to always find the matching fragment when Information Fragment Association functionality was in place, but failed a few times when using Baseline. After performing a paired t-test on the results it was revealed that the degree of failure with Baseline is significant enough to support the alternative hypothesis. Information Fragment Association provides a clear advantage.

Another measurement involved the number of keyword searches or number of Information Fragments viewed during browsing in Baseline prior to identifying the matching fragment (see Section 4.1.2). This measurement demonstrated the effort subjects had to make in order to accomplish this task in Baseline. More effort is expected to be necessary when users are dealing with large numbers of bookmarks.

In general the conclusions drawn from these measurements demonstrated that there are substantial benefits to employing Information Fragment Association. These benefits encourage us to believe that establishing an infrastructure for recording Information Fragment Associations would be a very useful enhancement to any tool used for information management and retrieval. This enhancement can conceivably be implemented in these tools by consuming a web service furnishing the Information Fragment Association data.

The measurements taken involved simple Information Fragment definitions and associations and did not involve the selection and assignment of any semantic descriptions for these associations. In other words, the subjects were not asked to establish more complex relationships, such as “A refutes B”, “A supports B” or “A updates B”. Had this been part of process, the results may have been slightly different. A further study involving an observation of the use of assigning semantic descriptions for the content of the Information Fragments as well as their relationship would be interesting, but more complex to carry out.

5.1.2 Estimation of the Accuracy of Association

This goal attempts to find out the degree to which associating Information Fragments can provide more content accuracy. If we are searching for a specific piece of information which happens to be in an Information Fragment, it is more accurate if we search only the Information Fragment and not the entire Information Entity. The differences in Accuracy of Association between Information Entities and Information Fragments were adequately measured with the web searching results. This goal was set without necessarily having Web searching in mind, but it was achieved using the Web searching task (Task 3) of this Study. Web search engines search the entire web page and not a specific Information Fragment in them. It was demonstrated that there is a significant distance within the search results between two Information Entities encompassing two interrelated Information Fragments (see Section 4.2.3).

However, the results would have illustrated the issue better had a further test been performed. It would have been desirable to use the collection of Information Fragment Associations generated by this study to compare the results of queries performed on the Information Fragments with queries performed on the entire Information Entities. This would have produced results more geared to determining the accuracy of association than queries against the entire universe of Information Entities on the Web.

Because this goal was handled by the same task as the Web search engine enhancement goal with which it in many ways coincides, more discussion and results analysis is provided in Section 5.1.4 below.

5.1.3 Evaluation of User Experience

User experience was evaluated with questionnaires and further discussion with the subjects. This goal was achieved, but the results were perhaps less lucid than in the cases of the other measurements. A variety of factors other than Information Fragment Association are likely to have affected these subjective assessments.

Hypothesis #2 was designed to help answer the question whether the overall usefulness and usability of the process of bookmarking and then of retrieving two Information Fragments within two Information Entities would be better when using an Information Fragment Association interface rather than using a mainstream social bookmarking tool.

During the Information Fragment definition process (Task 1) subjects were asked to provide rankings evaluating how easy, helpful and enjoyable each tool was for accomplishing the assigned task (see Section 4.3.1). A non-parametric Mann-Whitney U test was applied against the responses, and the results revealed that there is no statistically significant difference between the two tools in any one of these rankings, even though the tool supporting Information Fragment Association fared a little better. However, the analysis of the rankings provided during the Information Fragment retrieval process (See Section 4.3.2) reveals a statistically significant difference between the ratings for the two tools, demonstrating that FW, the tool with Information Fragment Association capabilities, provided higher user satisfaction with regards to the ease of finding the matching fragment, its effectiveness, and the degree to which subjects enjoyed using it.

Hypothesis #3 was designed to answer the question of how useful users would find the ability to establish Information Fragment Associations, to see Information Fragments side-by-side and to navigate from one Information Fragment to the other. This hypothesis was tested with responses to questions focusing on the Information Fragment Association. These questions asked for ratings of the ability to view Information Fragments side-by-side or to easily navigate between Information Fragments during the Information Fragment definition process (see Section 4.4). The means of the ratings were over 5 on a scale from 0 (the least favorable rating) to 6 (the most favorable rating). A rating of 4.8 was obtained when subjects were asked to evaluate how useful it was to them to have unique identifiers for each fragment for direct reference. Based on these responses the alternative hypothesis was accepted.

The usefulness of Information Fragment Association was also determined through free-text responses the subjects provided regarding their likes and dislikes of the two tools. Out of the subject responses, positive and negative comments about four categories of characteristics were gathered and calculated using a numeric scheme (see Figure 83). The only category of relevance to the Study was the retrieval functionality, but it was interesting to also see the comments on the other categories because they affected to some extent the overall evaluations the subjects provided. The comments on retrieval functionality showed that the tool supporting Information Fragment Association had a very big advantage.

Subjects found Information Fragment Association to be useful (see Figure 84), but when asked whether they can actually presently use it in their work (see Figure 85) the responses were far less enthusiastic. This may imply that this functionality may be of more value to specific audiences, such as the scholarly community. Some subjects felt that they could use the functionality at some point in the future. Some of them gave thought to scenario in which Information Fragment Association can be useful such as "getting different angles" of an issue. The number of positive and negative comments about each tool was calculated and presented in Figure 86.

A more accurate evaluation would have been carried out by presenting subjects with two identical interfaces, with the Information Fragment Association being the only differentiating factor.

This evaluation has provided useful insights for future directions. It has shown that users appreciate the improved retrieval functionality offered by Information Fragment Association enhancements. It has also shown that users feel more comfortable with familiar interfaces, and that a more successful application of Information Fragment Association would involve simulating these familiar interfaces while at the same time providing the new functionality. It has also shown perhaps a first step toward a successful implementation of this infrastructure may be to enhance tools being used every day, such as Web search engines. Another possibility may involve the customization of a tool with the purpose of serving a very narrow function for the benefit of a smaller community. If this customization serves needs not met by any other tool, it would generate a more enthusiastic group of users who would contribute to its future development.

5.1.4 Determination of the Need for Enhancement of Web Search Engine Results

The Study has demonstrated the inadequacies of mainstream search engines when it comes to bringing together Information Entities containing interrelated Information Fragments.

Hypothesis #4 addressed the difference with which Information Entities and Information Fragments are handled by Web search engines. The fourth null Hypothesis H_{4-0} stated that "more than or equal to 75% of a given set of pairs of related Information Fragments will have a rank difference less than 10 in the Web search results when a query designed to retrieve one of the Information Fragments is applied to a Web search engine".

Even though the Information Entity pairs involved in this Study were carefully selected to offer an abundance of points of thematic similarity, the results of Task 3 (see Section 4.2.3) revealed that queries devised with the purpose of retrieving an Information Fragment and its encompassing Information Entity failed in most of the cases to return a result set containing the second Information Entity within the same results page. Out of the 143 pairs of Information Fragments used, only 8 pairs or 5.6% were found to be in the same page which is lower than 75% and strongly supports the fourth alternative hypothesis. This means that even if the matching fragment is retrieved by the Web search engine, the user may still not be able to find it. Enhancing the results with Information Fragment Association would not create a redundancy of links, but every future design should take into consideration the possibility that a results page may provide two ways of getting to an Information Fragment, one supplied by the Web search engine and the other by the Information Fragment Association registry.

The Study went beyond this analysis to further analyze the results and to examine other useful information gathered during this task. One of these analyses involved the "specificity" of queries (see Section 4.2.1). Queries returning short result sets (less than 100) were considered "specific". Analysis of the results reveals that out of the 118 queries which failed to return the matching fragment 47 were not "specific". This means that the specificity of the query cannot always be used to explain the failure to retrieve the matching fragment.

The number of queries tested in the Web search engine by the subjects prior to submitting a query was counted (see Section 4.2.4). The results indicate that in most of the cases subjects used only one query in preparation for their final query submission. This suggests that they tended to be as specific as possible.

The number of visits to specific web pages (i.e. views of the full-text) prior to submitting a query was also counted (see Section 4.2.5). The mean of all subjects' visits for each query construction is 3 or less. This means that in most cases subjects felt comfortable with the information they were receiving in the snippet presented to them as part of the result set. An analysis of visits by topic revealed that subjects visited 38 pages while searching for a fragment on International news, 21 pages while searching for a fragment on Iraq news, 25 pages while searching for fragments on US Politics and 9 pages while searching for fragments on miscellaneous news.

The conclusion we can derive from this portion of the study is that although mainstream search engines today present us with valuable result sets, we are still very likely missing resources closely related individually to each one of the items in the result set we get. Since, as Task 1 and Task 2 demonstrated, humans can painlessly provide these associations through their everyday work, the merging of search engine results and association results would generate a more rich and valuable information retrieval environment.

The results have painted a much clearer picture of what needs to be implemented so that the benefits of Information Fragment Association may become evident to the wider public. By establishing an Information Fragment Association infrastructure and an interface allowing these associations to be inserted in a Web search engine results we would help users understand that establishing Information Fragment Associations has far reaching impact, beyond the realm of mere bookmarking.

The Study could have gone further than merely demonstrating what web search engines do poorly. It could have presented a solution integrating Information Fragment Associations with web search results and performed some tests against it. However, such tests would be more meaningful with the accumulation of larger numbers of Information Fragment Associations, and they would be more appropriate for future testing, if the establishment of an environment supporting Information Fragment Association ever comes to fruition.

5.1.5 Calculation of Recall and Precision of Searches in Social Bookmarking Tools

The Study was able to successfully demonstrate the difference of performing a search on an association-enhanced Information Fragment dataset as opposed to a simple Information Fragment dataset.

Hypotheses #5-7 were designed to provide an answer to the question whether better retrieval effectiveness (measured by Recall, Precision and F-measure) would be obtained when keyword searches are performed on a dataset containing user-defined Information Fragment pairs in a tool supporting Information Fragment Association rather than performed on the same dataset in a tool not supporting Information Fragment Association.

After performing these keyword searches, the Recall, Precision and F-measure were calculated. The Recall was 0.310512875 for Baseline and 0.69051255 for FW, the Precision was 0.979020979 for Baseline and 1 for FW and the F-measure was 0.425260703 for Baseline and 0.662715647 for FW. A paired t-test was performed on these results.

The t-test indicated that the Recall and F-measure achieved by FW are statistically significantly higher than those achieved by Baseline and that the null hypothesis H_{6-0} and the null hypothesis H_{7-0} have to be rejected.

The Precision for FW was higher than for Baseline but not statistically significantly higher, so the null hypothesis H_{5-0} could not be rejected. The measurement of Precision for Baseline was high because 17 queries which had failed were eventually counted after it was determined that they were caused by technical flaws in Baseline. This ensured that the lack of Information Fragment Association capabilities was the only factor considered. Even more reliable results would have been yielded using a single tool configured to be able to perform searches with or without Information Fragment Association.

SPURL (Baseline) stores the selected Information Fragment in the *Snip* field. In addition to the *Snip* field it has a *Description* field the purpose of which is presumably to allow the user to provide a description of the content of the selected Information Fragment or perhaps the entire Information Entity. A resourceful user can conceivably compensate for the Baseline's lack of Information Fragment Association capabilities by copying the content of an Information Fragment A to the *Description* field of Information Fragment B and vice versa. This would allow for the two Information Fragments to be retrieved together by a search. It would also demonstrate the desperate maneuvers a user would have to resort to make up for the lack of Information Fragment Association capabilities. In addition to the fact that this would be an error-prone approach as it would require cutting and pasting and that this would require a lot of effort on behalf of the user, such a workaround would produce less than optimal results. The user would be forced to use the content of an Information Fragment as the description of the other Information Fragment. One can reasonably argue that this would not be an advisable information management practice. Moreover, such a technique would not scale to multiple Information Fragment Associations on a single Information Fragment. The more one examines any such attempts at providing workarounds for the lack of Information Fragment Association the more evident it becomes that none of these workarounds is adequate for associating Information Fragments, and it makes the argument for the establishment of an Information Fragment Association even stronger.

5.2 ADVANTAGES AND DISADVANTAGES OF INFORMATION FRAGMENT ASSOCIATION

5.2.1 Advantages

Tests performed as part of this Study have demonstrated that Information Fragment Association can significantly improve the performance of retrieval of interrelated Information Fragment bookmarks. By retrieving one Information Fragment, a user can retrieve all interrelated Information Fragments right away without having to resort to time consuming keyword searching or browsing. The benefits of this improved functionality are reflected in the satisfaction of the users, as the responses of the subjects who participated in this Study indicate.

Information Fragment Association can be an invaluable enhancement to any tools widely used for information retrieval, such as social bookmarking tools and Web search engines. In the case of Web search engines, Information Fragment Association can resolve the problem caused by the fact that Web search engines perform searches on the entire Information Entities and not on the Information Fragments, thus very often keeping two Information Entities with interrelated Information Fragments very far apart within the result set

In social bookmarking tools, Information Fragment Association can improve the recall and precision of searches performed on Information Fragments. Information Fragments have a limited number of words and they can be easily missed by a keyword search not using those words. Information Fragment Association ensures that an Information Fragment is retrieved by a search which would have missed it had it not been associated with another Information Fragment.

5.2.2 Disadvantages

Defining Information Fragment Associations is an additional step in the bookmarking process, and although this Study has indicated that with the right tool this step is no burden to the user, it can still be considered a drawback even if it is a question of perception.

In spite of its advantages, Information Fragment Association may not be considered suitable for everybody or for everyday bookmarking. Although most subjects expressed the opinion the Information Fragment Association would be very useful, they were unsure whether they would use it themselves today if it was available to them.

5.3 LEARNING EXPERIENCES

5.3.1 Value of Using Familiar Interfaces

Every aspect of the Study has reinforced the notion that Information Fragment Association would constitute an invaluable component of future systems. It has proven that Information Fragment Association is not only providing the user with a new set of retrieval options, but it is also implementable at small cost to the bookmarking process or to the overall experience of the user. One lesson learned in this study, however, is that this functionality has to be incorporated within familiar interfaces, such as those used by traditional social bookmarking applications. In other words, the adoption of Information Fragment Association can be ensured only if blended together with functionality people are already using, maintaining the same look and feel.

5.3.2 User Targeting

Targeting the social bookmarking tool realm has proven that Information Fragment Association can improve mainstream tools significantly. However, it was apparent from the study that targeting the scholarly community would yield much more interesting results. Even the subjects of the study have pointed in that direction with their comments even though the Information Fragment Association was presented to them in its least sophisticated incarnation. It is evident that the concept of Information Fragment Association can better be grasped by a scholar whose more specific needs are met with such an interface that the average casual user. Regardless of the level of sophistication of the interface involved, however, the registry infrastructure can remain the same.

5.3.3 Appropriate Tools for Testing

Another lesson learned is one dealing with the Study's logistics and not with its objectives. In a case like the one at hand, in which one specific piece of functionality as opposed to the entire tool is being evaluated, it may be preferable to develop two test tools the one with the functionality being examined and the other without. That way the focus of the comparison will be placed on the object of interest and not on peripheral interface issues. This would also save a lot of manual work which is often necessary in those cases in which a system is not under the researcher's control, such was in the case of Task 4.

6 CONCLUSION

This dissertation represents an effort to determine the potential benefits of providing users with the capability of establishing and storing associations between Information Fragments which they have identified as bearing a certain relationship. These benefits were sought in everyday user activities such as social bookmarking and web searching. The effort concentrated in identifying the degree to which stored and easily retrievable Information Fragment Associations can be used to improve the performance and usability of the retrieval process and the search results of mainstream tools. The ultimate objective was to present these benefits as proof of the need for a comprehensive framework for the management of Information Fragment Associations.

Social bookmarking is a popular internet practice allowing users to share bookmarks to resources of interest. Many social bookmarking tools allow users to define an Information Fragment within the bookmarked page. However, the lack of the mechanism for recording the association between the Information Fragments prevents users who revisit these collections from recreating the mental association. This Study demonstrated that providing users with this missing functionality is indeed beneficial.

With the help of two subject groups, this Study has determined that users are able to find related Information Fragments much more easily when Information Fragment Association has been employed. Using questionnaires, the Study has revealed that users find that the presence of Information Fragment Association increases the ease of retrieval, the effectiveness and the enjoyment level of the process. Users also rated the usefulness of Information Fragment Association very high.

When two Information Fragments are explicitly identified as having some kind of relationship with each other, there is a reasonable expectation that within the universe of information resources there should be a relatively short navigational distance between these two Information Fragments. The results of this Study demonstrate unequivocally that with current Information Entity-based web searching tools the navigational distance between these two Informational Fragments is too large to be acceptable. The employment of Information Fragment Association infrastructure can make a significant contribution towards the resolution of this issue without undermining the current infrastructure.

The Study has also determined that the Recall and Precision of searches performed in datasets of Information Fragments is improved with the presence of Information Fragment Association. This has potential implications on future development of both social bookmarking and web searching tools.

Therefore, it is clear that Information Fragment Association implementation needs to move in two different directions. First, it needs to infiltrate social bookmarking tools and other mainstream applications such as search engines. Second it needs to be applied in the areas in which its full power can manifest itself, such as in tools designed specifically for scholarly work. In both of these cases, its success will depend on the degree to which the full feature set of the envisioned Morsoplexis Framework gets implemented.

Future directions should involve further testing, especially in the areas in which some weaknesses and omissions in the approaches taken by this Study have been detected. Based on such additional studies, implementation of fractions of the Morsoplexis Framework may take place. The Morsoplexis Framework as described in 0 calls for functionality beyond what was tested with this Study. The feasibility and usefulness of this functionality needs to be tested by future studies.

When Time Magazine decided that its “Person of the Year” for 2006 was “You” it signaled the recognition of the tremendous achievement generated by our collective work. Explaining this choice Lev Grossman wrote that the new Web is “a tool for bringing together the small contributions of millions of people and making them matter”[90]. However, instead of resting on our laurels, we need to ask ourselves: Are we really tapping into our full potential? Are there ways with which we can generate more contributions and make those contributions matter too? The Morsoplexis Framework provides a humble proposal for one possible way of taking advantage of users’ everyday interactions with information content.

APPENDIX A

Although this discussion has been moved to this appendix, it constitutes the core on which every other part of this dissertation was based.

0 A FRAMEWORK FOR GLOBAL IMPLEMENTATION OF FRAGMENT ASSOCIATION

This chapter presents the conceptual design for a framework aiming at providing a comprehensive solution supporting Information Fragment Association.

One of the major problems we are facing today is the lack of an easy way of collocating small pieces of interrelated information. A small fragment of a larger piece of work may contain specific information which is not necessarily of the same topic as the larger piece itself. It is therefore unlikely that with the interfaces currently available to users this specific piece of information is referenced adequately by other resources. The only tool currently available at our disposal is keyword searching. This is often an effective tool, but we can certainly easily conceive of two Information Fragments which according to a user bear a strong association with each other, yet contain no common keywords enabling them to be collocated by a keyword query.

As users browse the web, identify and consume useful information entities, they constantly create consciously or unconsciously mental associations between specific fragments within these Information Entities. Providing users with the capability of easily recording these associations will be beneficial both to them and to others. This can be achieved only by the introduction of a framework which incorporates all of the functionality necessary to make this association recording and management possible.

0.1 IMPORTANT CONCEPTS RELATING TO THE MORSOPLEXIS FRAMEWORK

These concepts along with concepts discussed earlier in this dissertation, especially in Section 3.7 for the basis on which this framework has been constructed.

0.1.1 Information Fragment Collection

It is often the case that a certain piece of information is contained in a multitude of content fragments scattered in different locations. These fragments complement each other in the creation of a single unit albeit not contiguous. The Information Fragment Collection has been conceived with the intention of providing a means of handling the need this situation generates. Of course, an Information Fragment Collection can be a collection of any Information Fragments regardless of the circumstances under which they have been brought together or the relationships between them. The assumption is that there is some commonality in them that let to their collection. As such, the collection can be treated and referred to as a single unit. In addition to Information Fragments, an Information Fragment Collection can contain other Information Fragment Collections. It can also contain Information Fragment Associations, which are defined in section 0.1.2 below.

0.1.2 Information Fragment Association

An Information Fragment Association is the result of the determination by a certain Individual that two Information Fragments bear a certain relationship which needs to be expressed with this association. The juxtaposition of these two interrelated Information Fragments makes some kind of a statement. Sometimes the statement is obvious, sometimes obscure. Information Fragment Associations can be classified in the following fashion:

- By Creator
    - Creator Associations
    - Third-Party Associations.
- By Complexity
    - Simple Associations
    - Complex Associations

0.1.2.1 Creator Associations vs. Third-Party Associations

Creator Associations are those between an Information Fragment of any Creator and an Information Fragment of the same Creator as the Creator of the Information Fragment Association itself. Third-Party Associations are those between two Information Fragments none of which having the same Creator as the Creator of the Association.

The different types of Information Associations can be used to organize collocated Information Fragments in different tiers. For example, let us look at the following scenario:

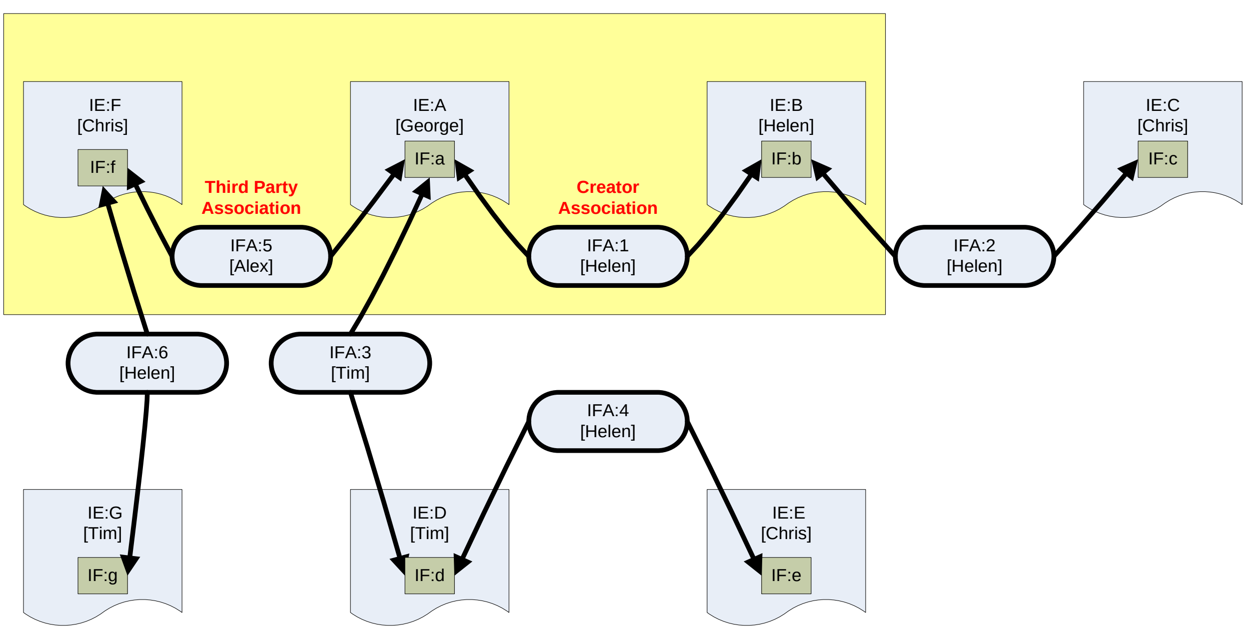


Figure 88 – Creator and Third Party Association

Figure 88 depicts a series of Information Fragment Associations as well as the creators of these Information Fragment Associations and the creators of the Information Fragments. A close look at this small cluster of Information Entities and Information Fragments demonstrates the difference between Creator and Third-Party Associations.

Helen is the creator of Information Fragment Association IFA:1 between Information Fragment IF:a and Information Fragment IF:b. She is also the creator of the Information Entity in which Information Fragment IF:b resides, while the Information Entity in which Information Fragment IF:a resides has been created by George. This is a “Creator” Association which means that Helen, the creator of the Information Fragment Association, has control over one of the two Information Entities involved. In this respect, it is similar to traditional web page linking or article referencing.

On the other hand, Alex is the creator of Information Fragment Association IFA:5 between Information Fragment IF:f and Information Fragment IF:a, but neither of the respective Information Entities in which these two Information Fragments reside has been created by Alex. This is an example of a “Third-Party” Association.

0.1.2.2 Simple vs. Complex Associations

A Simple Association is defined as one being established between two Information Fragments. A Complex Association is defined as one involving an Information Fragment Association or an Information Fragment Collection as at least one of the two items it associates.

The Morsoplexis Framework provides the capability of establishing an Information Fragment Association between an Information Fragment and a pre-existing Information Fragment Association or a pre-existing Information Fragment Collection. This functionality enables users to derive, refer to or comment on existing Information Fragment Associations and Collections.

For example, as Figure 89 demonstrates, a scholar X makes an Information Fragment Association between Information Fragments A and B, demonstrating the stylistic similarities of two passages of two well-known authors. While composing an Information Entity, another scholar Y makes an association between an Information Fragment C and the pre-existing association. Scholar Y, takes advantage of the fact that scholar X has already established this association. If scholar X is considered to be an expert in the field, using that association in order to make a certain point in Information Fragment C will provide this point with some support.

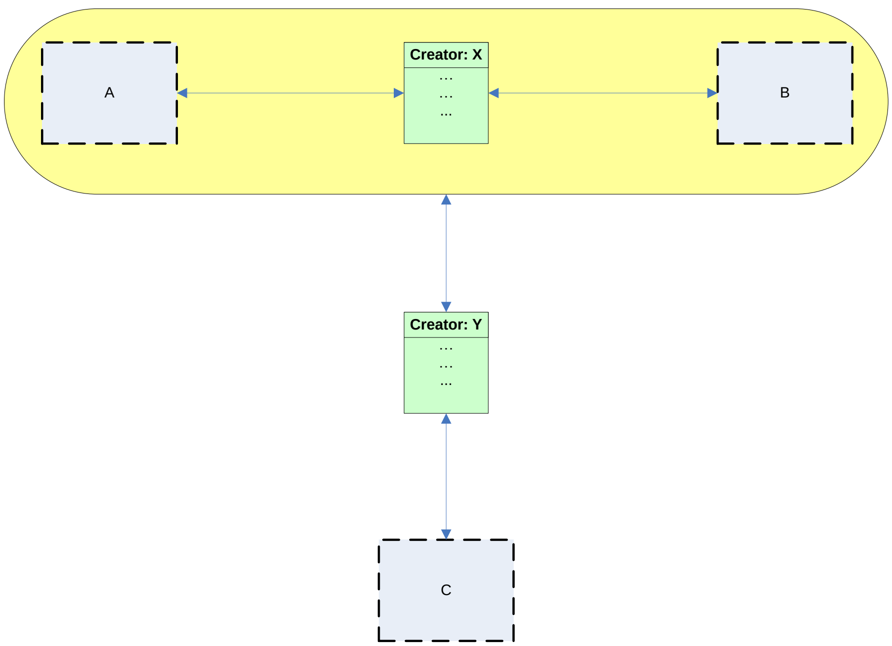


Figure 89 - Association between an Information Fragment and a Fragment Association

In brief, the following Information Fragment Associations are possible:

- Between an Information Fragment and an Information Fragment (Simple)
- Between an Information Fragment and an Information Fragment Association (Complex)
- Between an Information Fragment and an Information Fragment Collection (Complex)
- Between an Information Fragment Association and an Information Fragment Collection (Complex)

- Between an Information Fragment Association and an Information Fragment Association (Complex)
- Between an Information Fragment Collection and an Information Fragment Collection (Complex)

0.1.3 Trust and Reputation

Since the creation of Information Fragments, Information Fragment Associations and Information Fragment Collections is meant to express something, trusting the creator is particularly significant. The Morsoplexis Framework employs a simple but robust Identity Registry which allows for the definition of Identity Groups. These groups can be used to define what can be trusted and what cannot be trusted. Even though the determination of what can be trusted is external to the Morsoplexis Framework, the Identity Groups within the Morsoplexis Framework can store information conveying the degree to which an Identity and by extension the Information Fragment Association can be trusted.

0.2 THE NATURE OF THE CONCEIVED FRAMEWORK

This framework needs to extend the capabilities of the web, not provide an alternative to it. Therefore, it has to be open, functioning on top of the web and interacting with other resources and tools currently available on the web. Unlike other efforts dealing with content granularity, the essence of the envisioned framework lies in that it does not presuppose any pre-built structure in the resources it handles, thus becoming virtually universal in its scope. Any resource available on the web can be potentially available to users. This Framework places the definition of granularity with the consumer of the resource, not the author. This allows for an infinite number of possible fragments defined by information consumers. Although the author of a resource still has the capability of defining fragments and establishing fragment associations, any consumer of the resource also has the capability of defining fragments and optionally sharing them with other potential users.

0.3 FEATURES AND CHARACTERISTICS OF THE MORSOPLEXIS FRAMEWORK

In order to provide a better understanding regarding the functionality of this framework, it would be useful to outline the various features and characteristics this framework requires. The implementation of the features and characteristics listed below is deemed necessary in order to ensure that the envisioned functionality is achieved. Each one of these features will be built into and provided by a component of this framework. For this reason their discussion below is incorporated in the discussion of their encompassing framework components.

- General

  These features are introduced here as a general guidance for the design of the Morsoplexis Framework. They apply to or relate to all of the components. Each component of the Morsoplexis Framework is built with careful attention being paid to making sure that these features are available.

  - Simple implementation
  - Simple interface
  - Scalability
  - Global accessibility
  - Persistence of references (unique identifiers for Information Entities and Information Fragments)
  - Persistence of content (caching Information Entities)

    - Interaction with Expertise and Reputation Systems
    - The ability to publicly expose the Information Fragment Association data so that it may be used by other information systems such as search engines
- Creating Fragment Associations

  These features apply specifically to the interfaces designed for the purpose of allowing users to create and manipulate Information Fragment Associations
    - Ability to provide unique identifiers for its users
    - The ability to establish Associations between Information Fragments
    - The ability to establish Associations between an Information Fragment and a pre-existing Fragment Association
    - The ability to control access to (keep private or make public) Fragment Associations
    - The ability to attach attributes describing the nature of a Fragment Association
- Viewing and Maintaining Fragment Associations

  These features apply specifically to the interfaces designed for the purpose of allowing users to navigate through Information Fragment Associations. Such interfaces include extensions to traditional search engine result presentation.
    - The ability to view Information Fragments within Information Entities along with the Information Fragments with which they are associated
    - The ability to view Information Fragment Associations associated with an Information Fragment
    - The ability to delete an Information Fragment, Information Fragment Association or Information Fragment Collection. This is possible only if the item to be deleted is not itself one of the items being associated by another Information Fragment Collection. If this is the case, the option of retracting the item is available.
    - The ability to retract a Fragment Association

The flexibility of the Morsoplexis Framework allows for a variety of ways of interacting with it and taking advantage of the aforementioned functionality. This functionality is best demonstrated with an example. A conceivable usage scenario is the following:

- The user views an information entity A (or listens to or watches if the information entity is audio or video)
- At the same time, the user is viewing or creating another information entity B.
- The user selects an information fragment within the information entity A using a Fragment association client or utility.
    - The nature of this utility depends on the format of the information entity. For example, if the information entity is a document, this utility allows the user to select text in a fashion similar to the way a user normally selects text for copying and pasting.
    - Since this selection is meant to do more than copying and pasting it may be necessary that current browsers, word processors etc be enhanced to be able to interface with the proposed registries. Furthermore, if the information entity is audio or video, the same kind of capabilities will be needed using different types of "association creators".
- The user selects an information fragment within information entity B associating this fragment with the previously selected fragment.
- The same process is followed in the case of the creation of an association between an Information Fragment and a pre-existing Fragment Association. The interface should be able to handle this functionality as well
- The association information as well as any accompanying information entered by the user is saved in the registry
- The user is recorded as the creator of the fragment association

0.4 COMPONENTS OF THE MORSOPLEXIS FRAMEWORK

The purpose of the Morsoplexis Framework is to provide a globally accessible complement to the functionalities currently provided by the World Wide Web. It is a complement in the sense that it is not conceived of as replacing any of the resources currently available on the web or requiring that they be in any other format than the format they are currently in. Nor does it require that these resources be placed in some controlled enclosure. It simply has to work with what is in existence building on it instead of replacing it. The value of this framework chiefly relies on the combination of functionalities built into each one of its components. These components are interwoven providing a unique environment providing a set of capabilities currently unavailable to users in a comprehensive and easy to use package.

The components of the Morsoplexis Framework are the following:

- Global Distributed Registry
  - Distributed Registry of Registries
  - Distributed Qualifier Registry
  - Distributed Extra-Community Link Registry
- Community Registries
  - Identity Registry
  - Information Fragment Registry
  - Information Entity Caching Component
  - Local Registry of Registries
  - Local Qualifier Registry
  - Local Extra-Community Link Registry
  - Filter Registry
- Aggregators
- Information Fragment Association Creation Interfaces
- Information Fragment Association Viewing Interfaces

Each one of these components interacts and depends on the other components for full functionality to be achieved. Figure 90 provides a visual representation of these components, their relationships and place within the Morsoplexis Framework.

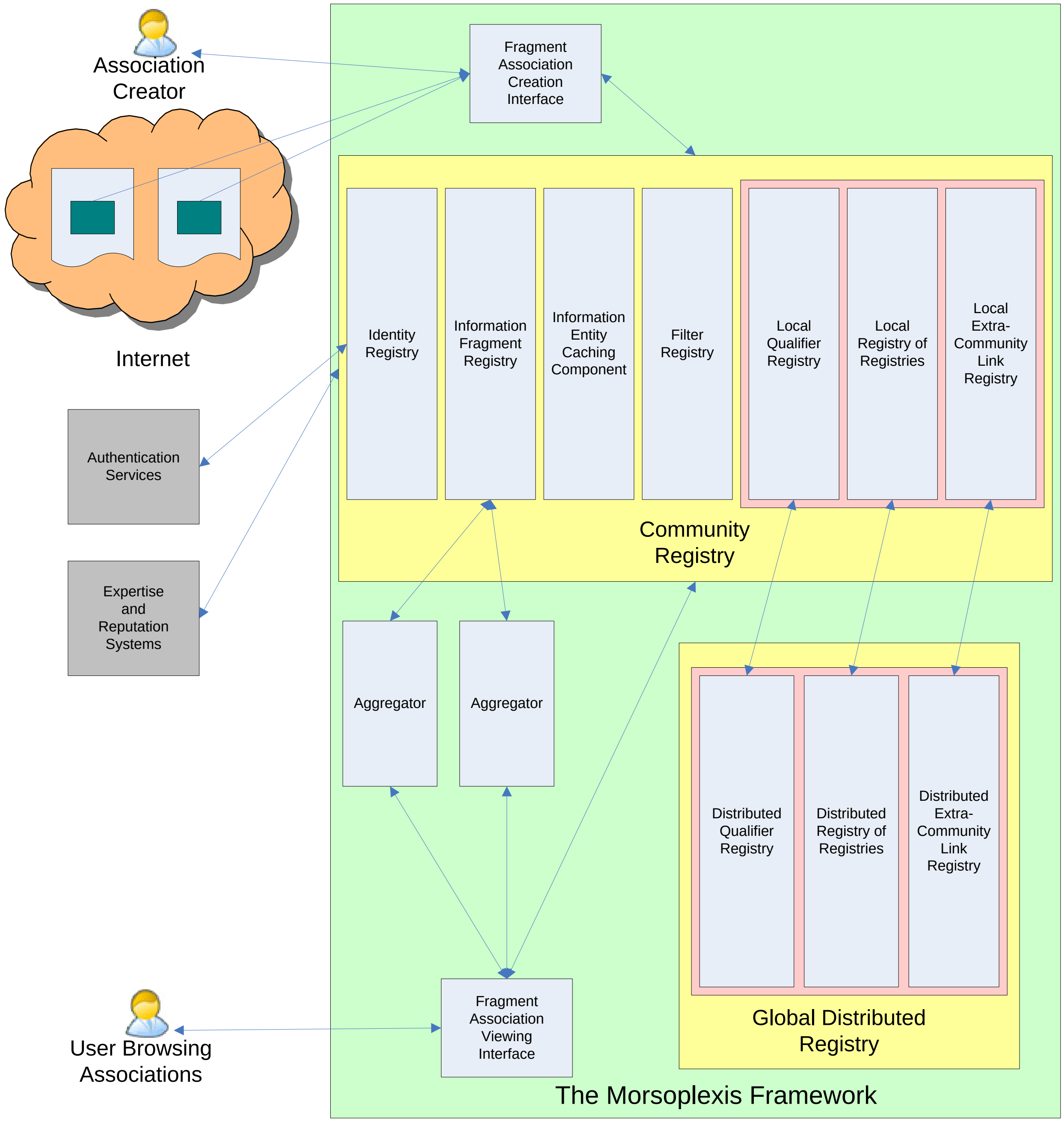


Figure 90 – Morsoplexis Framework Overview

The Morsoplexis Framework is designed to provide both self-sufficiency and universality. Its self-sufficiency lies in the capability of a Community Registry to function by itself with its entire feature set without necessarily interacting with the Global Distributed Registry or any other

Community Registries. Its universality lies in the capability of the Community Registries to fully integrate with the web and to interact with each other either directly or my means of the coordinating Global Distributed Registry. This coordination is necessary in order to ensure that the independent nature of Community Registries and the fact that they are built not to rely on a centralized resource does not come in the way of interoperability.

The core of the Community Registry is the *Information Fragment Registry*. This is where the definitions for Information Fragments, Information Fragment Associations and Information Fragment Collections are stored. An Information Fragment definition specifies the Information Entity to which that Information Fragment belongs. This allows for a viewer used to display an Information Fragment to query first the Information Fragment Registry, retrieve the Information Fragment definition, then using the identifier for the Information Entity to retrieve the Information Entity entry and consequently load the cached resource. Using again the Information Fragment definition, the boundaries of the Information Fragment are marked and it is displayed within the context of the cached Information Entity. The *Information Fragment Registry* and the *Information Entity Caching Component* are therefore being used side-by-side. They are also intertwined by the queries submitted to the system for the purpose of navigation from fragment to fragment. Each Information Entity will potentially contain a lot of Information Fragments.

A user can browse a listing of all Information Fragment Associations established for a specific Information Entity and can also filter these displayed results by applying Filters from the *Filter Registry*. Since Information Fragments and Information Fragment Associations specify the creator Identity and optionally incorporate Descriptive Qualifiers, these Filters, being lists of Identities and Qualifiers can narrow down the results of the query.

The *Identity Registry* provides authentication for entry to a Community Registry and subsequently ensures that every item created by an Identity remains under that Identity's control. The entry for each item such as an Information Fragment, Information Fragment Association etc always contains the identifier of the Identity who created it. The Identity Registry also handles the Identity Recommendations and the interaction with external Expertise and Reputation Systems as well as the interaction with external Authentication Services.

The *Local Qualifier Registry* contains collections of Qualifiers used to describe the content of Information Fragments, Information Fragment Associations and Information Fragment Collections (Descriptive Qualifiers) or used to describe the relationship between two associated items (Relationship Qualifiers). The Local Qualifier Registry is being synchronized with the Distributed Qualifier Registry. The same is the case with the *Local Registry of Registries* and the *Local Extra-Community Link Registry*, which synchronize with their respective distributed counterparts in the Global Distributed Registry. The Registry of Registries is the official list of Community Registries which also provides URI resolution. The Extra-Community Link Registry provides the ability of creating Fragment Association and Fragment Collections across Community Registries. These are handled as special kind of links the integrity of which is not completely guaranteed, but enough measures to maintain reliability are in place.

The interaction with the user is achieved through the *Fragment Association Creation Interface*, for the process of creating Information Fragment Associations and the *Fragment Association Viewing Interface* for the process of browsing and searching the contents of the registries.

Finally, the *Aggregators* interact with the Information Fragment Registry of every Community Registry, gathering the contents into a single searchable large registry. These Aggregators can be global or subject specific.

0.5 GLOBAL DISTRIBUTED REGISTRY

The Global Distributed Registry is a globally accessible component which binds and coordinates the otherwise independently operating Community Registries. It is distributed so that its availability and scalability may be ensured. It is conceived of as the central knot connecting the Community Registries. Its function is that of coordination and discovery assistance and the Morsoplexis Framework should still be able to function even if this component is unavailable or inaccessible.

The Global Distributed Registry would function in a fashion very similar to that of the DOI (Digital Object Identifier) System[91]. Each Information Fragment as well as any Morsoplexis information item of any type would be identified by a combination of that item's unique identifier as the suffix and the unique identifier of the Community Registry in which it resides as the prefix. The format of the Community Registry identifiers is "[a-zA-Z]{3}[0-9]{4}" and the format of prefix is that of the Universally Unique Identifier (UUID)[92] . For example, an Information Fragment Association can have an identifier:

"f81d4fae-7dec-11d0-a765-00a0c91e6bf6"

and belong to a Community Registry with an identifier "xyz7890"

The two identifiers are presented as a single unit separated by a "/" just as in the case of DOI names:

“xyz7890/f81d4fae-7dec-11d0-a765-00a0c91e6bf6”

Both the Distributed Registry of Registries and the Local Registry of Registries found in every Community Registry will be able to provide resolution for the Community Registry identifier providing the URI of the Community Registry.

The Distributed Registry of Registries and the Distributed Qualifier Registry are not necessarily related. They are being bundled together in the Global Distributed Registry not because of their need to interoperate but because of their position in the Morsoplexis Framework and the similarity of their mission. They simply operate centrally and they are global providing their respective services to the Community Registries. Their bundling does not necessarily have to be physical, it can be purely conceptual. However, all three subcomponents always share the following common properties:

- Lightweight (for manageability)
- Distributable (for service reliability, performance and scalability)
- Accessible to the Community Registries

Placing the Distributed Qualifier Registry physically together with the Distributed Registry of Registries has some technical management advantages, such as coordination of backup and replication. It also makes economic sense to keep these two relatively small registries on the same server and it is simpler for Community Registries to point to one server instead of two. On the other hand, it may prove beneficial in the future to physically separate the two, if the Distributed Qualifier Registry can be used for other purposes unrelated to the needs of the Morsoplexis Framework.

The Global Distributed Registry along with the Aggregators are resources expected to be of value to web search engines contributing to the quality of their results.

0.5.1 The Distributed Registry of Registries

The Distributed Registry of Registries is a kind of hyperdatabase providing a listing of all available Community Registries along with access and contact information for them. A Community Registry can be moved from one server to another (with different URI) but access to the Community Registry will not be affected. Retrieval tools can be designed to check the Distributed Registry of Registries if a direct attempt to retrieve something from a Community Registry fails. In addition to access and contact information, this registry also maintains other statistics regarding each Community Registry, such as technical service availability.

Each Community Registry has a unique identifier which is registered in the Distributed Registry of Registries. Within the global Framework, Information Fragment Associations and other items are referenced with their own unique identifier plus the unique identifier of the Community Registry in which they reside in a fashion similar to that used by the DOI (Digital Object Identifier) standard[93]. The same is the case with Identities and cached Information Entities. The Distributed Registry of Registries acts as a resolver retrieving information from the right Community Registry.

An entry of the Distributed Registry of Registries is of the following structure:

- Registry Identifier
- Registry URI
- Creation and update dates
- Registry Description
  - Description
  - Descriptive Qualifiers
- Contact Information
- Service availability information

The schema for the Registry of Registries is presented in Table 23 :

```xml
<?xml version="1.0" encoding="utf-8"?>
<xs:schema id="RegistryOfRegistries"
        targetNamespace="http://morsoplexis.org/schema/RegistryOfRegistries.xsd"
        elementFormDefault="qualified"
        xmlns="http://morsoplexis.org/schema/RegistryOfRegistries.xsd"
        xmlns:xs="http://www.w3.org/2001/XMLSchema">

  <xs:simpleType name="mpKey">
    <xs:annotation>
      <xs:documentation xml:lang="en">
        The Format of the Global Identifier used by most Morsoplexis items
      </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:pattern value="[a-fA-F0-9]{8}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{12}"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:simpleType name="mpCRKey">
    <xs:annotation>
      <xs:documentation xml:lang="en">
        The Format of the Community Registry Identifier
      </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:pattern value="[a-zA-Z]{3}[0-9]{4}"/>
    </xs:restriction>
  </xs:simpleType>


  <xs:element name="CommunityRegistry">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="CreationDateTime" type="xs:dateTime" nillable="false" />
        <xs:element name="UpdateDateTime" type="xs:dateTime" nillable="false" />
        <xs:element name="RegistryName" type="xs:string" default="" />
        <xs:element name="RegistryURI" type="xs:anyURI" />
        <xs:element name="Description" type="xs:string" />
        <xs:element name="Qualifier" minOccurs="0" maxOccurs="unbounded">
          <xs:complexType>
            <xs:attribute name="QualifierId" type="mpKey" />
          </xs:complexType>
        </xs:element>
        <xs:element name="ContactInformation">
          <xs:complexType>
            <xs:sequence>
              <xs:element name="ContactName" type="xs:string" />
              <xs:element name="ContactTitle" type="xs:string" />
              <xs:element name="ContactEmail" type="xs:string" />
            </xs:sequence>
          </xs:complexType>
        </xs:element>
        <xs:element name="ServiceAvailability">
          <xs:complexType>
            <xs:sequence>
              <xs:element name="LastTimeChecked" type="xs:dateTime" />
              <xs:element name="ServiceRating" type="xs:int" />
            </xs:sequence>
          </xs:complexType>
        </xs:element>
      </xs:sequence>
```

```
    <xs:attribute name="CommunityRegistryId" type="mpCRKey" />
  </xs:complexType>
 </xs:element>

</xs:schema>
```

Table 23 - Schema for the Registry of Registries

0.5.2 Distributed Qualifier Registry

This is a centralized depository of collections of descriptive terms to be used as attributes for Information Fragments, Information Fragment Associations and Information Fragment Collections. These terms are presented to the user as options during the fragment association process. Placing these simple term lists called "Qualifier Collections" in the Global Distributed Registry encourages consistency. Identities will have the capability of compiling Qualifier Collections specific to their particular disciplines, but they will also have the capability of importing standard well-established Thesauri and Taxonomies. This registry will adopt a minimalist approach to the structuring of such lists. The structure will attempt to address in a simple way the most basic features of a controlled vocabulary found in the NISO Z39.19 guidelines[94]. It is expected that Morsoplexis Framework implementations will have several major Thesauri and Taxonomies built into the Distributed Qualifier Registry. The test implementation to be produced by this project will attempt to include the *North American Industry Classification System* taxonomy[95] and if authorization is secured the *Thesaurus of ERIC Descriptors*[96] and the UNESCO Thesaurus[97].

Implementations are expected to replicate this registry locally in order to enhance performance. New Qualifier Collections are created in the Community Registries and then synchronized with the Distributed Qualifier Registry contributing to the creation of a large central depository of these simple Qualifier Collections. Flexibility is provided for locally defined Qualifier Collections to be used by Community Registries as "CommunityRegistryOnly", thus not being contributed to the Distributed Qualifier Registry. A Community Registry has the option of changing the designation of a Qualifier Collection to non local. The full structure and schema of this registry is provided in section 0.6.5 below in the discussion of the Local Qualifier Registry.

0.5.3 Distributed Extra-Community Link Registry

The Distributed External Link Registry coordinates the establishment of external links. A description of the process as well as the schema is provided below in section 0.6.6 discussing the Local Extra-Community Link Registry.

0.6 COMMUNITY REGISTRIES

The Community Registries are the main component of the Morsoplexis Framework. They are the location in which all of the user, association and information entity caching are stored. The Morsoplexis Framework provides individual communities with the capabilities needed for the creation of highly-functional registries. It also allows for the definition of the role these registries play in the wider space of all internet resources. The Identities representing the creators of Fragment Associations are expected to be ranked by Trust/Reputation systems in the future. Since the Morsoplexis Framework is global and the Community Registries are locally managed, it is expected that the Community Registries themselves will be ranked by Trust/Reputation systems as well.

Some of the main characteristics of Community Registries are the following:

- Each Community Registry has a unique identifier. This makes any Identities, Information Entities or Information Fragment Associations globally unique, since they are referenced by their unique identifier plus the identifier of their Community Registry.
- Two or more Community Registries are able to be merged into one. It is conceivable that two Community Registries likely need to be merged or conversely a single Community Registry likely needs to be split into two or more Registries. This need may stem out of organizational developments, for example two companies merging, or simply because of technical maintenance and management reasons.
- An Identity, the Information Fragments, Information Fragment Associations and other items created by this Identity as well as their cached content should be able to be migrated to another Community Registry. Administrative utilities available to Community Registry operators will allow the operator of Community Registry A to provide the operator of Community Registry B access to all data owned by a specific Identity for migration. All entries will be copied exactly as they are with the only adjustment being the change of their "CommunityRegistryId" attribute. This means that they will keep their existing unique identifier but they will now have a new Community Registry Id – Item Id combination which will resolve to the new Community Registry. The entries in the old Community Registry will be deleted and new entries will be added to the *Moved Items Registry*. This registry, described in section 0.6.2.4 below will record the ids of all items moved along with the id of the new Community Registry, and it will be used for redirection to the new Community Registry.

0.6.1 Identity Registry

Each Community Registry has a built-in Identity Registry. This registry contains the listing of all Identities used by the Community Registry as a whole, an Identity being a user and contributor to the Community Registry. Of Course, an Individual using the Morsoplexis Framework has the option of assuming multiple Identities, but only one Identity is available to the Individual in a given Fragment Association Creation Interface session. In other words, the user logs in as one

Identity, and during that given session every activity that user engages in is recorded as belonging to that Identity.

Identity Registries:

- make possible the registration of universal and unambiguous Identities to be used in the Fragment Association creation process, ensuring that each Fragment Association is attributed to an Identity
- interface with external identity/authentication services
- interface with external reputation/trust systems
- define hierarchical relationships between Identities (e.g. membership to a group)

These features make the Identity Registry an indispensable part of the Morsoplexis Framework by creating an environment which may provide more information useful for the ranking of the Fragment Associations.

The simplest implementation of an Identity Registry comprises of a listing of unique Identifiers, their relationships and optional basic personal information and username/password information. Proper implementations map Identities to external authentication systems, to take advantage of password encryption and better management tools, but the maintenance of the unique identifiers representing Identities always take place in the Identity Registry.

The structure of a Registry entry is approximately the following:

- Identifier (required)
- Creation and update dates/times
- Mapping to external Authentication Service (if this capability is implemented)
    - Username/Identifier
    - Authentication Service Identifier
- Descriptive Information
    - Name
    - Contact Information
        - E-mail (required)

        - Other Internet contact information, such as web page IM etc.
        - Mailing address
        - Other contact information
    - Affiliations
- Trust / Reputation Relationships

The schema for the Identity Registry is presented in Table 24:

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema id="Identities"
        targetNamespace="http://morsoplexis.org/schema/Identity.xsd"
        xmlns="http://morsoplexis.org/schema/Identity.xsd"
        elementFormDefault="qualified"
        xmlns:xs="http://www.w3.org/2001/XMLSchema">

  <xs:simpleType name="mpKey">
    <xs:annotation>
      <xs:documentation xml:lang="en">
        The Format of the Global Identifier used by most Morsoplexis items
      </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:pattern value="[a-fA-F0-9]{8}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{12}"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:simpleType name="mpCRKey">
    <xs:annotation>
      <xs:documentation xml:lang="en">
        The Format of the Community Registry Identifier
      </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:pattern value="[a-zA-Z]{3}[0-9]{4}"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:element name="Identity">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="CreationDateTime" type="xs:dateTime" />
        <xs:element name="UpdateDateTime" type="xs:dateTime" />

        <xs:element name="Name">
          <xs:complexType>
            <xs:sequence>
              <xs:choice>
                <xs:element name="Personal">
                  <xs:complexType>
                    <xs:sequence>
                      <xs:element name="LastName" type="xs:string" nillable="true" />
                      <xs:element name="FirstName" type="xs:string" nillable="true" />
                      <xs:element name="MiddleName" type="xs:string" nillable="true" />
                    </xs:sequence>
                  </xs:complexType>
                </xs:element>
```

```xml
        <xs:element name="Organization">
          <xs:complexType>
            <xs:sequence>
              <xs:element name="OrganizationName" type="xs:string" nillable="true" />
              <xs:element name="OrganizationFunction" type="xs:string" nillable="true" />
            </xs:sequence>
          </xs:complexType>
        </xs:element>
      </xs:choice>
    </xs:sequence>
  </xs:complexType>
</xs:element>

<xs:element name="Authentication">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="UserName" type="xs:string" nillable="true" />
      <xs:choice>
        <xs:element name="AuthenticationService" type="xs:string" nillable="true" />
        <xs:element name="Password" type="xs:string" nillable="true" />
      </xs:choice>
    </xs:sequence>
  </xs:complexType>
</xs:element>

<!-- Multiple Electronic Addresses are allowed -->
<xs:element name="Email" maxOccurs="unbounded">
  <xs:complexType>
    <xs:attribute name="EmailAddress" type="xs:string" />
  </xs:complexType>
</xs:element>
<xs:element name="HomePage" maxOccurs="unbounded">
  <xs:complexType>
    <xs:attribute name="HomePageAddress" type="xs:string" />
  </xs:complexType>
</xs:element>
<xs:element name="IM" maxOccurs="unbounded">
  <xs:complexType>
    <xs:attribute name="IMAddress" type="xs:string" />
  </xs:complexType>
</xs:element>

<xs:element name="PhysicalAddress">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="Street" type="xs:string" nillable="true" />
      <xs:element name="City" type="xs:string" nillable="true" />
      <!-- Choice of different Address Formats -->
      <xs:choice>
        <xs:element name="State" type="xs:string" nillable="true" />
        <xs:element name="Province" type="xs:string" nillable="true" />
      </xs:choice>
      <xs:choice>
        <xs:element name="Zip" type="xs:string" nillable="true" />
        <xs:element name="PostalCode" type="xs:string" nillable="true" />
      </xs:choice>
    </xs:sequence>
  </xs:complexType>
</xs:element>

<xs:element maxOccurs="unbounded" name="Affiliations">
  <xs:annotation>
```

```xml
        <xs:documentation xml:lang="en">
          Multiple Affiliations are allowed
        </xs:documentation>
      </xs:annotation>
      <xs:complexType>
        <xs:attribute name="Affiliation" type="xs:string"/>
      </xs:complexType>
    </xs:element>

    <xs:element maxOccurs="unbounded" name="Members">
      <xs:annotation>
        <xs:documentation xml:lang="en">
          Members of this Identity
        </xs:documentation>
      </xs:annotation>
      <xs:complexType>
        <xs:attribute name="MemberIdentityId" type="mpKey"/>
      </xs:complexType>
    </xs:element>

   </xs:sequence>

   <xs:attribute name="IdentityId" type="mpKey" use="required">
    <xs:annotation>
      <xs:documentation xml:lang="en">
        The unique identifier for the Identity Entry
      </xs:documentation>
    </xs:annotation>
   </xs:attribute>
   <xs:attribute name="CommunityRegistryId" type="mpCRKey" use="required">
    <xs:annotation>
      <xs:documentation xml:lang="en">
        The unique identifier for this Identity's Community Registry
      </xs:documentation>
    </xs:annotation>
   </xs:attribute>
  </xs:complexType>
 </xs:element>

</xs:schema>
```

Table 24 - Identity Registry Schema

The Morsoplexis Framework aims at providing the following features which are dependent of the Identity Registry:

- Creation of Identities
- Interface with External Authentication Services
- Interface with External Reputation/Trust systems
- Trust / Reputation Relationships

0.6.1.1 Creation of Identities

An Identity is a simple and unambiguous unique identifier. This identifier is attached to whoever logs into the Fragment Association Creation Interface with a certain set of credentials. An Individual logging in and assuming this Identity is permanently associated with the Fragment Associations s/he creates. The Fragment Association Creation Interface allows an Individual to create a new Identity which represents that Individual or a group of which this Individual is a member and designated editor.

0.6.1.2 Interface with External Authentication Services

Maintaining local usernames and passwords within an implementation of the Morsoplexis Framework would only contribute to the chaotic multitude of passwords in use today. Mapping to external multi-use authentication systems is the most desirable implementation. The Morsoplexis Framework makes provision for this capability and strongly recommends it. However, it does not dictate one authentication scheme over the other as long as the integrity of the Identities being created in the Identity Registry remains intact. The mapping is conceived of as being achieved by a design as simple as the inclusion of an identifier pointing to the authentication system of choice and the username/identifier that the authentication system uses in the Identity Registry entry. The implementation of the Fragment Association Creation Interface ensures that a user is able to assume a given framework Identity only when providing the right credentials to the authentication system specified in the entry for that Identity. An implementation incorporating the upcoming CardSpace[98] functionality would greatly enhance the user experience.

0.6.1.3 Interface with External Reputation/Trust systems

One of the most important functions of the Identity Registry will be its interaction with Expertise and Reputation systems. Trust and Reputation are social concepts going back to the beginning of civilization. Regardless of how one defines these two concepts, they are always related to the amount of information available about the object of trust.

In small communities trust and reputation are easily built because of direct interaction. In larger communities, it is inevitable to very often consult a third party, and often the acquaintance of this third party with the object of trust is also indirect. Although this indirect information is still very valuable, in the determination of trust, it is often complemented by the consideration of one's group affiliations.

Reputation is a factor contributing to the building of Trust which is the prerequisite for any interaction direct or indirect. This interaction can be the exchange of goods, services, or information. The advances in means of transportation, communication and storage and retrieval of the written word have significantly increased the challenge of establishing trust. The options are now too many to be able to rely on personal acquaintance or experience.

Reputation Systems have been introduced primarily to help identify trustworthy exchange partners, covering both business-to-business and business-to-consumer interactions. Their focus has been commercial, but in the cases in which the object of commercial transaction is an information resource, such as a book sold on Amazon, rankings can be useful outside the context of a pure financial transaction. However, when it comes to information resources, simple rankings are not sufficient. A consumer of a piece of merchandise may provide good feedback regarding that item. A consumer of a piece of information is very likely to be the author of other pieces of information. The relationships between these Information Entities and the rankings of their authors can be complex. Therefore a simple mechanism of recording these relationships is needed.

The proposed framework is not a reputation system, but rather a necessary complement to reputation systems by providing a simple and consistent interface between reputation systems and information seeking agents. The plethora of such reputation systems and the diversity of approaches involved make it difficult for search agents to be able to take advantage of them collectively without significant customization. Several reference models for reputation systems have been proposed[99, 100], but most of them are geared towards specific reputation contexts, such as e-commerce transactions and are in general complex and incompatible. The Identity Groups envisioned as part of the proposed framework offer a simple and scalable way of applying attributes containing reputation information regardless of the way this information has been collected or processed. A reputation ranking system may constitute an Identity Group in the Registry. This Identity Group may contain Identity Groups for each one of the ranked categories. The ranking system Identity Group may itself be contained in Other Identity Groups specifying its context-relevance and reliability as a ranking system.

0.6.1.4 Trust / Reputation Relationships

The Relationships defined in the Identity Registry are purely meronymic. They are dealing with membership in a collection and not membership in a class. Winston, Chaffin and Herrmann[101] draw the distinction between membership in a collection and membership in a class: “membership in a collection is determined by spatial or temporal proximity or by a social connection (e.g., tree-forest, cow-herd), characteristics which are extrinsic to the individual members themselves. Membership in a class, in contrast, is determined by similarity to the other members on one or more intrinsic property”.

An Identity is thus defined as being a member of a certain group as opposed to belonging to a class of Individuals having a common property. In other words, for example a certain Identity representing a specific woman is defined as a member of a professional society to which she belongs, but not as belonging to the class “women”. Furthermore, an Identity is either an Individual or a group of Individuals containing Individuals or other groups.

Since the prerequisite for inclusion in a group is usually the fulfillment of a certain set of conditions, trust for an individual may be inherited from the group that the individual belongs to. As Figure 91 shows, I may trust Author C and Author D who belong to Professional Society X more than other authors because I happen to know that all members of the Professional Society X are required to meet certain criteria for admission, one of those criteria being the requirement that they have received an award for one of their writings. However, the real reason why I trust Professional Society X is the fact that it is being recommended by the Academic Department Y, which I know I can trust because I happen to be a student in that department and I have placed Academic Department Y in a Filter I have defined for myself. It may happen that there is information residing outside the Morsoplexis Framework that I do not have. For example, as this figure depicts, the listing of all Award Recipients is not inside the Morsoplexis Framework, therefore I do not know that Author A has received an award. The membership of Professional Society X is available to the Morsoplexis Framework and it provides me with the information, albeit incomplete. All members of Professional Society X are Award Recipients, but not all Award Recipients are members of Professional Society X. Ideally, if some organization is compiling a list of all Award Recipients this list can be made directly available to the Morsoplexis Framework.

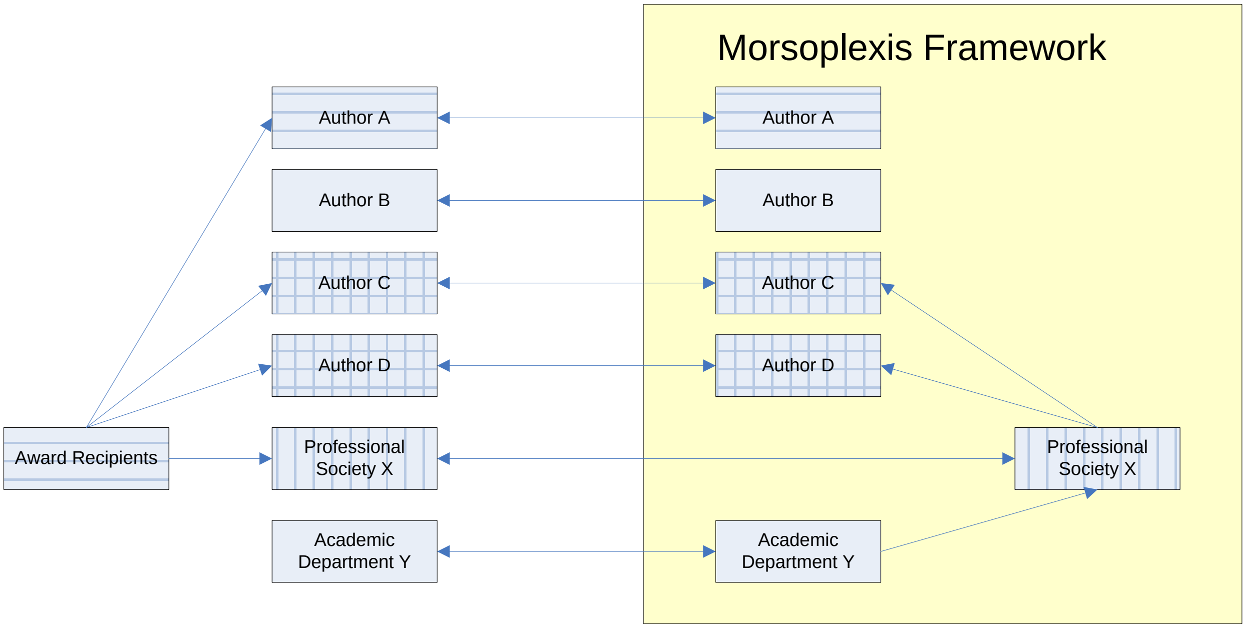


Figure 91 - Example of Part-Whole Relationships within and out the Morsoplexis Framework

The above example is rather simple, taking into consideration just one criterion, that of receiving an Award. The existence of this information is important and certainly much better than no information, but a thorough evaluation of an author is often a complex process involving the consideration of a large number of criteria. This process yields a ranking relevant to some specific query. This ranking can presented to the Morsoplexis Framework by the Reputation/Trust systems residing outside the Morsoplexis Framework. In

Figure 92, such a Trust/Reputation system is being roughly depicted. It represents the process of evaluation of authors by Peer Review Panel Z. The relationship between this system and the Morsoplexis Framework is such that an Identity is being created in the Morsoplexis Framework by the Trust/Reputation System. This Identity is called “Ranked List of Authors Compiled by the Peer Review Panel Z”. This Identity contains as members the Identities of the Authors it lists. The Identities of the Authors also have to be in the Morsoplexis Framework. This ranked list is being produced by the Peer Review Panel Z, the rankings of which are recommended by Academic Department Y which I know I can trust and which I have specified in my Identity Filter. Another ranked list in the Morsoplexis Framework, produced by Peer Review Panel W is not recommended by Academic Department Y, so its ranking will be given much lower weight

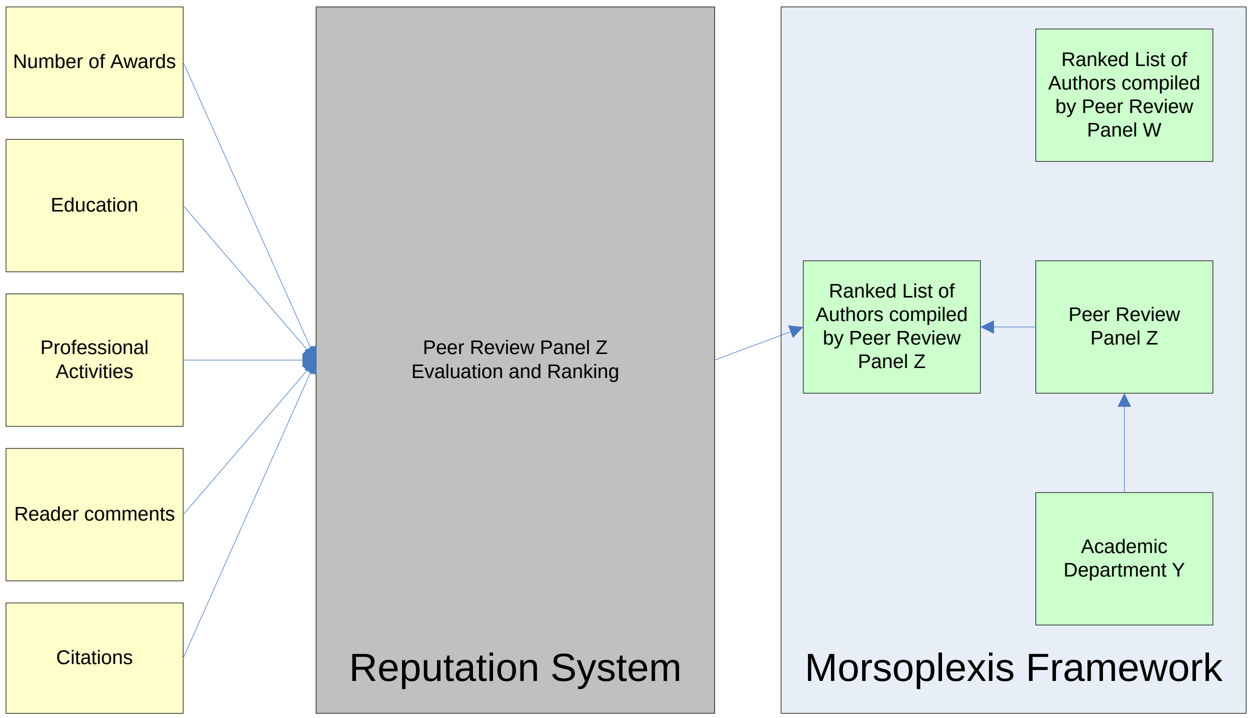


Figure 92 - Trust/Reputation System Interaction

The example in

Figure 92 above demonstrates the contribution of Trust/Reputation systems to the Morsoplexis Framework. However, it has to be pointed out that this relationship is conceived of as reciprocal. Fragment Associations created by users within the Morsoplexis Framework may be used by Reputation Systems as another factor contributing to their ratings. For example, any Fragment Associations created within the Morsoplexis Framework involving the Identities of these Authors and affecting their reputation positively or negatively can be used by a Reputation System. In the example above, we know that Peer Review Panel Z can be found in the

Reputation System. If a highly ranked author receives a lot of negative comments through Fragment Associations, the Reputation System will take that into consideration as it recompiles its rankings.

Individual users, Aggregators and Search Engines are interacting with the Community Registries. The anticipated accumulation of a large number of Fragment Associations and the desire to focus on information that is both relevant with respect to a subject area and reliable with respect to a community with common goals and needs leads to the need for establishing usage Filters. These Filters restrict and rank the information yielded by searches involving Fragment Associations.

The Morsoplexis Framework is not a reputation system, nor do any of its components attempt to simulate the functionality of reputation systems. What it offers is an interface allowing for the utilization of data extracted from reputation systems for the purpose of filtering and sorting Information Fragments. A reputation system determines a rating for a specific person by taking into consideration a wide range of factors involving that person’s various characteristics and capabilities. Gail Rein provides a very lucid reference model for reputation information systems[92].

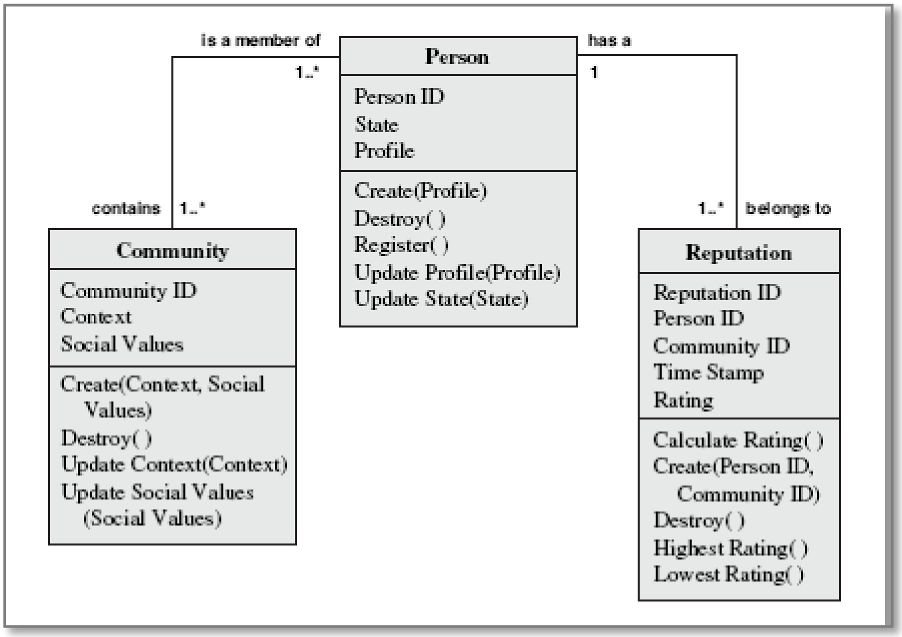


Figure 93 - Gail Rein's person, community and reputation classes[43]

The classes displayed in Figure 93 contain detailed information about the person and the community the person belongs to. For example, "Profile" which appears as an attribute of the "Person" class contains four arrays of statements describing a person's knowledge, experience, credentials and connections. This information along with "Context" and "Social Values" is used for calculating the rating. Morsoplexis does not need the detailed information processed by the reputation systems. What it needs is the ability to map to three basic properties:

- the identifier of the person
- the rating
- (and optionally) a single term descriptive of that person's ranked capabilities and expertise

In order to demonstrate this mapping and to conceptualize the meronymic and recommendation relationships between Identities, it will be helpful to express the example of

[43] From [99] G. L. Rein, "A reference model for designing effective reputation information systems," *Journal of Information Science,* vol. 31, pp. 365-380, October 1, 2005 2005. http://jis.sagepub.com/cgi/content/abstract/31/5/365

Figure 92 first using Resource Description Framework (RDF) syntax, and then present the way the relationship it expresses can be imported into and handled by the Morsoplexis Framework. Table 25 is an RDF representation of the above examples. Assuming that Academic Department Y (item #6, urn-ADY, shaded) is the Organization we trust, we can set it as our starting point.

We see that items #8 and #9 contain statements asserting that two other resources, the Data Security Experts Group (urn-DSE) and Peer Review Panel Z (urn-PRPZ) are being recommended by Academic Department Y. For our purposes, the Peer Review Panel Z is some kind of reputation system, even if as its name suggests some of the calculation of one's reputation is contributed directly by humans. We see that Peer Review Panel Z (urn-PRPZ) is listed as the recommender in items #10, #11, #12 and #13. These items specify that the respective Authors have each their designated trustworthiness rating (xyz:recommendationLevel) and the expertise area in which the recommendation is applicable (xyz:isRecommendedFor). This provides a functionality context for each one of these recommendations specifying not just how reliable an Identity is but also how relevant this reliability is to the specific need at hand. In the case of the Data Security Experts Group, we see that item #5 contains its description and lists its membership, which consists of Author C and Author D. In this case, the meronymic relationship between the Data Security Experts Group and the two authors is an implicit recommendation.

In this example we see that Author B and Author C are being highly recommended as Authors by the Academic Department Y, but Author C is also a member of the Data Security Experts Group which itself is recommended for "Software: Data Security" (code number 17264250) as listed in Thomas Register[102]. Therefore, Author C can be ranked higher because s/he is highly recommended both for knowledge on Information Systems in general (ID 2478 in OECD's Macrothesaurus)[103], and specifically for Data Security.

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:fa="http://purl.org/xyz/elements/1.0/"
xmlns:dc="http://purl.org/dc/elements/1.0/">

1.  <rdf:Description rdf:ID="urn-A">
      <xyz:name>Author A</xyz:name>
    </rdf:Description>

2.  <rdf:Description rdf:ID="urn-B">
      <xyz:name>Author B</xyz:name>
    </rdf:Description>

3.  <rdf:Description rdf:ID="urn-C">
      <xyz:name>Author C</xyz:name>
    </rdf:Description>

4.  <rdf:Description rdf:ID="urn-D">
      <xyz:name>Author D</xyz:name>
    </rdf:Description>

5.  <rdf:Description rdf:ID="urn-DSE">
      <xyz:name>Data Security Experts Group</xyz:name>
      <xyz:hasMember rdf:resource="urn-C" />
      <xyz:hasMember rdf:resource="urn-D" />
    </rdf:Description>

6.  <rdf:Description rdf:ID="urn-ADY">
      <xyz:name>Academic Department Y</xyz:name>
    </rdf:Description>

7.  <rdf:Description rdf:ID="urn-PRPZ">
      <xyz:name>Peer Review Panel Z</xyz:name>
      <xyz:recommends rdf:resource="urn-XRLM" />
    </rdf:Description>

8.  <rdf:Description rdf:about="urn-DSE">
      <xyz:isRecommendedFor rdf:resource="http://www.thomasnet.com/products/software-data-security-
17264250-1.html" />
      <xyz:isRecommendedBy rdf:resource="urn-ADY" />
      <xyz:recommendationLevel rdf:resource="http://purl.org/xyz/elements/1.0/#RecommendationLevel3" />
    </rdf:Description>

9.  <rdf:Description rdf:about="urn-PRPZ">
      <xyz:isRecommendedFor rdf:resource="http://purl.org/xyz/elements/1.0/#General_Recommendation" />
      <xyz:isRecommendedBy rdf:resource="urn-ADY" />
      <xyz:recommendationLevel rdf:resource="http://purl.org/xyz/elements/1.0/#RecommendationLevel1" />
    </rdf:Description>

10. <rdf:Description rdf:about="urn-B">
      <xyz:isRecommendedFor rdf:resource=" http://info.uibk.ac.at/info/oecd-macroth/en/2478.html" />
```

```
    <xyz:isRecommendedBy rdf:resource="urn-PRPZ" />
    <xyz:recommendationLevel rdf:resource="http://purl.org/xyz/elements/1.0/#RecommendationLevel5" />
  </rdf:Description>

11. <rdf:Description rdf:about="urn-C">
    <xyz:isRecommendedFor rdf:resource=" http://info.uibk.ac.at/info/oecd-macroth/en/2478.html" />
    <xyz:isRecommendedBy rdf:resource="urn-PRPZ" />
    <xyz:recommendationLevel rdf:resource="http://purl.org/xyz/elements/1.0/#RecommendationLevel5" />
  </rdf:Description>

12. <rdf:Description rdf:about="urn-A">
    <xyz:isRecommendedFor rdf:resource=" http://info.uibk.ac.at/info/oecd-macroth/en/2478.html" />
    <xyz:isRecommendedBy rdf:resource="urn-PRPZ" />
    <xyz:recommendationLevel rdf:resource="http://purl.org/xyz/elements/1.0/#RecommendationLevel4" />
  </rdf:Description>

13. <rdf:Description rdf:about="urn-D">
    <xyz:isRecommendedFor rdf:resource=" http://info.uibk.ac.at/info/oecd-macroth/en/2478.html" />
    <xyz:isRecommendedBy rdf:resource="urn-PRPZ" />
    <xyz:recommendationLevel rdf:resource="http://purl.org/xyz/elements/1.0/#RecommendationLevel1" />
  </rdf:Description>

</rdf:RDF>
```

Table 25 - RDF Representation of Relationships Example

The above RDF representation has demonstrated how reputation system ratings and Identity meronymic relationships can be used together. Within the Morsoplexis Framework the ratings are imported into the Identity Recommendation component of the Identity Registry. The purpose of interfacing with recommendation systems is to filter and sort Information Fragments. Figure 94 outlines the process of filtering and sorting using the ratings imported as Identity Recommendations, the meronymic relationships between Identities and the related Qualifiers.

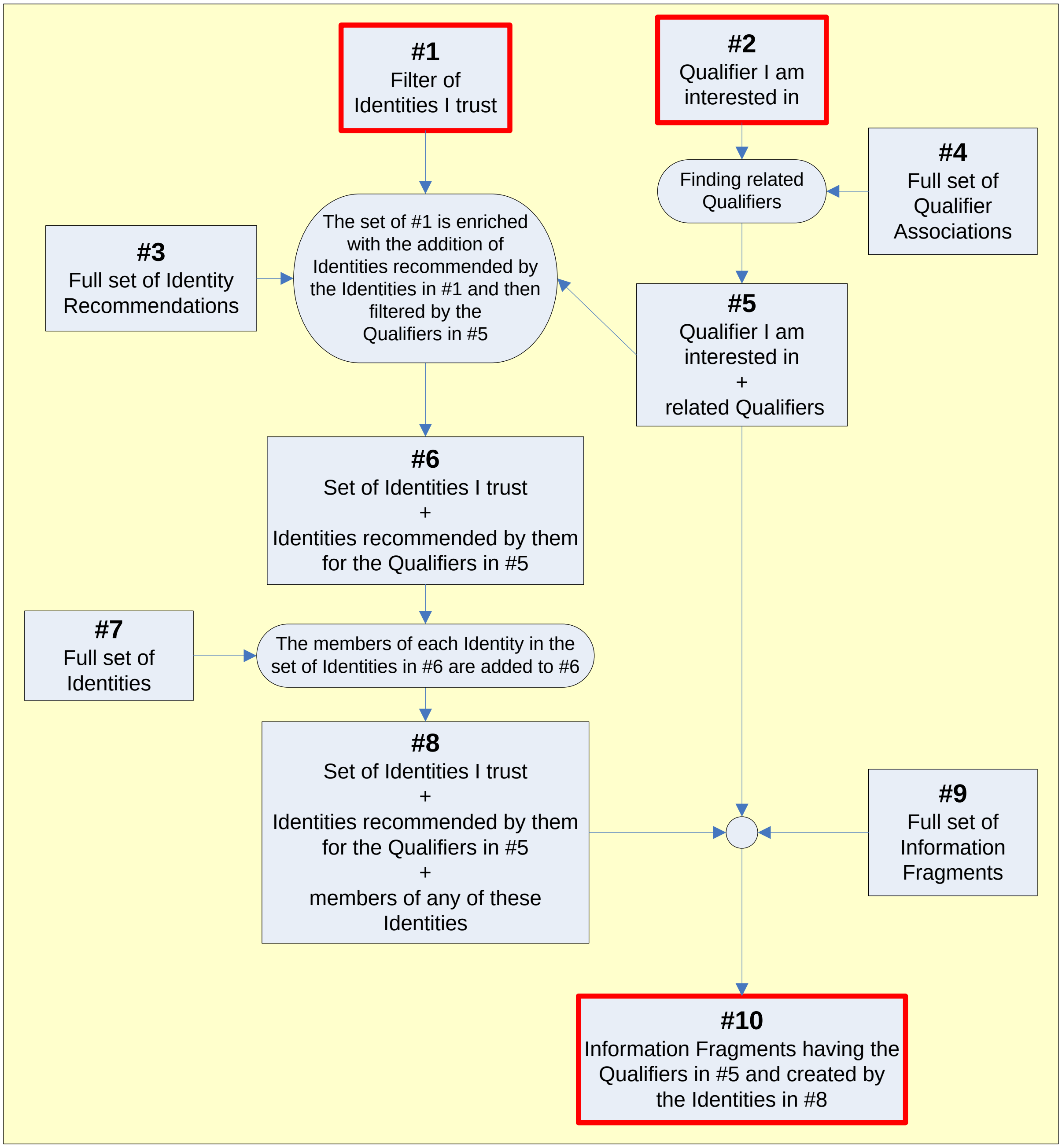


Figure 94 – Filtering and sorting using recommendations imported from reputation systems

The filtering process is the following

#1. The user specifies a Filter which includes the Identity whose recommendations s/he trusts

```
<Filter xmlns="http://morsoplexis.org/schema/Filter.xsd" FilterId="F:66" CommunityRegistryId="ABC1234"
 FilterType="IdentityFilter" >
  <Title>Identities whose recommendations I trust</Title>
  <OwnerIdentityId>ID:1</OwnerIdentityId>
```

```
 …
 <Identity IdentityId="urn:ADY" />  (Academic Department Y)
</Filter>
```


#2. The user specifies the Qualifier s/he is interested in searching for.

#5. Any Qualifiers related to the Qualifier the user is interested in are added to the set.

#6. The next step involves the retrieval of Identities recommended by the Identity the user has in the Filter and in turn the retrieval of Identities recommended by those recommended Identities. The set is then narrowed down by limiting it to the Qualifiers in the set of #5.


```
<IdentityRecommendation xmlns=”http://morsoplexis.org/schema/IdentityRecommendation.xsd">
  <RecommendedIdentityId>urn:DSE</RecommendedIdentityId>
  <RecommendedFor>Q:56</RecommendedFor>
  <RecommendingIdentityId>urn-ADY</RecommendingIdentityId>
  <RecommendationLevel>3</RecommendationLevel>
</IdentityRecommendation>

<IdentityRecommendation xmlns="http://morsoplexis.org/schema/IdentityRecommendation.xsd">
  <RecommendedIdentityId>urn:PRPZ</RecommendedIdentityId>
  <RecommendedFor>Q:56</RecommendedFor>
  <RecommendingIdentityId>urn-ADY</RecommendingIdentityId>
  <RecommendationLevel>1</RecommendationLevel>
</IdentityRecommendation>

<IdentityRecommendation xmlns="http://morsoplexis.org/schema/IdentityRecommendation.xsd">
  <RecommendedIdentityId>urn:B</RecommendedIdentityId>
  <RecommendedFor>Q:56</RecommendedFor>
  <RecommendingIdentityId> urn:PRPZ</RecommendingIdentityId>
  <RecommendationLevel>5</RecommendationLevel>
</IdentityRecommendation>

<IdentityRecommendation xmlns="http://morsoplexis.org/schema/IdentityRecommendation.xsd">
  <RecommendedIdentityId>urn:C</RecommendedIdentityId>
  <RecommendedFor>Q:56</RecommendedFor>
  <RecommendingIdentityId> urn:PRPZ</RecommendingIdentityId>
  <RecommendationLevel>5</RecommendationLevel>
</IdentityRecommendation>

<IdentityRecommendation xmlns="http://morsoplexis.org/schema/IdentityRecommendation.xsd">
  <RecommendedIdentityId>urn:A</RecommendedIdentityId>
  <RecommendedFor>Q:56</RecommendedFor>
  <RecommendingIdentityId> urn:PRPZ</RecommendingIdentityId>
  <RecommendationLevel>4</RecommendationLevel>
</IdentityRecommendation>

<IdentityRecommendation xmlns="http://morsoplexis.org/schema/IdentityRecommendation.xsd">
  <RecommendedIdentityId>urn:D</RecommendedIdentityId>
  <RecommendedFor>Q:56</RecommendedFor>
```

```xml
  <RecommendingIdentityId> urn:PRPZ</RecommendingIdentityId>
  <RecommendationLevel>1</RecommendationLevel>
</IdentityRecommendation>
```

#8. The set is augmented with any Identities specified as members of any of the Identities in the Identity set of #6.

```xml
<Identity xmlns=http://morsoplexis.org/schema/Identity.xsd IdentityId="ID:44" CommunityRegistryId="ABC1234">
  …
  <Members MemberIdentityId="ID:78"/>
  <Members MemberIdentityId="ID:89"/>
  <Members MemberIdentityId="ID:31"/>
</Identity>
```

#10. Finally, the entire set of Information Fragments is searched for Information Fragments created by any of the Identities in the Identity set of #8 and containing one of the Qualifiers in the Qualifier set of #5.

```xml
<Fragment FragmentId="F:69"  xmlns="http://morsoplexis.org/schema/Fragment.xsd">
  <Title>???</Title>
  <IdentityId>ID:89</IdentityId>
  <FragmentDescriptiveQualifiers FragmentDescriptiveQualifier="Q:67" />
  <FragmentDescriptiveQualifiers FragmentDescriptiveQualifier="Q:524" />
</Fragment>
```

The schema for Identity Recommendation is the one displayed in Table 26:

```xml
<?xml version="1.0" encoding="utf-8"?>
<xs:schema id="IdentityRecommendations"
        targetNamespace="http://morsoplexis.org/schema/IdentityRecommendation.xsd"
        xmlns="http://morsoplexis.org/schema/IdentityRecommendation.xsd"
        elementFormDefault="qualified"
        xmlns:xs="http://www.w3.org/2001/XMLSchema">

  <xs:simpleType name="mpKey">
    <xs:annotation>
      <xs:documentation xml:lang="en">
        The Format of the Global Identifier used by most Morsoplexis items
      </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:pattern value="[a-fA-F0-9]{8}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{12}"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:simpleType name="RecommendationLevels">
```

```
    <xs:annotation>
      <xs:documentation xml:lang="en">
        The Morsoplexis Associable Item Types
      </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:enumeration value="1"/>
      <xs:enumeration value="2"/>
      <xs:enumeration value="3"/>
      <xs:enumeration value="4"/>
      <xs:enumeration value="5"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:element name="IdentityRecommendation">
    <xs:complexType>
      <xs:sequence>
      <xs:element name="RecommendedIdentityId" type="mpKey" nillable="false" />
      <xs:element name="RecommendedFor" type="mpKey" nillable="false" />
      <xs:element name="RecommendingIdentityId" type="mpKey" nillable="false" />
      <xs:element name="RecommendationLevel" type="RecommendationLevels" nillable="false" />
      </xs:sequence>
    </xs:complexType>
  </xs:element>

</xs:schema>
```

Table 26 - Identity Recommendation Schema

Finally, it is fitting to add a note regarding the trustworthiness of the Community Registries themselves. Given that Community Registries are envisioned to require a logon to update, their maintenance has to be controlled by reputable organizations. Just like the envisioned Identities have varying degrees of trustworthiness, Community Registries also most likely end up being ranked based on their commitment to data integrity and service availability.

0.6.2 Information Fragment Registry

This is the core Registry of the Morsoplexis Framework. Its function is to record the Fragment Association information. This information consists of a representation of the boundaries of each one of the two fragments being associated. The nature of this representation varies depending on the format of the two entities bearing the two fragments being associated. The Morsoplexis

Framework should allow for the association of fragments of different formats. For example, a section of an image of a reproduction of a Renaissance painting with a mythological motif can be associated with a few lines from Ovid's Metamorphoses. For the purposes of demonstrating the functionality of this Framework, however, we will concentrate only on textual fragments residing in web pages. The representation of the boundaries will consist of the respective offsets of the beginning and ending of the fragment.

A typical Information Fragment Registry entry has approximately the following structure:

- Information Fragment
  - Fragment Description
    - Creation and update dates/times – the date and time of creation and updates of the Information Fragment
    - Optional Title and Abstract for the Fragment
    - Designation as Public or Private – denoting whether the Information Fragment is for the viewing of the creator only or shared with the rest of the world
    - Designation as Retracted or not – denoting whether this Information Fragment has been retracted by its creator. This option is to be used in the cases in which the Information Fragment has already been associated by another creator to another Information Fragment, Information fragment Association or Information Fragment Collection
    - The unique identifier of the Identity of the creator of the Fragment
    - The unique identifier of the Information Entity containing the Fragment
    - Descriptive Qualifiers
  - Fragment Definition
    - Boundaries of the two Fragments – designations of boundaries for each Information Fragment; different techniques are employed for

each format and consequently different schemas with varying data structures need to be defined for each format

- Information Fragment Association
    - Association Description
        - Association Identifier – a unique identifier for the described Information Fragment Association
        - Creation and update dates/times – the date and time of creation and updates of the Information Fragment Association
        - An Abstract describing the Association
        - Designation as Public or Private – denoting whether the Information Fragment Association is for the viewing of the creator only or shared with the rest of the world
        - Designation as Retracted or not – denoting whether this Information Fragment Association has been retracted by its creator. This option is to be used in the cases in which the Information Fragment Association has already been associated by another creator to another Information Fragment, Information fragment Association or Information Fragment Collection
        - Creator Identity Identifier – the unique identifier of the creator of the Information Fragment Association
        - Relationship Qualifiers – qualifiers describing the relationship between the two associated fragments
        - Descriptive Qualifiers – qualifiers describing the content of the association
    - Association Definition
        - Identifiers for the two associated Information Fragments
        - An Association is defined by providing a pair of any of the following:
            - Fragment
            - Fragment Association
            - Fragment Collection

- Information Fragment Collection
    - Collection Description
        - Collection Identifier – a unique identifier for the described Information Fragment Collection
        - Creation and update dates/times – the date and time of creation and updates of the Information Fragment Collection
        - Designation as Public or Private – denoting whether the Information Fragment Collection is for the viewing of the creator only or shared with the rest of the world
        - Designation as Retracted or not – denoting whether this Information Fragment Collection has been retracted by its creator. This option is to be used in the cases in which the Information Fragment Collection has already been associated by another creator to another Information Fragment, Information fragment Association or Information Fragment Collection
        - Creator Identity Identifier – the unique identifier of the creator of the Information Fragment Collection
        - Descriptive Qualifiers – qualifiers describing the content of the collection
    - Collected Fragments – a series of identifiers of Information Fragments collected together by a single user into a collection with a unique identifier
- Moved Item
    - Item Id – The id of the item regardless of its type
    - Item Type – The type of the item. Specified whether this item is an Information Fragment, an Information Entity, etc.
    - Identity Id – The identifier of the Identity of the owner of these items
    - New Community Registry Id – The identifier of the Community Registry to which these items are being moved

The following tables present the schemata for the Information Fragment Registry. Each one of the three schemata constituting the Information Fragment Registry contains a description section followed by a definition section and finally a qualifier section. The description section contains descriptive information about the Fragment, Association or Collection, such as title update dates etc. The definition section contains the information controlling the functionality this item provides. The qualifier section provides formal descriptors.

These schemata are not comprehensive, but they are designed to demonstrate adequately the capabilities of the conceived Framework. They contain only the bare essentials but they have been designed with expandability in mind.

0.6.2.1 Information Fragment Schema

The role of the Information Fragment Schema is to define a data model providing a unique identifier for each Information Fragment and to record the technical parameters defining its boundaries. This information is store so that it may be subsequently used to recreate the fragment for viewing as it was originally defined. The design has taken care of providing the flexibility of building new definitions of any possible structure. This is in recognition of the fact that some content formats require much more complex definitions than those needed by text and images. This schema, presented in Table 27, can be extended with any number of format definitions, such as video, audio maps etc.

```xml
<?xml version="1.0" encoding="utf-8"?>
<xs:schema id="Fragments"
        targetNamespace="http://morsoplexis.org/schema/Fragment.xsd"
        elementFormDefault="qualified"
        xmlns="http://morsoplexis.org/schema/Fragment.xsd"
        xmlns:xs="http://www.w3.org/2001/XMLSchema">

  <xs:simpleType name="mpKey">
    <xs:annotation>
```

```
    <xs:documentation xml:lang="en">
     The Format of the Global Identifier used by most Morsoplexis items
    </xs:documentation>
   </xs:annotation>
   <xs:restriction base="xs:string">
    <xs:pattern value="[a-fA-F0-9]{8}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{12}"/>
   </xs:restriction>
  </xs:simpleType>

  <xs:simpleType name="mpCRKey">
   <xs:annotation>
    <xs:documentation xml:lang="en">
     The Format of the Community Registry Identifier
    </xs:documentation>
   </xs:annotation>
   <xs:restriction base="xs:string">
    <xs:pattern value="[a-zA-Z]{3}[0-9]{4}"/>
   </xs:restriction>
  </xs:simpleType>

  <xs:group name="TextFragment">
   <xs:annotation>
    <xs:documentation xml:lang="en">
     Definition for a Text Fragment
    </xs:documentation>
   </xs:annotation>
   <xs:sequence>
    <xs:element name="FragmentStart" type="xs:int" nillable="false" default="0"/>
    <xs:element name="FragmentEnd" type="xs:int" nillable="false" default="0" />
   </xs:sequence>
  </xs:group>

  <xs:group name="ImageFragment">
   <xs:annotation>
    <xs:documentation xml:lang="en">
     Definition for an Image Fragment
    </xs:documentation>
   </xs:annotation>
   <xs:sequence>
    <xs:element name="FragmentTop" type="xs:decimal" nillable="false" default="0" />
    <xs:element name="FragmentLeft" type="xs:decimal" nillable="false" default="0" />
    <xs:element name="FragmentWidth" type="xs:decimal" nillable="false" default="0" />
    <xs:element name="FragmentHeight" type="xs:decimal" nillable="false" default="0" />
   </xs:sequence>
  </xs:group>


  <xs:element name="Fragment">
   <xs:complexType>
    <xs:sequence>

     <!-- DESCRIPTION -->
     <xs:element name="CreationDateTime" type="xs:dateTime" nillable="false" />
     <xs:element name="UpdateDateTime" type="xs:dateTime" nillable="false" />
     <xs:element name="Title" type="xs:string" default="" />
     <xs:element name="Description" type="xs:string" default="" />
     <xs:element name="Private" type="xs:boolean" default="0" />
     <xs:element name="Retracted" type="xs:boolean" default="0" />
     <xs:element name="IdentityId" type="mpKey" nillable="false" />
     <xs:element name="EntityId" type="mpKey" nillable="false" />

     <!-- DEFINITION -->
```

```
      <!-- Choice of various Fragment formats or another Association-->
      <xs:choice>
        <xs:element name="TextFragment">
          <xs:complexType>
            <xs:group id="TextDefinition" ref="TextFragment" />
          </xs:complexType>
        </xs:element>
        <xs:element name="ImageFragment">
          <xs:complexType>
            <xs:group id="ImageDefinition" ref="ImageFragment" />
          </xs:complexType>
        </xs:element>
      </xs:choice>

      <!-- QUALIFIERS -->
      <!-- Allows the entry of Qualifiers describing the content of the Fragment -->
      <xs:element name="FragmentDescriptiveQualifiers" maxOccurs="unbounded">
        <xs:complexType>
          <xs:attribute name="FragmentDescriptiveQualifier" type="mpKey">
          </xs:attribute>
        </xs:complexType>
      </xs:element>

    </xs:sequence>
    <xs:attribute name="FragmentId" type="mpKey" use="required" />
    <xs:attribute name="CommunityRegistryId" type="mpCRKey" use="required" />
  </xs:complexType>
 </xs:element>
</xs:schema>
```

Table 27 – Information Fragment Schema

0.6.2.2 Information Fragment Association Schema

The role of the Information Fragment Association Schema is to define a data model providing a unique identifier for each Information Fragment Association and the means of recording the two items being associated. An Information Fragment Association in essence defines a relationship between two objects each represented by a unique identifier and qualifies this relationship with descriptive and relationship qualifiers. As its name suggests, the two objects being associated are primarily two Information Fragments. However, an important aspect of the Morsoplexis Framework is that it allows for "Complex" Associations. A Complex Association is one involving the designation another Information Fragment Association or an Information Fragment

Collection as one of the two associated objects. This means that a user is able to identify a Fragment Association of interest and link it to a content Fragment. This allows for the establishment of multi-dimensional relationships between Information Fragments, Information Fragment Associations and Information Fragment Collections. An Information Fragment Association can serve as a launching point for comments or the juxtaposition of a variety of other related fragments. The schema, presented in Table 28, accomplishes this functionality by requiring two instances of the Association Definition element while providing a choice between Information Fragment, Information Fragment Association or Information Fragment Collection as the content of this definition.

```
<?xml version="1.0" encoding="utf-8" ?>
<xs:schema id="FragmentAssociations"
        targetNamespace="http://morsoplexis.org/schema/FragmentAssociation.xsd"
        elementFormDefault="qualified"
        xmlns="http://morsoplexis.org/schema/FragmentAssociation.xsd"
        xmlns:xs="http://www.w3.org/2001/XMLSchema">

  <xs:simpleType name="mpKey">
    <xs:annotation>
      <xs:documentation xml:lang="en">
        The Format of the Global Identifier used by most Morsoplexis items
      </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:pattern value="[a-fA-F0-9]{8}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{12}"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:simpleType name="mpCRKey">
    <xs:annotation>
      <xs:documentation xml:lang="en">
        The Format of the Community Registry Identifier
      </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:pattern value="[a-zA-Z]{3}[0-9]{4}"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:simpleType name="AssociableItemTypes">
    <xs:annotation>
      <xs:documentation xml:lang="en">
        The Morsoplexis Associable Item Types
      </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:enumeration value="Fragment"/>
```

```xml
      <xs:enumeration value="FragmentAssociation"/>
      <xs:enumeration value="FragmentCollection"/>
    </xs:restriction>
  </xs:simpleType>


  <xs:element name="FragmentAssociation">
    <xs:complexType>
      <xs:sequence>

        <xs:element name="CreationDateTime" type="xs:dateTime" nillable="false" />
        <xs:element name="UpdateDateTime" type="xs:dateTime" nillable="false" />
        <xs:element name="Title" type="xs:string" nillable="true" />
        <xs:element name="Description" type="xs:string" nillable="true" />
        <xs:element name="Private" type="xs:boolean" default="false" />
        <xs:element name="Retracted" type="xs:boolean" default="false" />
        <xs:element name="IdentityId" type="mpKey" nillable="false" />

        <!-- DEFINITION -->
        <!-- Two Items must be defined for each Association -->
        <xs:element name="LinkedItem1" minOccurs="1" maxOccurs="1">
          <xs:complexType>
            <xs:attribute name="LinkedItemId" type="mpKey" />
            <xs:attribute name="LinkedItemCommunityRegistryId" type="mpKey" />
            <xs:attribute name="LinkedItemType" type="AssociableItemTypes" />
          </xs:complexType>
        </xs:element>
        <xs:element name="LinkedItem2" minOccurs="1" maxOccurs="1">
          <xs:complexType>
            <xs:attribute name="LinkedItemId" type="mpKey" />
            <xs:attribute name="LinkedItemCommunityRegistryId" type="mpKey" />
            <xs:attribute name="LinkedItemType" type="AssociableItemTypes" />
          </xs:complexType>
        </xs:element>

        <!-- QUALIFIERS -->
        <!-- Allows the entry of Qualifiers describing the relationship between the two Fragments -->
        <xs:element name="RelationshipQualifiers" maxOccurs="unbounded">
          <xs:complexType>
            <xs:attribute name="RelationshipQualifier" type="mpKey">
            </xs:attribute>
          </xs:complexType>
        </xs:element>
        <!-- Allows the entry of Qualifiers describing the content of the Assiciation -->
        <xs:element name="AssociationDescriptiveQualifiers" maxOccurs="unbounded">
          <xs:complexType>
            <xs:attribute name="AssociationDescriptiveQualifier" type="mpKey">
            </xs:attribute>
          </xs:complexType>
        </xs:element>

      </xs:sequence>
      <xs:attribute name="FragmentAssociationId" type="mpKey" use="required" />
      <xs:attribute name="CommunityRegistryId" type="mpCRKey" use="required" />
    </xs:complexType>
  </xs:element>
</xs:schema>
```

Table 28 - Information Fragment Association Schema

0.6.2.3 Information Fragment Collection Schema

The role of the Information Fragment Collection Schema, presented in Table 29, is to define a data model providing a unique identifier for each Information Fragment Collection and the means of recording the identifiers of the Information Fragments being collected. Its structure is similar to that of the Information Fragment schema with the difference that the definition contains a list of the identifiers of the fragments belonging to this collection.

```
<?xml version="1.0" encoding="utf-8" ?>
<xs:schema id="FragmentCollections"
          targetNamespace="http://morsoplexis.org/schema/FragmentCollection.xsd"
          elementFormDefault="qualified"
          xmlns="http://morsoplexis.org/schema/FragmentCollection.xsd"
          xmlns:xs="http://www.w3.org/2001/XMLSchema">

  <xs:simpleType name="mpKey">
    <xs:annotation>
      <xs:documentation xml:lang="en">
        The Format of the Global Identifier used by most Morsoplexis items
      </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:pattern value="[a-fA-F0-9]{8}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{12}"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:simpleType name="mpCRKey">
    <xs:annotation>
      <xs:documentation xml:lang="en">
        The Format of the Community Registry Identifier
      </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:pattern value="[a-zA-Z]{3}[0-9]{4}"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:element name="FragmentCollection">
    <xs:complexType>
      <xs:sequence>

        <xs:element name="CollectionDescription" maxOccurs="1" minOccurs="1">
          <xs:complexType>
            <xs:sequence>
              <xs:element name="CreationDateTime" type="xs:dateTime" nillable="false" />
              <xs:element name="UpdateDateTime" type="xs:dateTime" nillable="false" />
              <xs:element name="Abstract" type="xs:string" nillable="true" />
              <xs:element name="Private" type="xs:boolean" default="0" />
              <xs:element name="Retracted" type="xs:boolean" default="0" />
              <xs:element name="IdentityId" type="mpKey" nillable="false" />
            </xs:sequence>
```

```
      </xs:complexType>
     </xs:element>

     <!-- DEFINITION -->
     <!-- Several Fragments can be defined in each Collection -->
     <xs:sequence maxOccurs="unbounded">
      <xs:element name="CollectedFragment">
       <xs:complexType>
        <xs:attribute name="CollectedFragmentId" type="mpKey" />
       </xs:complexType>
      </xs:element>
     </xs:sequence>

     <!-- QUALIFIERS -->
     <!-- Allows the entry of Qualifiers describing the content of the Collection -->
     <xs:element name="CollectionDescriptiveQualifiers" maxOccurs="unbounded">
      <xs:complexType>
       <xs:attribute name="CollectionDescriptiveQualifier" type="mpKey" />
      </xs:complexType>
     </xs:element>

    </xs:sequence>
    <xs:attribute name="CollectionId" type="mpKey" use="required" />
    <xs:attribute name="CommunityRegistryId" type="mpCRKey" use="required" />
   </xs:complexType>
  </xs:element>


</xs:schema>
```

Table 29 - Information Fragment Collection Schema

0.6.2.4 Moved Item Registry Schema

The Moved Item Registry records items moved to another Community Registry. The Morsoplexis Framework provides the capability of moving all items created by a specific Identity from one Community Registry to another. The entries in this registry ensure that a permanent redirection to the new Community Registry is in place. Each item's Identifier and entire contents remain the same while the Community Registry Id changes.

```
<?xml version="1.0" encoding="utf-8" ?>
<xs:schema id="MovedItems"
         targetNamespace="http://morsoplexis.org/MovedItems.xsd"
         elementFormDefault="qualified"
         xmlns="http://morsoplexis.org/MovedItems.xsd"
         xmlns:xs="http://www.w3.org/2001/XMLSchema">

 <xs:simpleType name="mpKey">
```

```xml
    <xs:annotation>
      <xs:documentation xml:lang="en">
        The Format of the Global Identifier used by most Morsoplexis items
      </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:pattern value="[a-fA-F0-9]{8}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{12}"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:simpleType name="mpCRKey">
    <xs:annotation>
      <xs:documentation xml:lang="en">
        The Format of the Community Registry Identifier
      </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:pattern value="[a-zA-Z]{3}[0-9]{4}"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:simpleType name="ItemTypes">
    <xs:annotation>
      <xs:documentation xml:lang="en">
        The Morsoplexis Item Types
      </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:enumeration value="Entity"/>
      <xs:enumeration value="Filter"/>
      <xs:enumeration value="Fragment"/>
      <xs:enumeration value="FragmentAssociation"/>
      <xs:enumeration value="FragmentCollection"/>
      <xs:enumeration value="Qualifier"/>
      <xs:enumeration value="QualifierAssociation"/>
      <xs:enumeration value="QualifierCollection"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:element name="MovedItem">
    <xs:complexType>
      <xs:attribute name="ItemId" type="mpKey"></xs:attribute>
      <xs:attribute name="ItemType" type="ItemTypes"></xs:attribute>
      <xs:attribute name="IdentityId" type="mpKey"></xs:attribute>
      <xs:attribute name="NewCommunityRegistryId" type="mpCRKey"></xs:attribute>
    </xs:complexType>
  </xs:element>

</xs:schema>
```

0.6.3 Information Entity Caching Component

The function of this component is one of capturing Information Entities in the state in which they are at the moment of the creation of the Fragment Association and storing this captured copy to be used whenever that given Fragment Association is used, ensuring persistence of this reference to the Fragment Association within the Web environment which is notorious for its lack of persistence.

Caching implementations may vary and they may utilize any technology available meeting the needs of the Morsoplexis Framework. An implementation of the Morsoplexis Framework needs to ensure that caching has the following functionality:

- Maintaining the same content – in the case of a web page this means the same textual and graphical content
- Maintaining the same format – in the case of a web page this means capturing any style sheets used by this web page
- Maintaining the same dynamic functionality to a degree that is reasonable – in the case of a web page this means ECMAscript content. Some server-based functionality may not be cacheable

The Morsoplexis Framework is being conceived of and designed with the understanding that technical constraints may prevent the adequate caching of every possibly available resource. If the Morsoplexis Framework proves to be of value, technical workarounds may be devised to achieve optimal caching results. The technology which will be used to provide some demonstration of this capability is "MIME Encapsulation of Aggregate Documents, such as HTML" IETF specification[44].

[44] http://www.apps.ietf.org/rfc/rfc2557.html

There are two issues with which the design of the Information Entity Caching Component has been concerned. The one is the issue of versioning and the other is the issue of caching per se. The consideration of these issues has been crucial in the selection of the appropriate mechanism.

0.6.3.1 Versioning

It was deemed essential that the versioning issue be addressed first. A proper design aiming at persistency has to adopt an appropriate versioning approach. As we refer to an Information Entity in this framework, it is important to discuss how it fits within the realm of existing version models. Our examination needed to answer two questions:

- what the version model to be employed by this framework is
- what role an Information Entity plays within this model.

The two primary types of versioning are extensional versioning (often called state-based versioning) and intensional versioning (often called task-based or change-based versioning) [30]. The versioning issue has been discussed more thoroughly in the Literature Review. For the purpose of our discussion here, we will use the terms state-based and task-based.

In spite of the benefits of task-based versioning, when it comes to the required characteristics of the Information Entity suitable for the Morsoplexis Framework, state-based versioning is a better fit. An Information Entity has to be a single version (revision) of an item.

If the vision outlined here is to be realized, there are two inviolable requirements for links between Information Fragments:

- The links should be adequately reliable (provided that the technical infrastructure is operational)
- A reference to an Information Fragment within Information Entity A created by the author of Information Entity B should be guaranteed to always bring up the

very same version of Information Entity A that the author of Information Entity B had consulted at the time of the creation of the link. The ability to carry the reference over to later versions is highly desirable but it cannot be a requirement because it the imprecise nature of such an operation.

If these references have to be absolutely reliable, then the Information Entities being interlinked on both sides need to be frozen. If version 2 of a document A has a reference to version 1 of document B, then we need to make sure that version 1 of document B remains unaltered even if document B has subsequently evolved into newer versions. If the Information Entities at the two ends of a link are frozen but additional attributes can be attached to the link between them by the creator of the link, then the integrity of the relationship has been preserved. This arrangement renders unnecessary the definition of special links such as what Østerbye [104] calls "substance links" and Vitali [33] calls substantial links differentiating them from "annotation" links. Drawing a distinction between links and arbitrarily determining which links are guaranteed and which are not

Using a state-based model as part of the description of the essence of an Information Entity does not obviate the employment of a change-based mechanism. A change-based mechanism can be a useful complement, but not a requirement.

0.6.3.2 Caching

With the versioning approach having been selected, we can proceed with the consideration of caching options. The proposal of adopting a state-based versioning model brings another question. How does one ensure that a version of the information resource being referenced is the same? In an open web environment this cannot be guaranteed unless references are made to a very limited number of sites known to preserve older versions. The only practical albeit somewhat costly solutions are:

- using a global archive such as the Internet Archive [105]
- using a mechanism generating a registered cached copy of the version of the information entity being referenced at the time of the reference creation

Caching was one of the features envisioned by Ted Nelson for his Xanadu [106] system. More complex content storing solutions are the several hyperbase systems, which incorporate the linkbase model along with an elaborate mechanism for storing content. This storage subsystem is often called HyperBase Management System (HBMS). An example of a hyperbase system was HyperDisco. [38] In spite of the fact that some of these hyperbase systems were considered open systems, in their ability to incorporate additional tools and content, their goal was to provide a superior information storage system and not to integrate with a wide and dynamic environment such as the Web.

Lessons learned from linkbase and hyberbase systems can be used to conceive of a new model incorporating some of their most valuable features, but taking into consideration that in order for such a model to be successful it cannot diverge from the simplicity that the Web provides today

0.6.3.3 The Chosen Approach and Structure

The approach chosen involves the capture of the Information Entity in the state in which it is at the time of the Fragment Association. This solution calls for storing the Information Entity in a format preserving its full functionality. It also calls for storage space for these Information Entities as well as a registry keeping track of the stored information and controlling the unique identifier attached to them.

The structure of an entry in this registry is the following

- EntityId – a unique identifier for a cached Information Entity
- EntityURI – the URI of the resource cached

- CacheDateTime – the Date and time the caching took place
- IdentityId – the unique identifier of the Identity who created this cached Information Entity
- Unacheable – an administrative mark designating the failure of caching addressing imperfections in the caching implementation

The schema is presented in Table 30:

```xml
<?xml version="1.0" encoding="utf-8"?>
<xs:schema id="Entities"
        targetNamespace="http://morsoplexis.org/schema/Entity.xsd"
        elementFormDefault="qualified"
        xmlns="http://morsoplexis.org/schema/Entity.xsd"
        xmlns:xs="http://www.w3.org/2001/XMLSchema">

  <xs:simpleType name="mpKey">
    <xs:annotation>
      <xs:documentation xml:lang="en">
        The Format of the Global Identifier used by most Morsoplexis items
      </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:pattern value="[a-fA-F0-9]{8}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{12}"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:simpleType name="mpCRKey">
    <xs:annotation>
      <xs:documentation xml:lang="en">
        The Format of the Community Registry Identifier
      </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:pattern value="[a-zA-Z]{3}[0-9]{4}"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:element name="Entity">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="EntityURI" type="xs:anyURI" />
        <xs:element name="Title" type="xs:string" />
        <xs:element name="CacheDateTime" type="xs:dateTime" />
        <xs:element name="IdentityId" type="mpKey" />
        <xs:element name="Uncacheable" type="xs:boolean" default="false" />
      </xs:sequence>
      <xs:attribute name="EntityId" type="mpKey" use="required" />
      <xs:attribute name="CommunityRegistryId" type="mpCRKey" use="required" />
    </xs:complexType>
  </xs:element>

</xs:schema>
```

Table 30 - Information Entity Caching Schema

0.6.4 Local Registry of Registries

The Local Registry of Registries is a copy of the Distributed Registry of Registries replicated for performance and resolution reliability.

0.6.5 Local Qualifier Registry

The Local Qualifier Registry contains the lists of terms used to describe the relationships between Information Fragment Associations. Its structure is identical to that of the Distributed Qualifier Registry described in section 0.5.2.Its two components are:

- Qualifier Collections – named collections of qualifiers used to present users with their lists of qualifiers for assignment
- Qualifiers – a list of all qualifiers each one along with the Qualifier Collection to which it belongs

The Local Qualifier Registry synchronizes its Qualifier Collections and Items with the Distributed Qualifier Registry. Each Qualifier Collection bears the unique identifier of the Community Registry in which it was created. It also bears a sharing designation. By being designated as "CommunityRegistryOnly", a Qualifier Collection is not synchronized with the Distributed Qualifier Registry. As mentioned in section 0.5.2 above, some mainstream Thesauri and Taxonomies will be part of the Distributed Qualifier Registry and they will be available to all Local Qualifier Registries. The schema of the Local Qualifier Registry is the one shown in Table 31.

```xml
<?xml version="1.0" encoding="utf-8" ?>
<xs:schema id="Qualifiers"
        targetNamespace="http://morsoplexis.org/schema/Qualifier.xsd"
        elementFormDefault="qualified"
        xmlns="http://morsoplexis.org/schema/Qualifier.xsd"
        xmlns:xs="http://www.w3.org/2001/XMLSchema">
```

```
  <xs:simpleType name="mpKey">
   <xs:annotation>
    <xs:documentation xml:lang="en">
     The Format of the Global Identifier used by most Morsoplexis items
    </xs:documentation>
   </xs:annotation>
   <xs:restriction base="xs:string">
    <xs:pattern value="[a-fA-F0-9]{8}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{12}"/>
   </xs:restriction>
  </xs:simpleType>

  <xs:simpleType name="mpCRKey">
   <xs:annotation>
    <xs:documentation xml:lang="en">
     The Format of the Community Registry Identifier
    </xs:documentation>
   </xs:annotation>
   <xs:restriction base="xs:string">
    <xs:pattern value="[a-zA-Z]{3}[0-9]{4}"/>
   </xs:restriction>
  </xs:simpleType>

  <xs:element name="Qualifier">
   <xs:complexType>
    <xs:sequence>
     <xs:element name="CreationDateTime" type="xs:dateTime" />
     <xs:element name="UpdateDateTime" type="xs:dateTime" />
     <xs:element name="QualifierName" type="xs:string" nillable="false" />
     <xs:element name="QualifierDescription" type="xs:string" default="" />

     <xs:element name="OtherIdentifier" type="xs:string" default="">
      <xs:annotation>
       <xs:documentation xml:lang="en">
        OtherIdentifier is used for miscelleneous identifiers mostly in the case of importing external Thesauri and
Taxonomies
       </xs:documentation>
      </xs:annotation>
     </xs:element>

     <xs:element name="QualifierCollectionId" type="mpKey" nillable="false">
      <xs:annotation>
       <xs:documentation xml:lang="en">
        The Qualifier Collection to which this Qualifier belongs
       </xs:documentation>
      </xs:annotation>
     </xs:element>

     <xs:element name="IdentityId" type="mpKey" />

    </xs:sequence>
    <xs:attribute name="QualifierId" type="mpKey" use="required" />
    <xs:attribute name="CommunityRegistryId" type="mpCRKey" use="required" />
   </xs:complexType>
  </xs:element>

</xs:schema>
```

Table 31 - Qualifier Schema

The schema for the Qualifier Collection is shown in Table 32.

```
<?xml version="1.0" encoding="utf-8" ?>
<xs:schema id="QualifierCollections"
           targetNamespace="http://morsoplexis.org/schema/QualifierCollection.xsd"
           elementFormDefault="qualified"
           xmlns="http://morsoplexis.org/schema/QualifierCollection.xsd"
           xmlns:xs="http://www.w3.org/2001/XMLSchema">

 <xs:simpleType name="mpKey">
  <xs:annotation>
   <xs:documentation xml:lang="en">
    The Format of the Global Identifier used by most Morsoplexis items
   </xs:documentation>
  </xs:annotation>
  <xs:restriction base="xs:string">
   <xs:pattern value="[a-fA-F0-9]{8}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{12}"/>
  </xs:restriction>
 </xs:simpleType>

 <xs:simpleType name="mpCRKey">
  <xs:annotation>
   <xs:documentation xml:lang="en">
    The Format of the Community Registry Identifier
   </xs:documentation>
  </xs:annotation>
  <xs:restriction base="xs:string">
   <xs:pattern value="[a-zA-Z]{3}[0-9]{4}"/>
  </xs:restriction>
 </xs:simpleType>

 <xs:simpleType name="QualifierCollectionTypes">
  <xs:annotation>
   <xs:documentation xml:lang="en">
    The two types of Qualifier Collections
   </xs:documentation>
  </xs:annotation>
  <xs:restriction base="xs:string">
   <xs:enumeration value="Relationship"/>
   <xs:enumeration value="Descriptive"/>
  </xs:restriction>
 </xs:simpleType>

 <xs:simpleType name="SharingTypes">
  <xs:annotation>
   <xs:documentation xml:lang="en">
    Specifying the extent to which this Qualifier Collection is to be shared
   </xs:documentation>
  </xs:annotation>
  <xs:restriction base="xs:string">
   <xs:enumeration value="Private">
    <xs:annotation>
     <xs:documentation xml:lang="en">
      Private - For use by the owner Identity only
     </xs:documentation>
    </xs:annotation>
   </xs:enumeration>
   <xs:enumeration value="Limited">
    <xs:annotation>
     <xs:documentation xml:lang="en">
```

```
          Limited to a certain group of Identities
        </xs:documentation>
      </xs:annotation>
    </xs:enumeration>
    <xs:enumeration value="CommunityRegistryOnly">
      <xs:annotation>
        <xs:documentation xml:lang="en">
          Public within this Community Registry Only
        </xs:documentation>
      </xs:annotation>
    </xs:enumeration>
    <xs:enumeration value="Global">
      <xs:annotation>
        <xs:documentation xml:lang="en">
          Shared with other Community Registries
        </xs:documentation>
      </xs:annotation>
    </xs:enumeration>
  </xs:restriction>
 </xs:simpleType>


 <xs:element name="QualifierCollection">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="CreationDateTime" type="xs:dateTime" nillable="false" />
      <xs:element name="UpdateDateTime" type="xs:dateTime" nillable="false" />
      <xs:element name="CommunityRegistryId" type="mpKey" nillable="false" />
      <xs:element name="QualifierCollectionName" type="xs:string" nillable="false" />
      <xs:element name="QualifierCollectionDescription" type="xs:string" default="" />
      <xs:element name="QualifierCollectionType" type="QualifierCollectionTypes" nillable="false" />
      <xs:element name="Sharing" type="SharingTypes" default="CommunityRegistryOnly" />
      <xs:element name="IdentityId" type="mpKey" nillable="false" />

      <xs:element name="UsedBy" minOccurs="1" maxOccurs="unbounded">
        <xs:annotation>
          <xs:documentation xml:lang="en">
            Keeping track of the Community Registries using this Qualifier Collection
          </xs:documentation>
        </xs:annotation>
        <xs:complexType>
          <xs:attribute name="CommunityRegistryId" type="mpCRKey" use="required" />
        </xs:complexType>
      </xs:element>

    </xs:sequence>
    <xs:attribute name="QualifierCollectionId" type="mpKey" use="required" />
    <xs:attribute name="CommunityRegistryId" type="mpCRKey" use="required" />
  </xs:complexType>
 </xs:element>
</xs:schema>
```

Table 32 - Qualifier Collection Schema

In order to simulate to some basic degree the functionality of traditional Thesauri and Taxonomies, Descriptive Qualifiers will be able to be associated with other Descriptive

Qualifiers and their association will be qualified by appropriate Relationship Qualifiers such as "Narrower Term" and "Use for". The schema for these Qualifier Associations is presented in Table 33.

```xml
<?xml version="1.0" encoding="utf-8" ?>
<xs:schema id="QualifierAssociations"
        targetNamespace="http://morsoplexis.org/schema/QualifierAssociation.xsd"
        elementFormDefault="qualified"
        xmlns="http://morsoplexis.org/schema/QualifierAssociation.xsd"
        xmlns:xs="http://www.w3.org/2001/XMLSchema">

 <xs:simpleType name="mpKey">
  <xs:annotation>
   <xs:documentation xml:lang="en">
    The Format of the Global Identifier used by most Morsoplexis items
   </xs:documentation>
  </xs:annotation>
  <xs:restriction base="xs:string">
   <xs:pattern value="[a-fA-F0-9]{8}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{12}"/>
  </xs:restriction>
 </xs:simpleType>

 <xs:simpleType name="mpCRKey">
  <xs:annotation>
   <xs:documentation xml:lang="en">
    The Format of the Community Registry Identifier
   </xs:documentation>
  </xs:annotation>
  <xs:restriction base="xs:string">
   <xs:pattern value="[a-zA-Z]{3}[0-9]{4}"/>
  </xs:restriction>
 </xs:simpleType>

 <xs:element name="QualifierAssociation">
  <xs:complexType>
   <xs:sequence>

    <!-- DESCRIPTION -->
    <xs:element name="CreationDateTime" type="xs:dateTime" nillable="false" />
    <xs:element name="UpdateDateTime" type="xs:dateTime" nillable="false" />
    <xs:element name="Title" type="xs:string" nillable="true" />
    <xs:element name="Description" type="xs:string" nillable="true" />
    <xs:element name="Private" type="xs:boolean" default="false" />
    <xs:element name="Retracted" type="xs:boolean" default="false" />
    <xs:element name="IdentityId" type="mpKey" nillable="false" />

    <!-- DEFINITION -->
    <!-- Two Items must be defined for each Association -->
    <xs:element name="LinkedItem1" minOccurs="1" maxOccurs="1">
     <xs:complexType>
      <xs:attribute name="LinkedItemId" type="mpKey" />
     </xs:complexType>
    </xs:element>
    <xs:element name="LinkedItem2" minOccurs="1" maxOccurs="1">
```

```xml
        <xs:complexType>
          <xs:attribute name="LinkedItemId" type="mpKey" />
        </xs:complexType>
      </xs:element>

    </xs:sequence>
    <xs:attribute name="QualifierAssociationId" type="mpKey" use="required" />
    <xs:attribute name="CommunityRegistryId" type="mpCRKey" use="required" />
  </xs:complexType>
 </xs:element>
</xs:schema>
```

Table 33 - Qualifier Association

0.6.6 Local Extra-Community Link Registry

The Local Extra-Community Link Registry in conjunction with the Distributed Extra-Community Link Registry enables the coordination of the process of establishing Information Fragment Associations and Information Fragment Collections across Community Registries. If a user in Community Registry CR:A creates an Information Fragment Association IFA:a involving an Information Fragment IF:b in Community Registry CR:B, Community Registry CR:B will be apprised of this action through a concise entry in this registry. This entry will specify that IFA:a in CR:A has as one of its two linked items IF:b which resides in CR:B. This entry will originally be created in the Local Extra-Community Link Registry of CD:A, then copied over to the Distributed Extra-Community Link Registry and then in turn copied over to the Local Extra-Community Link Registry of CR:B.

This approach is taken in order to obviate the need of keeping a centralized resolver of every single information item. Using this registry, two Community Registries are able to exchange the lists of the items involved in these links.

The structure of an entry in this registry is the following

- Creation and Update Date and Time
- Linking Item

    - The Identifier of the Linking Item
    - The Identifier of the Community Registry to which the Linking Item belongs
    - The Type of the link (i.e. Information Fragment Association or Information Fragment Collection)
- Linked Item
    - The Identifier of the Linked Item
    - The Identifier of the Community Registry to which the Linked Item belongs
    - The Type of the link (i.e. Information Fragment, Information Fragment Association or Information Fragment Collection)
- A designation regarding the status of the Linking Item as deleted or not. If the Linking item is deleted, the Community Registry receiving the entry will ensure that the previously linked item is no longer appearing as linked-to.

```xml
<?xml version="1.0" encoding="utf-8"?>
<xs:schema id ="ExtraCommunityLinks"
        targetNamespace="http://morsoplexis.org/schema/ExtraCommunityLink.xsd"
        elementFormDefault="qualified"
        xmlns="http://morsoplexis.org/schema/ExtraCommunityLink.xsd"
        xmlns:xs="http://www.w3.org/2001/XMLSchema">


  <xs:simpleType name="mpKey">
    <xs:annotation>
      <xs:documentation xml:lang="en">
        The Format of the Global Identifier used by most Morsoplexis items
      </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:pattern value="[a-fA-F0-9]{8}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{12}"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:simpleType name="mpCRKey">
    <xs:annotation>
      <xs:documentation xml:lang="en">
        The Format of the Community Registry Identifier
      </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:pattern value="[a-zA-Z]{3}[0-9]{4}"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:simpleType name="LinkingItemTypes">
    <xs:annotation>
      <xs:documentation xml:lang="en">
        Linking Item Types
      </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
```

```xml
      <xs:enumeration value="FragmentAssociation"/>
      <xs:enumeration value="FragmentCollection"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:simpleType name="LinkedItemTypes">
    <xs:annotation>
      <xs:documentation xml:lang="en">
        Linked Item Types
      </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
      <xs:enumeration value="Fragment"/>
      <xs:enumeration value="FragmentAssociation"/>
      <xs:enumeration value="FragmentCollection"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:element name="ExtraCommunityLink">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="CreationDateTime" type="xs:dateTime" nillable="false" />
        <xs:element name="UpdateDateTime" type="xs:dateTime" nillable="false" />
        <xs:element name="LinkingItemId" type="mpKey" />
        <xs:element name="LinkingCommunityRegistryId" type="mpCRKey" />
        <xs:element name="LinkingType" type="LinkingItemTypes" />
        <xs:element name="LinkedItemId" type="mpKey" />
        <xs:element name="LinkedCommunityRegistryId" type="mpCRKey" />
        <xs:element name="LinkedType" type="LinkedItemTypes" />
        <xs:element name="Deleted" type="xs:boolean" />
      </xs:sequence>
    </xs:complexType>
  </xs:element>

</xs:schema>
```

Table 34 - Local Extra-Community Link Registry Schema

0.6.7 Filter Registry

The Filter Registry stores Filters created by Identities. These Filters provide the means by which the number of Fragment Associations displayed is restricted. Tools will be able to utilize this registry to allow focusing on specific subjects or on Fragment Associations created by a certain list of Identities. This capability is important in order to ensure that relevant information is being retrieved. It also helps improve the level of reliability of the information by restricting the display to Fragment Associations created by trustworthy Identities. The simple schema for this registry is as show in Table 35.

```xml
<?xml version="1.0" encoding="utf-8"?>
<xs:schema id="Filters"
        targetNamespace="http://morsoplexis.org/schema/Filter.xsd"
        elementFormDefault="qualified"
        xmlns="http://morsoplexis.org/schema/Filter.xsd"
        xmlns:xs="http://www.w3.org/2001/XMLSchema">

 <xs:simpleType name="mpKey">
  <xs:annotation>
   <xs:documentation xml:lang="en">
    The Format of the Global Identifier used by most Morsoplexis items
   </xs:documentation>
  </xs:annotation>
  <xs:restriction base="xs:string">
   <xs:pattern value="[a-fA-F0-9]{8}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{12}"/>
  </xs:restriction>
 </xs:simpleType>

 <xs:simpleType name="mpCRKey">
  <xs:annotation>
   <xs:documentation xml:lang="en">
    The Format of the Community Registry Identifier
   </xs:documentation>
  </xs:annotation>
  <xs:restriction base="xs:string">
   <xs:pattern value="[a-zA-Z]{3}[0-9]{4}"/>
  </xs:restriction>
 </xs:simpleType>

 <xs:simpleType name="FilterTypes">
  <xs:annotation>
   <xs:documentation xml:lang="en">
    The Morsoplexis Filter Types
   </xs:documentation>
  </xs:annotation>
  <xs:restriction base="xs:string">
   <xs:enumeration value="IdentityFilter"/>
   <xs:enumeration value="QualifierFilter"/>
  </xs:restriction>
 </xs:simpleType>

 <xs:element name="Filter">
  <xs:complexType>
   <xs:sequence>
    <xs:element name="Title" maxOccurs="1" />
    <xs:element name="OwnerIdentityId" />
    <xs:element name="Private" type="xs:boolean" default="0" />
    <xs:choice minOccurs="1" maxOccurs="1">
     <xs:sequence>
     <xs:element name="Identity" minOccurs="0" maxOccurs="unbounded">
      <xs:complexType>
        <xs:attribute name="IdentityId" type="mpKey" use="required" />
      </xs:complexType>
     </xs:element>
     </xs:sequence>
      <xs:sequence>
```

```xml
        <xs:element name="Qualifier" minOccurs="0" maxOccurs="unbounded">
          <xs:complexType>
            <xs:attribute name="QualifierId" type="mpKey" use="required" />
          </xs:complexType>
        </xs:element>
        </xs:sequence>
      </xs:choice>
    </xs:sequence>
    <xs:attribute name="FilterId" type="mpKey" use="required" />
    <xs:attribute name="CommunityRegistryId" type="mpCRKey" />
    <xs:attribute name="FilterType" type="FilterTypes" />
  </xs:complexType>
 </xs:element>

</xs:schema>
```

Table 35 - Filter Registry Schema

Users have the ability of designating a Filter as public or private, just as in the case of Information Fragments. A Filter is always editable only by the Identity specified as "Owner Identity". When designated as "private" it is viewable only to the Owner Identity. If designated as "public" the Filter is able to be used by any Identity in that Community Registry, but only the owner is able to make changes. Any existing Filter is derivable, which means it can be used as a template for the creation of a new Filter. This allows users to take advantage of existing Filters that others in the same discipline have created and build on them with changes to meet their own needs.

0.7 THE AGGREGATORS

Information Fragment Associations and cached Information Entities residing in a given Community Registry are globally accessible, provided that the Registry is open to the public and not restricted to its registered users. However, even in the cases in which this data is publicly exposed, Information Fragment Association Viewing Interfaces can be of more value if they could query Fragment Associations from multiple Community Registries. This could be accomplished through sequential querying of each Community Registry, but this would not be practical performance-wise. A reasonable solution is to point Information Fragment Association Viewing Interfaces to Aggregators. These Aggregators combine the public contents of the Information Fragment Registries into larger combined registries which are queried with a single efficient query and they still point to their respective Community Registries' Caches to retrieve the cached content. The aggregation design is similar to that of the Real Simple Syndication (RSS), the Open Archives Initiative (OAI), the Atom Syndication Format and other such schemes. The data accumulated by the Aggregators is read only and used only for dissemination. Aggregation is conceived of as being a simple process. Each Community Registry exposes the entries which have been added or updated since a given date and time, and the Aggregator adds or updates these entries to its unified registry. All entries in this unified registry, of course, also bear the unique identifier of the Community Registry to which they belong.

Although some Aggregators are expected to be comprehensive, most are more likely to be selective, performing aggregation based on some conditions defining specific registries and specific Identity restrictions. The Aggregators are tools useful for querying descriptive content of Information Fragments such as titles, descriptions and Descriptive Qualifiers. They may use information from the Distributed Registry of Registries in order to select Community Registries within a specific discipline. However, even the comprehensive Aggregators, i.e. those which aggregate the content of all Community Registries are only tools. They do not have the central authoritative function that the Registry of Registries has.

0.8 POSSIBLE APPLICATION EXAMPLES

In order to better illustrate the possible impact the proposed Framework may have on current practices, the following examples are being presented. These examples do not incorporate every feature of the Morsoplexis Framework, nor does the Morsoplexis Framework aspire to provide a definitive solution for the needs of these practices. However, the presentation of these examples demonstrates that the conceived capabilities can be invaluable for a variety of uses and for diverse communities.

0.8.1 Associating Thematic Variations

Literature abounds with instances of thematic variations appearing scattered within the content of different works. Whether an author is directly or indirectly inspired by a theme found in an existing work or the two themes simply have serendipitous similarities stemming out of a similar frame of mind or similar circumstances, it is often extremely valuable to scholars studying these two works to establish the points of similarity and comment on them. Even more valuable is the ability to ensure that anybody visiting either of the two instances of the theme is made aware of the other instance. Regardless of literary genre, one of the most colorful sources of such thematic variations is Mythology. Variants of myths have existed first through the oral tradition and then they began being incorporated in literary works. Different oral tradition branches often distorted some of the details in a myth creating variants which would be recognized as the same myth but sometimes with substantial differences. As myths started being incorporated into literary works the creativity of the author would also come into play providing intentional variations. These similarities and differences have a lot to tell us as we study these literary works. We tend to wonder whether a small changed detail has some larger hidden significance.

A wonderful example is the account of the Flood as related by the book of Genesis in the Bible and by the Epic of Gilgamesh. When the tablets containing the Gilgamesh story first started being unearthed and being translated in the nineteenth century, it was the Flood account that created a big sensation. The two accounts bear remarkable similarities but also several differences. With a Framework like the one envisioned here, a scholar could establish numerous Fragment Associations between the two texts pointing out the similarities and the differences and commenting on each one of them. In the first example, (Figure 95), we see an association between the entire bird scout section in the Noah story in the Bible and a single reference to the sending of a swallow in the Epic of Gilgamesh. The fact that the fragment defined in Gilgamesh focuses on the swallow and it is linked to a fragment containing the entire section in Genesis reveals the intention of the fragment creator. It appears that this Fragment Association is meant to point out that there is no mention of a swallow in the Bible. This may or may not have any significance in this particular case, but this example draws attention to it.

And it came to pass at the end of forty days, that Noah opened the window of the ark which he had made: And he sent forth a raven, which went forth to and fro, until the waters were dried up from off the earth. Also he sent forth a dove from him, to see if the waters were abated from off the face of the ground; But the dove found no rest for the sole of her foot, and she returned unto him into the ark, for the waters were on the face of the whole earth: then he put forth his hand, and took her, and pulled her in unto him into the ark. And he stayed yet other seven days; and again he sent forth the dove out of the ark; And the dove came in to him in the evening; and, lo, in her mouth was an olive leaf pluckt off: so Noah knew that the waters were abated from off the earth. And he stayed yet other seven days; and sent forth the dove; which returned not again unto him any more.
Genesis 8 verses 6-12

When the seventh day dawned I loosed a dove and let her go. She flew away, but finding no resting place she returned. I loosed a swallow, and she flew away but finding no resting place she returned. I loosed a raven, she saw that the waters had retreated, she ate, she flew around, she cawed, and she did not come back. Then I threw everything open to the four winds, I made a sacrifice and poured out a libation on the mountain top.
Gilgamesh Tablet 11

Figure 95 - The Flood - Example 1

The same two parts of the myth can invite a variety of Fragment Associations, each concentrating on a specific detail. The example in Figure 96 draws attention to another difference between the two versions of the myth. By juxtaposing the two fragments one does not fail to observe that the very well-known image of the dove bringing back an olive branch is missing from the Gilgamesh version. Scholars may have several theories regarding this omission. With the conceived Framework this entire scholarly activity can be brought together.

And it came to pass at the end of forty days, that Noah opened the window of the ark which he had made: And he sent forth a raven, which went forth to and fro, until the waters were dried up from off the earth. Also he sent forth a dove from him, to see if the waters were abated from off the face of the ground; But the dove found no rest for the sole of her foot, and she returned unto him into the ark, for the waters were on the face of the whole earth: then he put forth his hand, and took her, and pulled her in unto him into the ark. And he stayed yet other seven days; and again he sent forth the dove out of the ark; And the dove came in to him in the evening; and, lo, in her mouth was an olive leaf pluckt off: so Noah knew that the waters were abated from off the earth. And he stayed yet other seven days; and sent forth the dove; which returned not again unto him any more. **Genesis 8 verses 6-12**	When the seventh day dawned I loosed a dove and let her go. She flew away, but finding no resting place she returned. I loosed a swallow, and she flew away but finding no resting place she returned. I loosed a raven, she saw that the waters had retreated, she ate, she flew around, she cawed, and she did not come back. Then I threw everything open to the four winds, I made a sacrifice and poured out a libation on the mountain top. **Gilgamesh Tablet 11**

Figure 96 - The Flood - Example 2

One can imagine a scholar establishing this Fragment Association providing some commentary about the issue and subsequently other scholars discussing the issue in separate articles and creating Fragment Associations between fragments within their discussion and this Fragment Association. Figure 97 demonstrates this capability along with the respective XML entries for each item involved. The two Information Entity entries (“Genesis” and “Gilgamesh”) are created in the Information Entity Caching Component along with the cached copies of the Information Entities. The two Information Entities have two distinct creator Identities, “Identity 1” and “Identity 2”. However, the two Information Fragments in each one of these two Information Entities have been created by the same person (“Scholar X”). They were created by

“Scholar X” along with Information Fragment association “FA-1”, which is the one associating the two Information Fragments. When “Scholar X” logged into the Fragment Association Creation Interface, everything s/he created was assigned the Identity Id “Scholar X”. The XML entries of the Information Fragments and Information Fragment Association were saved as part of the association process in the respective components of the Information Fragment Registry. Since these entries have “Scholar X” as their creator, only “Scholar X” can alter them. In addition to defining the boundaries of the two Information Fragments and establishing the Information Fragment Association, “Scholar X” assigns a Relationship Qualifier “q1” and two Descriptive Qualifiers “q2” and “q3” to the Information Fragment Association. The Relationship Qualifier in this case can stand for a Qualifier such as “Is Similar To”. In other words, it will be the predicate in a statement regarding the two Information Fragments. The two Descriptive Qualifiers can be terms describing the subject matter of this Information Fragment Association. Both Relationship and Descriptive Qualifiers were picked from lists populated by members of specific Qualifier Collections. In this particular example, “Scholar X” is likely to have used a Qualifier Collection relevant to Mythology, Literature or Religion.

Subsequently, “Scholar A” finds Information Fragment Association “FA-1” and creates another Information Fragment Association “FA-2” which associates “FA-1” with Information Fragment “Fragment a” within an article s/he is writing. This means that an end-user encountering “Fragment 1” using a Fragment Association Viewing Interface would be able to navigate to “Fragment 2” and “Fragment a”. However, if this user has applied a Filter which contains “Scholar X” but not “Scholar A”, only “Fragment 2” is visible.

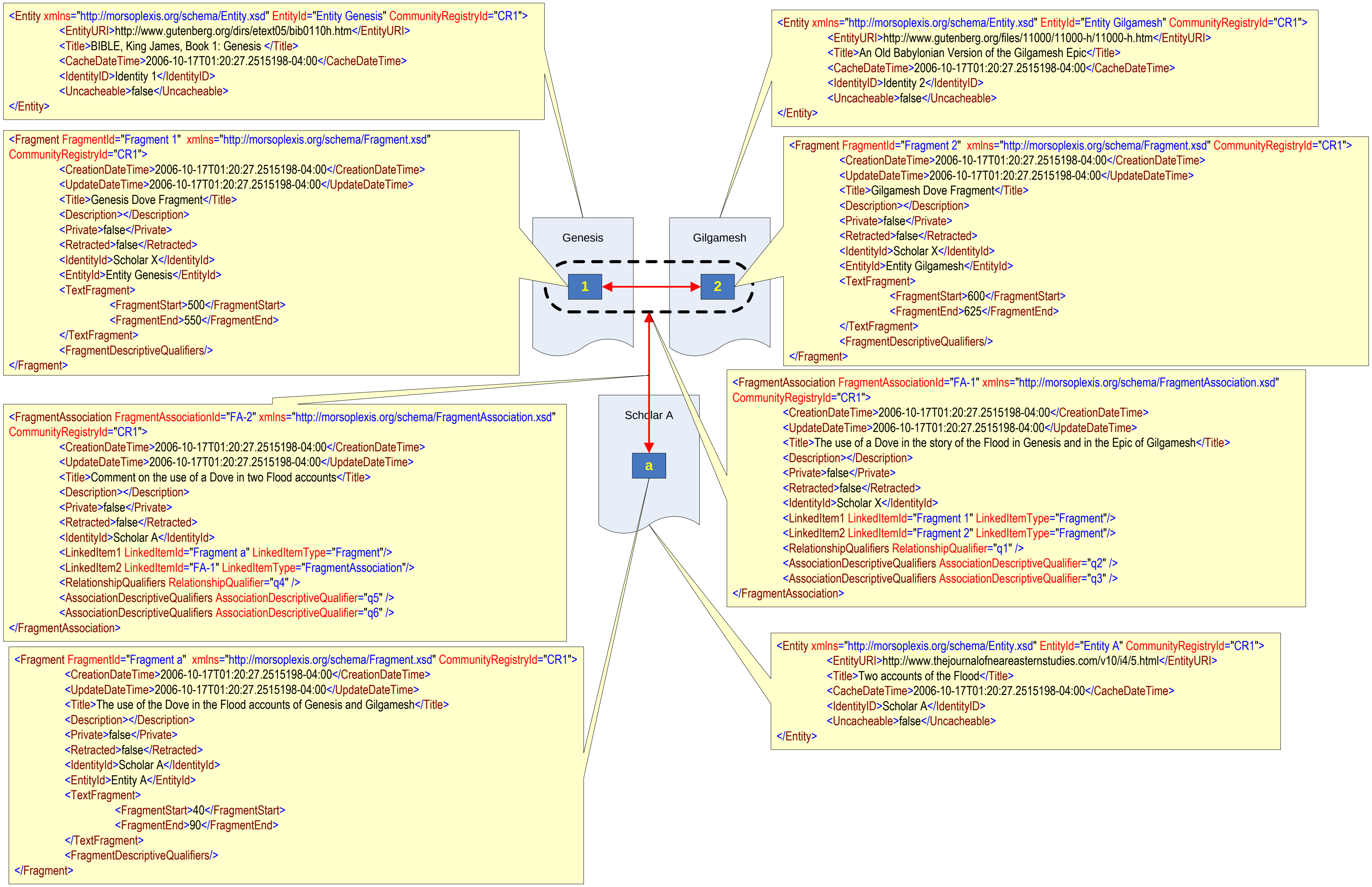


Figure 97 - Flood account example XML

These capabilities generate a type of commentary which breaks the barriers of traditional commentaries. A commentary has been traditionally a type of work comprising of a set of well-organized comments referring to specific fragments within a main primary work. The primary work has sometimes not even been included, yet it still constitutes the backbone of the commentary. The traditional commentary was primarily the work of a single person, or at least a compilation of comments made by multiple contributors but still put together by a single editorial hand. The advent of the computer and the network environment made it easy for the traditional commentary to be opened to contributions by multiple individuals. Numerous systems allow for the addition of comments and even responses to comments. Yet the main backbone of the commentary remains the single primary text which gave birth to the commentary. With the envisioned framework, every single Information Entity has the potential of being both the subject of a distributed commentary and itself a contributor to a multitude of distributed commentaries. In other words, as Figure 98 demonstrates Information Fragments constituting a commentary exist in parallel and they assume the primary text-comment relationship only when the focus is placed on one particular Information Fragment. When viewing Information Entity A, the Information Fragments b1, c1, d1 and e1 become parts of the commentary of Information Entity A. When viewing Information Entity B, Information Fragment a1 becomes part of the commentary of Information Entity B.

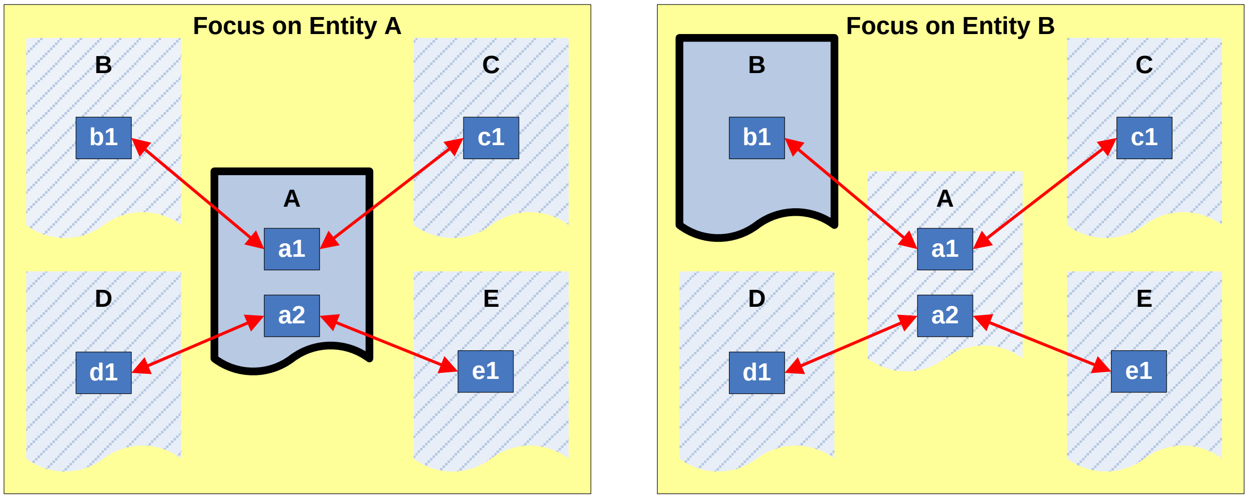


Figure 98 - Change of Focus

This commentary functionality as displayed in

Figure 98 is accomplished in the following way within the Morsoplexis Framework:

- A query is initiated against the Information Entity Caching Component using the identifier for Entity “A”
- The cached Entity is displayed in a browser session of the Fragment Association Viewing Interface
- Queries are initiated against the Information Fragment, Information Fragment Association and Information Fragment Collection components of the Information Fragment Registry
- If the user has enabled any Filters, the queries are narrowed down by applying these Filters from the Filter Registry
- The boundaries of the Information Fragments on this Information Entity are retrieved from the Information Fragment entries making the display of individual Information Fragments possible
- The user is able to see the Information Fragments contained in this Information Entity as well as a tree of Information Fragment Associations and Information Fragment Collections. In this example, we see Information Fragments a1 and a2 and a tree of associated fragments b1, c1, d1 and e1

- When the focus changes by following one of the links on this tree, the entire process is being repeated by loading another Information Entity, retrieving the associated Information Fragment's boundaries and displaying it.

This kind of commentary bridges several works, both primary and secondary sources, in one large seamless resource the structure of which transcends the rigidity of the Information Entities which constitute it. This means that "thematic" Information Fragment groups serve as important information resources themselves even though they are the creations of different authors, residing within different works and have been created asynchronously. "Thematic Information Fragment Groups" means Information Fragments and Information Fragment Associations associated together on account of the fact that they are dealing with a common or related theme. The example demonstrated in Figure 96 can be expanded further to demonstrate a Thematic Information Fragment Group. For this example the following Information Fragments are being considered (the two examples of comments by scholars being fictitious and being simply used in support of the example and not as pieces of real scholarship):

Scholar A:
The olive branch is symbol of peace, signaling a truce between God and mankind. As such, it has no place in the Gilgamesh story in which the polytheistic establishment and the difference of opinion among the deities prevents such a truce from being so patently announced

Scholar B:
The olive branch derives its peace symbolism from the Genesis story and there is no indication that this symbolism predates it. Therefore, the Genesis story does not employ the dove and olive branch as a symbol but rather in its literal sense. Consequently, the absence of the olive branch from the Gilgamesh story cannot be given

any particular significance.

Apollonius Rhodius - Argonautica Book 2 Lines 328-339:

First entrust the attempt to a dove when ye have sent her forth from the ship. And if she escapes safe with her wings between the rocks to the open sea, then no more do ye refrain from the path, but grip your oars well in your hands and cleave the sea's narrow strait, for the light of safety will be not so much in prayer as in strength of hands. Wherefore let all else go and labour boldly with might and main, but ere then implore the gods as ye will, I forbid you not. But if she flies onward and perishes midway, then do ye turn back;

Apollonius Rhodius - Argonautica Book 2 Lines 556-572:

… and then Euphemus grasped the dove in his hand and started to mount the prow; and they, at the bidding of Tiphys, son of Hagnias, rowed with good will to drive Argo between the rocks, trusting to their strength. And as they rounded a bend they saw the rocks opening for the last time of all. Their spirit melted within them; and Euphemus sent forth the dove to dart forward in flight; and they all together raised their heads to look; but she flew between them, and the rocks again rushed together and crashed as they met face to face. And the foam leapt up in a mass like a cloud; awful was the thunder of the sea; and all round them the mighty welkin roared. The hollow caves beneath the rugged cliffs rumbled as the sea came surging in; and the white foam of the dashing wave spurted high above the cliff. Next the current whirled the ship round. And the rocks shore away the end of the dove's tail-feathers; but away she flew unscathed.

Scholar C:

A dove serves as a forerunner, its fate serving as a prophecy for the fate of those following

Scholar D:
Scholar C makes an interesting observation regarding the Argonautica story which can be applied cross-culturally to a pair of Near Eastern stories…

Table 36- Thematic Information Fragment Group Example

The above fragments can be associated together in the fashion shown in Figure 99. This figure illustrates the grouping together of several Information Fragments in Information Entities (such as Genesis and the Argonautica) which would normally not have been seen together. This Thematic Information Fragment Group stands alone within the information universe in that it tackles a very specific theme. The figure is presented out of scale to aid with the presentation, but the reality is that these fragments are very small fragments in large bodies of content. The Morsoplexis Framework provides the capability of creating potentially extensive groupings of thematically interrelated fragments residing within a vast space of large mostly unrelated works. This example demonstrates several aspects of the Morsoplexis Framework:

- A Fragment Association is established between Information Fragments "1" and "2". Information Fragment "a" is then associated with the Fragment Association of "a" and "b", providing an example of an association between an Information Fragment and an Information Fragment Association. Information Fragment "a" is commenting on the relationship, similarities and differences between Information Fragments "1" and "2", therefore it is fitting that it be associated with the Fragment Association between them and not with each one of them individually
- Information Fragments "3" and "4" are associated providing an example of two Information Fragments within the same Information Entity being associated together in order to condense the content thematically. In this case Information Fragment "3" contains Phineus' prophecy and advice for using a dove, and Information Fragment "4" contains the incident of actually using the dove.

- Information Fragment “c” is associated as “a” with the Fragment Association of “3” and “4” providing a comment.
- Scholar D creates a Fragment Association between Fragment Association of “1” and “2” and Information Fragment “c” and associates this new Fragment association with a comment in an article he is writing (Information Fragment “d”).

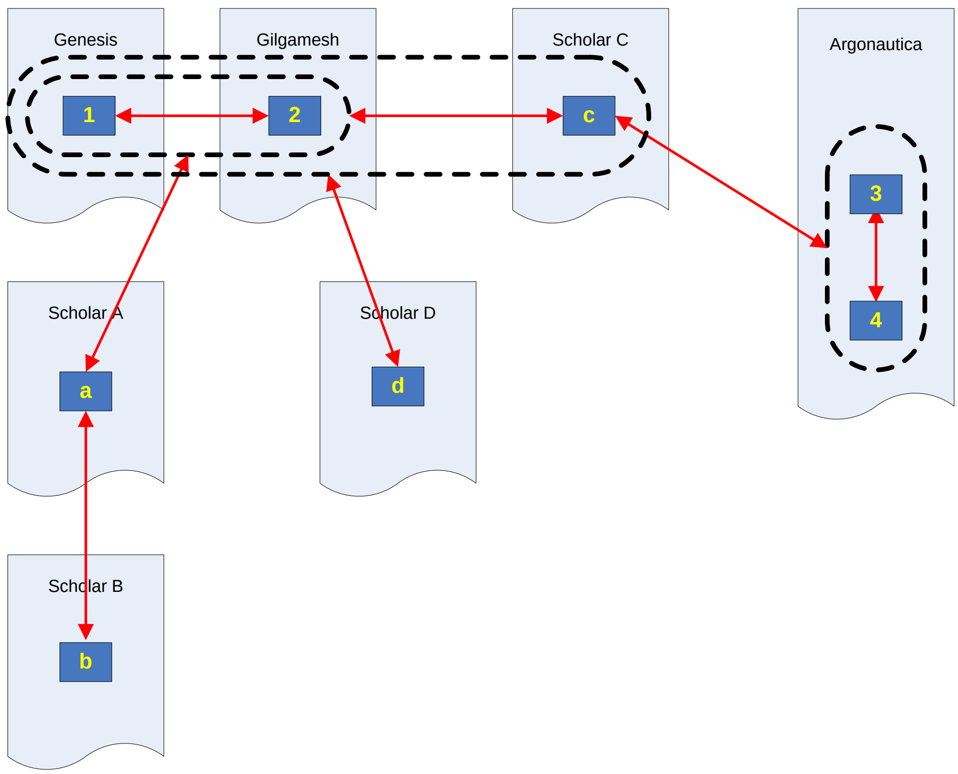


Figure 99 - Thematic Information Fragments Group

0.8.2 Focused Content Navigation

Traditionally News Organizations, Archives and other institutions have been in the habit of collecting newspaper and magazine clippings on very specific topics. These clippings have been usually kept in vertical files and they often consisted of partial photocopies of any possible resource, including books. These photocopies have often been highlighted in the sections discussing the specific topic being focused on.

Examples of such focused collections include gatherings of carefully selected bits and pieces of resources mentioning somebody's name, announcements, comments and pictures of a specific event such as the dedication of a new athletic facility, and useful information regarding some specific political or social issue such as the use of facial recognition software or the negative nature of television commercials used in political campaigns.

What is particularly interesting about such topics is that more often than not their discussion appears within the discussion of a larger more newsworthy issue. For example, the discussion regarding the use of facial recognition software has appeared over the last few years within context related to terrorism. It has also appeared within the reporting of major sporting events as one of the security measures being employed and within the discussion of measures taken to prevent child abduction. The traditional practice would have aggregated that scattered information in a single vertical folder highlighting the paragraphs dealing with the topic of interest. The intention of the proposed Framework is to provide the capability for this same kind of focused content aggregation in an environment which allows the users of this information to be also contributors to this aggregation effort.

The success of Social Bookmarking sites provides evidence of the fact that ordinary internet users are now ready to join forces in organizing information and publicly sharing the fruits of their efforts provided that the effort is rendered minimal by a simple and efficient interface design. This opens the door for additional efforts which may further enhance information organization sharing.

One of the main drawbacks of traditional bookmarks is the fact that they usually refer to a large body of content, making it difficult for somebody visiting the bookmark to identify the purpose the resource has been bookmarked for. It is not uncommon for somebody to bookmark a resource and then fail to remember the reason why s/he had bookmarked that resource. The bookmark refers to an entire resource, but the purpose for bookmarking may have been a specific piece of information in a small subset of the content of that resource.

The Morsoplexis Framework does not only provide the capability of focusing on the intended content, but it also enables users to navigate from the one Information Fragment to the other because of the associations users are creating between them. This navigation overcomes the limits of traditional navigation and allows users seeking information to take advantage of the painstaking work others may have done to identify these useful fragments.

In order to demonstrate the focused content navigation capability the following example is provided. The Information Fragments listed below are dealing with instances in which GPS (Global Positioning System) devices malfunctioned or their reliability was challenged:

http://www.crowsey.com/news.asp Peterson Trial - GPS evidence exclusion hearing
From Fox News - Tuesday, February 12, 2004, New York, NY — Many cities and towns across the United States are turning to technology to help monitor house-arrest prisoners and to keep jails from busting at the seams. Some are even using global positioning satellites (search) to keep track of offenders. "It's what we consider a viable alternative to actual incarceration," said Lt. Wayne Garner of Louisiana's LaFourche Parish Sheriff's Office, which has monitored 422 offenders in 2.5 years. "It keeps the jail space open for the more serious offenders." But the technology has its detractors. In the legal case of accused double murderer Scott Peterson, defense attorneys tried to convince a judge Wednesday that GPS is inaccurate and unreliable. They claim their client, who is accused of killing his wife and unborn child, was tracked by GPS devices placed by authorities in vehicles he drove after Laci disappeared on Christmas Eve in 2002. Geragos wants all the GPS tracking evidence excluded from the trial.

http://www.scottisinnocent.com/Trial/Pretrial/loomis.htm Peterson Trial Transcript – GPS and motion detector – First Fragment
Mark Geragos: Now, I had picked 1-D. If we go to the next portion of this, 1-G, which appears also to be 1-9, and another series of entries. It also appears that we have got the same issue here where there is speeds and indication of no motion, 51 miles an hour, 25 miles an hour, 31 miles an hour, that shows no motion on that column, yet the speed being, probably on surface streets, of a substantial amount of speed; isn't that correct? Peter Loomis: These are what these data sets show, yes.

http://www.scottisinnocent.com/Trial/Pretrial/loomis.htm Peterson Trial Transcript – GPS and motion detector – Second Fragment
Peter Loomis: Okay. I notice many cases here where, of course, as we expect little or no speed, and no motion was reported; but there are other cases where there is substantial speed. I see one here, 24 miles per hour 30 miles per hour, and the no motion is also indicated. Mark Geragos: So there would be a, what appears to be a longitude and latitude, 24 miles per hour, yet it says there is no motion whatsoever; isn't that correct? Peter Loomis: That's what it says.

http://www.scottisinnocent.com/Trial/Pretrial/loomis.htm Peterson Trial Transcript – GPS and motion detector – Third Fragment
Mark Geragos: There appears to be some problems there as well; is that not correct? We have one 38,000 miles per hour; is that correct? Peter Loomis: This is interesting. This could indicate what Orion is actually doing here. This particular instance is an instance where it is just exactly one degree longitude. If Orion is calculating their speeds by a change in

position over those five seconds, it went, oh, 60 miles or so over five seconds. And that very well might be 38,000 miles per hour. If they have a separate motion detector in the device, that's an actual electronic little accelerometer, tells whether they are moving or not.

http://transcripts.cnn.com/TRANSCRIPTS/0601/05/ltm.06.html Tragedy At Sago Mines – GPS difficulties (CNN Transcripts)
But finding them was very difficult. You said you could probably have been ready to drill in about 10 or 11 hours after the incident. It wasn't until 21 hours after the explosion that you were actually to begin boring into the ground. That's a long time, but it was difficult for them to pinpoint the location, wasn't it? ROSS: It was. The surveyors were having problems with the satellites on the GPS system and everything. It was very cloudy. There was lightening and a lot of bad conditions on the surface. And we were standing by. And then we had an original site that was done on the GPS by a local individual here and then we had to move it about 30 feet.

http://www.wral.com/news/3400765/detail.html N.C. Boaters Urged Not To Use GPS This Weekend
POSTED: 4:58 pm EDT June 9, 2004 PORTSMOUTH, Va. -- Boaters will be unable to rely on Global Positioning System equipment and possibly cell phones because of scheduled disruptions from Friday to Sunday that will affect mariners transiting the waters of North Carolina and parts of Florida. GPS interference testing conducted during an exercise by the Department of Defense will make the GPS signal unreliable and may affect cell phone signals from about 50 to 60 miles off the coast of Cape Hatteras, N.C., and approximately 80 miles out between Jacksonville and Melbourne, Fla.

Table 37 - Focused Content Navigation Example

These Information Fragments are found in unrelated contexts but they are themselves related. Identifying these fragments with traditional web searching techniques is an arduous task. By enabling users to establish Information Fragment Associations between these Information Fragments, the Morsoplexis Framework provides the capability of creating browsing interfaces for quick navigation between associated Information Fragments. This example demonstrates the use of what is a typical set of Information Fragment Associations for the purpose of enhancing browsing. An Information Fragment Collection is created consisting of Information Fragments from the transcripts of one of the preliminary hearings of a well-known criminal trial. The three Information Fragments highlight the points of the discussion during which some statements were made regarding the problematic readings produced by a GPS tracking device used in that case. The three Information Fragments (urn-6, urn-7 and urn-8) are placed in an Information Fragment Collection because their contents complement each other in dealing with this particular issue. Subsequently a Fragment Association is created between this Information Fragment Collection and an Information Fragment (urn-4) within an article covering this portion of the trial. This Fragment Association is in turn associated with an Information Fragment within the transcripts of a news interview regarding the Sago mine tragedy and that one with another news story. The ability to connect these fragments together in a meaningful way and to enable navigation between them while preserving the context is an invaluable asset for anybody visiting any of these fragments. Figure 100 illustrates a collapsible menu system built on the established relationships described above.

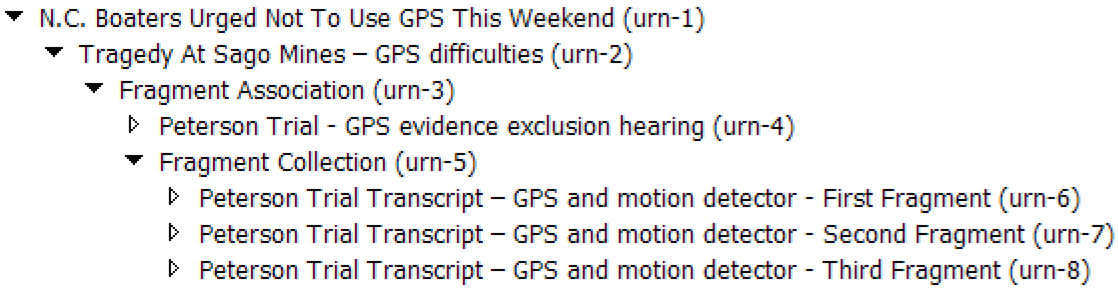


Figure 100 - Focused Content Navigation

Each menu item retrieves and display the Information Fragment it corresponds to as shown by Figure 101.

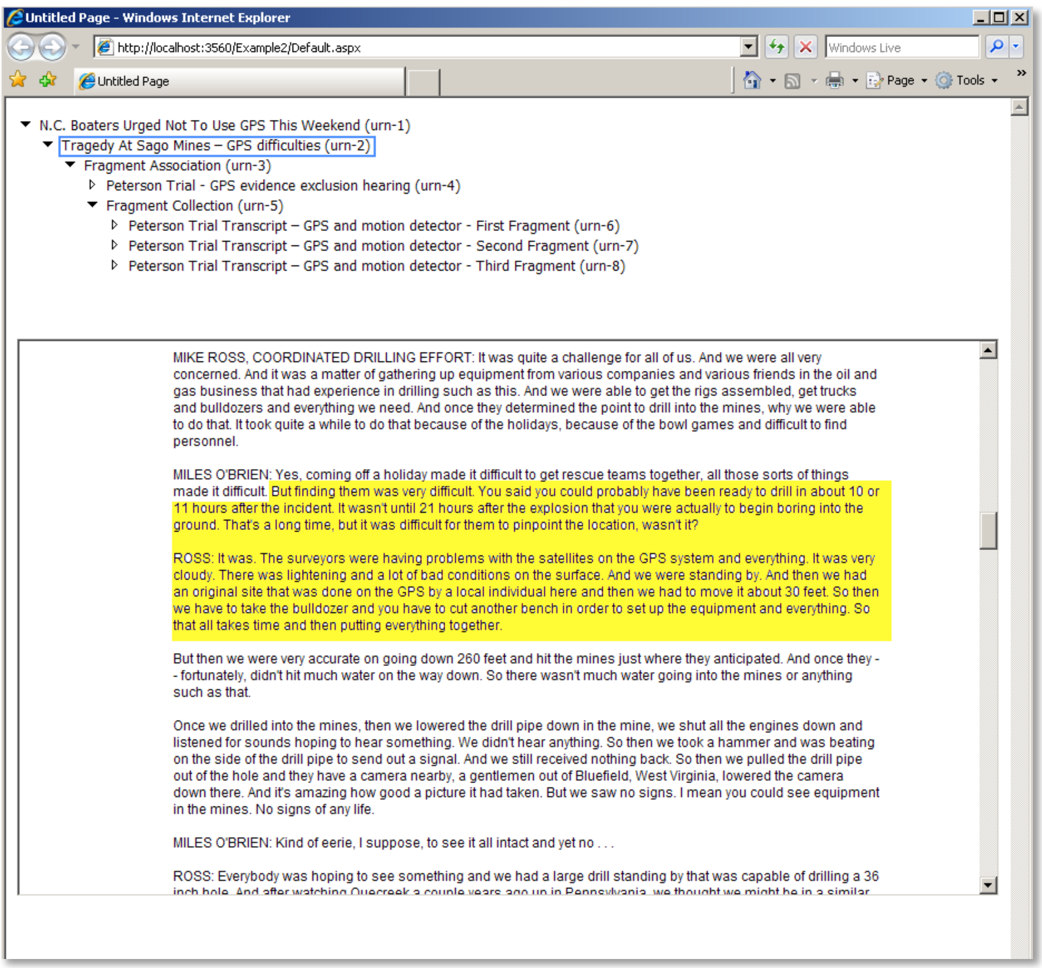


Figure 101 - Focused Content Display

The above example demonstrates that by identifying one relevant Information Fragment a user can easily take advantage of the effort others have already put into bringing together relevant small pieces of information. A user is also able to adjust the set of Information Fragments being displayed with these navigation interfaces. Since the envisioned Framework supports Filters which limit the viewable Fragment Associations to those created by Identities belonging to a certain group, navigation varies depending on the profile being used.

This navigation is made possible by queries to the Information Fragment, Information Fragment Association and Information Fragment Collection components of the Information Fragment Registry. The tree, with a possibly customizable depth of branches, is being built with these queries. Just like in the case of the previous example the composition of this tree depends on the Filters the end-user has applied. The query used to produce this tree can be narrowed down by applying these Filters from the Filter Registry.

0.8.3 Potential Contribution to Search Engines

The desired solution provides well-defined, relevant and focused content fragments on both ends of a link, thus allowing existing search algorithms to be enhanced to take advantage of them. In their ARC system, Chakrabarti et al.[73] have weighted links by considering text surrounding the anchor on the referring page. One could envision the benefits of being able to have relevant information fragments residing in different information entities linked directly as opposed to having a simple link pointing to the entire information entity.

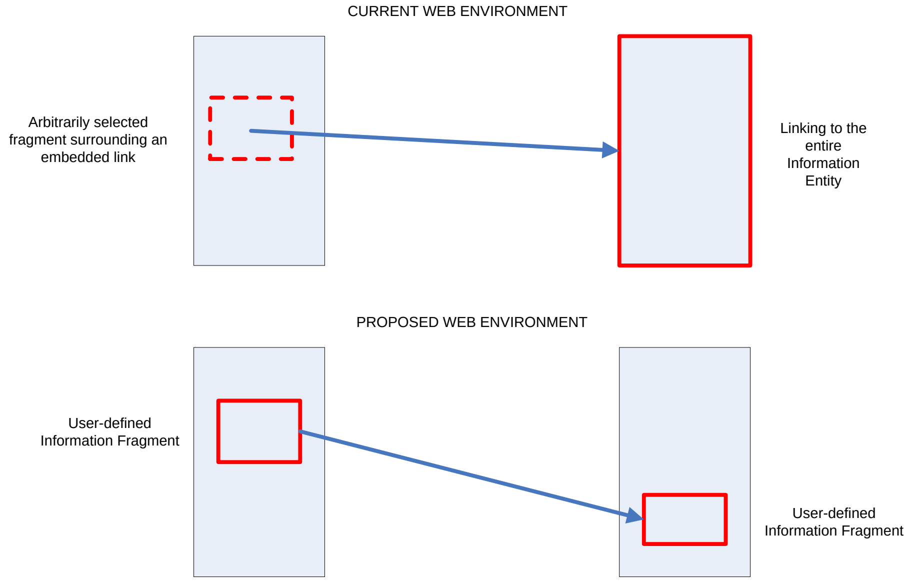


Figure 102 - Information Entity linking vs. Information Fragment linking

As Figure 102 demonstrates, what the current web environment provides is a link between a point in an information entity and another information entity. Search algorithms may take into consideration an arbitrarily defined Information Fragment surrounding this link. On the other end of the link, what we have is a reference to the entire Information Entity. The Morsoplexis Framework provides a link between a user-defined information fragment and another user-defined information fragment. The precision provided effortlessly by the user is a very useful tool available to search algorithms for the purpose of better weighing and producing better results

A more immediate and more easily demonstrable feature is one involving the blending of traditional search engine results with information derived from the registry. For example, an interface using currently available search engine web service-based APIs presents results to users expanding each one of the search result items. Conceivably, each one of these items can be expanded even with something as simple as an additional link. In the example shown in Table 38 below, a "Fragment Associations" link is added to a standard Google search result item. This link presumably leads to a list of Fragment Associations established for that particular Information Entity.

PADI - **Intellectual property rights** management This topic deals with the intellectual property rights and copyright issues which relate to preserving access to digital information. www.nla.gov.au/padi/topics/28.html - 124k - Cached Similar pages Fragment Associations🡐

Table 38 - Possible Google Results Example

APPENDIX B

QUERIES SUBMITTED BY SUBJECT GROUP C

Query Number	Query	Fragment Character Length	Total Results in Result Set	Number of Phrases in Query	Number of Words in Query	Phrase Type	Max Phrase Character Length	Total Phrase Character Length
1	british prime minister blair "power-sharing deal" "northern ireland's main"	151	56	2	4	Both	23	41
2	us visa iranian "President Mahmoud Ahmadinejad" un security council programme "extra sanctions against iran"	392	6	2	7	Both	29	57
3	New York, San Francisco, Portland 3,200 US troops.	329	801	0	9	Word	0	0
4	hanged dawn tuesday saddam hussein's 148 shiites dujail	212	820	0	8	Word	0	0
5	firing eight prosecutors. u.s. attorney general aide invoked constitutional right	211	5665	0	10	Word	0	0
6	valerie plame glamorous democrats' politically motivated smear husband	285	525	0	8	Word	0	0
7	"forever" stamp 41 cents "beginning in may"	265	88	2	3	Both	16	23
8	"plunged to a lifetime low" Vonage Holdings Corp.	249	6	1	3	Both	25	25
9	The Smithsonian draws more than 23 million visitors a year, most of them to museums along Washington, D.C.'s	358	33	0	18	Word	0	0
10	iranians abused russia relations "harmed our image"	376	9	1	4	Both	16	16

Query Number	Query	Fragment Character Length	Total Results in Result Set	Number of Phrases in Query	Number of Words in Query	Phrase Type	Max Phrase Character Length	Total Phrase Character Length
11	10 years in prison soldier iraqi detainee "freed and told to run before being shot"	227	30	1	7	Both	39	39
12	spring 2008 timeline withdrawal deadlines	376	4359	0	5	Word	0	0
13	"Few in the Basque Country doubt ETA's days are numbered whether talks prosper or not."	719	2	1	0	Phrase	85	85
14	boarding checking Indian merchant vessel routine	406	746	0	6	Word	0	0
15	"sectarian bloodshed was leading to civil war"	491	6	1	0	Phrase	44	44
16	"Vilsacks became among the most high-profile backers of Clinton's bid."	585	39	1	0	Phrase	69	69
17	proposed security council mission kosovo serbia the Ahtisaari churkin	824	93	0	9	Word	0	0
18	death green zone Katyusha rocket	598	6899	0	5	Word	0	0
19	jan. 28 chlorine gas attack	593	19843	0	5	Word	0	0
20	political groundwork veto showdown democratic	616	4099	0	5	Word	0	0
21	tuskegee black elite aviators world war	518	227	0	6	Word	0	0
22	wal-mart chicago "thousands of job applicants"	749	25	1	2	Phrase	27	27
23	stubborn inflation could upend	509	123	0	4	Word	0	0
24	companies' DVR services for fresh TV episodes	655	10676	0	7	Word	0	0
25	Hollywood film+"cultural and psychological warfare"+ISNA	685	22	1	3	Both	34	34
26	U.S. accuse +nuclear reactor outside the southern city of Bushehr is not part of Irans dispute with the U.N.	673	326	0	19	Word	0	0
27	Retreat +"British troops on the run"+"Menzies Campbell"+a base outside Basra	659	2	2	5	Both	25	41
28	"EADS' share price down 2% "+"from 25 to only nine "	362	1	2	0	Phrase	26	47
29	"fbi-patriot-act " + "illegally used"	554	25	2	0	Phrase	16	30
30	"Iranian militants seized the U.S. Embassy in Tehran following the 1979 Islamic revolution"+"Mehdi Karroubi"	572	66	2	0	Phrase	89	103
31	"radical Shiite cleric Muqtada al-Sadr"+"attack against al-Darraji"	325	20	2	0	Phrase	37	62

Query Number	Query	Fragment Character Length	Total Results in Result Set	Number of Phrases in Query	Number of Words in Query	Phrase Type	Max Phrase Character Length	Total Phrase Character Length
32	"a provision"+"Stephen Hadley" +"get a safe haven in Iraq "	687	6	3	0	Phrase	25	50
33	Airbus+2006+"another substantial loss"	318	4	1	1	Phrase	24	24
34	'one of us'+ Stanley Crouch+"first black president"	373	198	1	5	Both	21	21
35	"Romania's two main ruling parties met early Monday"	177	2	1	0	Phrase	50	50
36	"Blair warned Iran on Tuesday"	194	99	1	0	Phrase	28	28
37	"BAGHDAD (Reuters) - Bomb attacks killed 77 people in Iraq on Tuesday"	165	23	1	0	Phrase	68	68
38	Bush's proposal+"coal-to-liquids"+"natural gas liquids"	361	21	2	2	Phrase	19	34
39	"Burger King Holdings Inc. said"+"confine their animals in cages"	199	6	2	0	Phrase	30	60
40	Riyadh summit+"recognising a Palestinian partner"	345	2	1	2	Phrase	33	33
41	" ABU DHABI"+"not allow anyone to use its territory"	225	4	2	0	Phrase	37	47
42	"Chief government advisers accepted"+655,000 Iraqis died	484	11	1	4	Both	34	34
43	"truck bombings in Tal Afar"+"Wednesday, killing as many as 60"	371	177	2	0	Phrase	32	58
44	"lawsuit against"+Donald Rumsfeld+"Two human rights groups"	485	31	2	2	Phrase	23	38
45	Hillary Clinton+"remarkable seven days" +"raised $2.6 million"	335	1	2	2	Phrase	21	40
46	"Beazer Homes USA Inc."+FBI+"federal prosecutor"	304	227	2	1	Phrase	21	39
47	SAN FRANCISCO (AP) +approved a ban+"plastic grocery bags"	484	374	1	6	Both	20	20
48	a young militant Islamist commander its leader in Mogadishu as fighting raged for a second day in the coastal capital	179	254	0	20	Word	0	0
49	zimbabwe gideon gono price petrol 200%	160	138	0	6	Word	0	0
50	Moqtada al-Sadr killing U.S soldiers Kerbala January	208	420	0	7	Word	0	0
51	states moving 2008 primaries early February	148	52390	0	6	Word	0	0

Query Number	Query	Fragment Character Length	Total Results in Result Set	Number of Phrases in Query	Number of Words in Query	Phrase Type	Max Phrase Character Length	Total Phrase Character Length
52	guru Charles Simonyi thursday nervous appearances space	216	101	0	7	Word	0	0
53	judge blow internet pornography striking 1998 U.S. law	231	14722	0	8	Word	0	0
54	talks north korea abruptly thursday "no progress"	83	526	1	5	Both	11	11
55	rocket prime minister's thursday iraq first	151	19630	0	6	Word	0	0
56	rocket prime minister's thursday iraq first	244	19630	0	6	Word	0	0
57	democrats split liberals undermining U.S. troops war-fighting	222	334	0	7	Word	0	0
58	january 2006 hansen complained NASA deutsch former intern no training	301	59	0	10	Word	0	0
59	News Corp "unveil YouTube"	47	176	1	2	Phrase	14	14
60	U.S. forces Iraq Iran-made	52	4288	0	4	Word	0	0
61	Christopher Dodd thursday hearing subprime mortgage	179	643	0	6	Word	0	0
62	"texas senate"	13	52254	1	0	Phrase	12	12
63	recalled products menu foods "United States, the Food and Drug Administration has said. "	174	12	1	4	Both	58	58
64	oracle sued sap ag thursday "steal copyrighted software	180	125	0	8	Word	0	0
65	olmert "within ten days" protocols	270	24	1	2	Phrase	15	15
66	iranian navy gulf "Admiral Sajad Kushaki told state television."	391	5	1	3	Both	44	44
67	"the spectre of a broader insurgency involving foreign Islamic extremists linked to Osama bin Laden's terror network."	229	13	1	0	Phrase	116	116
68	U.N. Secretary-General "Unhurt After Nearby Attack In Baghdad's Green Zone"	76	12	1	2	Phrase	50	50
69	"that they will only lay down their arms when they have a final timetable for the withdrawal of foreign troops from Iraq. "	227	2	1	0	Phrase	121	121

Query Number	Query	Fragment Character Length	Total Results in Result Set	Number of Phrases in Query	Number of Words in Query	Phrase Type	Max Phrase Character Length	Total Phrase Character Length
70	"Stark County Recorder's records show that New Century has done more than $100 million in business here during the last decade. "	296	77	1	0	Phrase	128	128
71	"British Police Arrest 3 Suspects in Deadly London Transit Bombings "	66	41	1	0	Phrase	67	67
72	Margerie, Iraq, Oil	944	683	0	3	Word	0	0
73	U.S. troops killed five insurgents	200	256343	0	5	Word	0	0
74	UN, iraq civilian violence	1000	328380	0	4	Word	0	0
75	Obama, Clinton	271	177070	0	2	Word	0	0
76	Ahmadinejad, Iran, "technological progress"	238	250	1	2	Phrase	22	18
77	Tareq al-Hashemi, Iraq 'should talk to militants'	864	350	0	7	Word	0	0
78	two million people displaced, Iraq, BBC	1000	59699	0	6	Word	0	0
79	"U.S. soldiers killed five suspected militants," Baghdad	177	47	1	1	Phrase	46	46
80	immigration card, $137, terrorism, "Free and Secure Trade," US	763	40	1	5	Both	22	22
81	President Chirac and Nicolas Sarkozy have long had a stormy relationship	340	134	0	11	Word	0	0
82	Somali insurgents dragged soldiers' bodies through the streets of Mogadishu	198	5355	0	10	Word	0	0
83	Moonwalker Buzz Aldrin, Hualapai Reservation	538	42	0	5	Word	0	0
84	Pakistan, nuclear-capable cruise missile, 700 km (435 miles)	110	93	0	8	Word	0	0
85	Health Ministry official detained, Moqtada al-Sadr's Mehdi Army militia	221	356	0	9	Word	0	0
86	Baghdad, car bomb, Iranian embassy	74	58867	0	5	Word	0	0
87	John Edwards, wife, personal and professional life	179	197656	0	7	Word	0	0
88	Apple 1984 YouTube	269	107137	0	3	Word	0	0
89	"eliminate remaining barriers on airline ownership," EU, US	308	55	1	2	Phrase	50	50
90	world's largest retailer, 13,400 workers	110	130	0	5	Word	0	0
91	Sina-1, Shahab-3, Iran	186	175	0	3	Word	0	0

Query Number	Query	Fragment Character Length	Total Results in Result Set	Number of Phrases in Query	Number of Words in Query	Phrase Type	Max Phrase Character Length	Total Phrase Character Length
92	Ayatollah Ali Khamenei, "But if they take illegal actions, we too can take illegal actions and will do so."	200	225	1	3	Both	81	81
93	Zamili, Ministry of Health official, arrest, corruption	144	212	0	7	Word	0	0
94	Bush is standing by embattled Attorney General Alberto Gonzales	273	12518	0	9	Word	0	0
95	The Brent crude contract for May jumped $1.09 to $61.86 a barrel on the ICE Futures exchange in London	102	14	0	19	Word	0	0
96	"Tehran failed" + "February deadline to suspend enrichment" + "Russia, China, the United States"	373	90	3	0	Phrase	39	84
97	"Farzana Sayid Saidi, a 29-year-old reporter and colleague of Samiei"	351	6	1	0	Phrase	67	67
98	"The Iraq war has killed more than 3,200 U.S. military personnel and tens of thousands of Iraqis."	227	9	1	0	Phrase	96	96
99	"Insurgents in western Iraq set off three chlorine gas car bombs"	196	228	1	0	Phrase	63	63
100	"Fire swept through a nursing home in southern Russia after the night watchman ignored two alarms Tuesday"	214	193	1	0	Phrase	104	104
101	"Ties between the two Koreas, chilled by Pyongayng's decision to launch missiles and its first nuclear test in October 2006"	239	4	1	0	Phrase	122	122
102	"other women's rights activists staged a meeting in front of parliament in Tehran on Thursday afternoon"	213	2	1	0	Phrase	102	102
103	"Adel Abdul-Mahdi, one of two vice presidents, urged international support for the Iraq Compact"	291	22	1	0	Phrase	94	94
104	"George Stephanopoulos" + "he has joined Unity '08, a group that would like to elect a bipartisan ticket to the White House."	171	50	2	0	Phrase	97	118
105	"Obama's profile boasted more than 67,000 friends."	49	69	1	0	Phrase	49	49
106	"That gives Thompson every incentive to hang around and wait"	266	23	1	0	Phrase	59	59

Query Number	Query	Fragment Character Length	Total Results in Result Set	Number of Phrases in Query	Number of Words in Query	Phrase Type	Max Phrase Character Length	Total Phrase Character Length
107	"A gas explosion at the Ulyanovskaya mine in Kemerovo region kills 106 people."	210	1	1	0	Phrase	77	77
108	"On Monday South Africa, a council member, circulated amendments to the latest draft"	280	10	1	0	Phrase	83	83
109	"An aide to European Union foreign policy chief Javier Solana"	190	24	1	0	Phrase	60	60
110	"Steinmeier, who is due to meet U.S. Secretary of State Condoleezza Rice on Monday"	226	4	1	0	Phrase	81	81
111	"Britain's Prince Harry is undergoing his last major stint of army training "	189	19	1	0	Phrase	75	75
112	The adults then parked next to a market in the Adamiya area of Baghdad" + "according to the general and another defense official"	197	44	0	22	Word	0	0
113	"And more than half of those interviewed said they would support the use of British troops"	165	2	1	0	Phrase	89	89
114	"Such interviews would be private and conducted" + "Fielding said in a letter to the Senate and House Judiciary committees"	254	58	2	0	Phrase	70	116
115	"North Korea boycotted the six-party negotiations for over a year"	205	6	1	0	Phrase	64	64
116	"All households will be eligible to request up to two $40 discount coupons to buy converter boxes until $990 million"	318	36	1	0	Phrase	115	115
117	"Christopher Dell denied Washington was actively seeking regime change"	216	4	1	0	Phrase	69	69
118	"President Bush was informed of his father's condition on Sunday night"	209	60	1	0	Phrase	69	69
119	"Families who opted for compensation from the federal fund had to give up the possibility of suing the airlines"	296	2	1	0	Phrase	110	110
120	"mediators on kosovo rejected on monday"	206	19	1	0	Phrase	38	38
121	Tehran "banning Iranian arms"	310	447	1	1	Phrase	20	20
122	"Palestinian child" Gaza Hamas	212	10339	1	2	Phrase	17	17
123	"Iraqi Deputy Prime Minister" surgery	178	657	1	1	Phrase	27	27

Query Number	Query	Fragment Character Length	Total Results in Result Set	Number of Phrases in Query	Number of Words in Query	Phrase Type	Max Phrase Character Length	Total Phrase Character Length
124	"state farm" "hurricane katrina" "class action" policyholders	203	820	3	1	Phrase	17	39
125	"suspending the campaign" "Jennifer Palmieri"	269	4	2	0	Phrase	23	36
126	"Federal Communications Commission" AT&T Comcast "extra fees"	241	237	2	2	Phrase	33	43
127	Arbil "Revolutionary Guard-Qods Force"	380	52	1	1	Phrase	30	30
128	pentagon "state department" "joint command"	286	521	2	1	Phrase	16	29
129	"house democrats" "artificial date"	251	290	2	0	Phrase	15	30
130	houdini "persistent rumors" "take a second look"	216	12	2	1	Phrase	18	35
131	"federal judge" vonage friday rival	325	5698	1	3	Both	13	13
132	"ed markey" "questioned best buy"	987	1	2	0	Phrase	19	28
133	Kirchner anti-bush	642	5016	0	2	Word	0	0
134	technology "cellulosic biofuels"	799	567	1	1	Phrase	19	19
135	somali "refugee agency" 57,000	366	118	1	3	Both	14	14
136	Yemeni coastline smugglers	206	265	0	3	Word	0	0
137	"u.s. military" "five soldiers" "separate attacks"	1000	167	3	0	Phrase	16	42
138	"arrested two men suspected" car bombings	923	25	1	2	Phrase	26	26
139	"nouri al-maliki" "jalal talabani" "accountability and justice" "joint statement"	594	32	4	0	Phrase	26	70
140	fray "john sununu" "fire the attorney general"	483	18	2	1	Phrase	25	36
141	thompson "former marine officer" "november election	605	4	1	3	Both	21	21
142	"also thursday" "menu foods" "wet pet foods"	290	34	3	0	Phrase	13	36
143	"minimum wages" "region to region" china	510	123	2	1	Phrase	16	29

Table 39 - Queries Submitted by Subject Group C

APPENDIX C

SUBJECT COMMENTS

Upon completion of the assigned tasks subjects were asked to offer their opinion regarding their experience and the usefulness of Information Fragment Association.

Subject	Subject Comments	Subject Considers Fragment Association Useful? (Summarized assessment)
1	n/a	Yes (Assessment from recollection)
2	FW works way easier, it's a lot more user-friendly, but I think [Baseline] looks better, so it would be more easy to market, I guess. It's more visually appealing, but the other one [FW] definitely works a lot better. [To be able to refer to a specific fragment] would be really useful, especially when it came to articles, research, or something like that.	Yes
3	[Having the ability to associate two different fragments] definitely saves a lot of time, and the FW was a lot easier to use than the [Baseline] because it brought them up and it was a lot quicker. I did not have to continuously search for the same thing like if I would do it in like Google or whatever. So if you are writing like a paper or something that would definitely - instead of printing off the article you could use that instead, you know, so it would save both paper and time. [That is something I one could use] for College or something, especially since the FW highlighted the fragments. When I went back I noticed that. So, I mean, it's actually like the same as taking a highlighter to stuff you are printing off when writing a paper, so if you wanted to do everything all online, that's where I can see it used.	Yes

Subject	Subject Comments	Subject Considers Fragment Association Useful? (Summarized assessment)
4	Both [tools] seem to be helpful for looking at news stuff. Normally when I look for news things I just go to Google News or something like that so I don't know if I would use it but for someone who is maybe in Political Science or something doing a lot of intensive research it might be useful. [Going directly to a fragment instead of a full resource] is pretty cool. I think it would be useful if you are researching one story if you want to get different angles on the same story. If it was just everyday use I probably wouldn't use it but if I was doing a project or researching something then I would consider something like that. [Baseline] seemed to be more like a search engine so for me personally it would probably help, but [FW] was pretty nice too. I guess I would use them both if I was doing any intensive research type of thing. [Baseline] is a lot like a search engine and if I was just looking for an article on some current event or something I probably would use that because it is very easy. FW is a little less intuitive.	Maybe, but not for me
5	n/a	Maybe, but not for me (Assessment based on recollection)
6	I guess [associating fragments] might help. I cannot really see a scenario where I would need to have two different articles and have links between them. Usually I mark stuff, I print stuff.	No
7	n/a	Not sure (Assessment based on recollection)
8	n/a	Yes (Assessment based on recollection)
9	I thought that [Baseline] was more difficult to use and that FW was very basic, because I have just learned how to do this, so FW was much easier to retrieve the information and it highlighted it in the context whereas [Baseline] you saw the selection and then you had to go to the context. So, in order to remember where my mind was when I selected that I thought that FW was more effective. [Fragment association] can be useful for research. I personally don't do a lot of that myself, but knowing that these are out there that could be useful to people involved with academics and finding certain articles and what they liked about certain articles. And it also saves a lot of time, so they would be	Yes

Subject	Subject Comments	Subject Considers Fragment Association Useful? (Summarized assessment)
	able to retrieve the URL instead of having to write them down or copy them into a Word document.	
10	I liked [the idea of Fragment Association]. I found my searches very quickly. FW gave me my actual searches and it was easier to use than the [Baseline] one. I would definitely choose FW. It probably would not [at this point be applicable to] my work. Maybe for Grad School - I'm still undergrad, but definitely for Grad School or if I go to Law School I would definitely be useful.	Yes
11	[The two tools were] pretty helpful. It's pretty nice that they are lined up and that they are organized for you. The one that lines up the two fragments side-by-side for you FW is a little bit easier for you to know what article it came from. [The concept of Fragment Association] is nice especially when you are writing papers and you need to pull different ideas or concepts from many sources or whatever, it's easy to have it all right there lined up for you, and being able to store it in case you have to go back to the article.	Yes (Not a very strong endorsement)
12	I thought that for the FW program that it works really well. Just how you can go back so easy and find them and pull them up. It gave you a lot more information a lot quicker like sort of like searching all over again. I thought that was really good. [Having the ability to go from fragment to fragment] would be good. Obviously every news source is somewhat biased and having the ability to go back and forth that easily kind of lets you get more sides using it that way, so I think it is useful.	Yes

Table 40 - Subject Comments - Transcripts

APPENDIX D

LIKES AND DISLIKES

	Baseline likes	Baseline dislikes	Baseline Attractive	Baseline Intuitive	Baseline Display	Baseline Retrieval
1	can easily search for all fragments that a certain word/phrase	fragments aren't connect together in anyway.		1		1
2	It is very easy to access full web articles. The interface looks better	It's not very user friendly; it's hard to find things	1	-1	1	-1
3	I liked how it was able to show me all my fragments, and from what websites they were from	I disliked how SPURL took a long time to load when searching for fragments, or for the original websites.			1	-1
4	Attractive intuitive interface	Results were limited - took many attempts with different keywords/phrases to find the related article	1			-1
5	Colors. Layout of features. Search feature. I like how you use it though their Internet site.	Search feature is limited compared to FW	1			-1
6	Easy to search. Sort by time. Good color scheme, easy on eyes	Can't associate two fragments and stories	1	1		0
7	I can compare other articles each other	not simple much		-1		

	Baseline likes	Baseline dislikes	Baseline Attractive	Baseline Intuitive	Baseline Display	Baseline Retrieval
8	I like the interface of spurl. It is very friendly	The match ability of spurl is not good. You should choose several keywords to find the related fragments		1		-1
9	seems to have more options lists your links	more difficult to find exact saved information, cannot read it within the context			0	-1
10	The graphics	Longer time searching. Would not recognize keywords.	1			-1
11	I like that SPURL stores all articles you have selected fragments from in 3 separate folders	that it did not line your similar matching fragments up for you. You had to find them			1	-1
12	More control over what you could see, weren't locked into only the related article	Too many windows to navigate, sometimes hard to find the connections.		-1	1	

Table 41 - Baseline Likes and Dislikes

	FW likes	FW dislikes	FW Attractive	FW Intuitive	FW Display	FW Retrieval
1	Fragments are linked together. Easy to use interface	I didn't notice a clickable link to the original article. A more noticable link or adding the link would be useful.		1		1
2	Side by side comparrison makes finding corresponding fragments simple	Access to full web articles not readily apparent, less visually appealing	-1	-1	1	1
3	I liked how FW was quick in bringing up my articles and fragments. I did not feel as though I was sitting there waiting for it to load like I did with the SPURL. I also like how it automatically brought up both my fragments side by side when I originally searched for my fragment. That feature saved a lot of time.	I did not dislike anything about FW. I liked it a lot better than SPURL. If I had to choose one to use in the future I would definitely use FW.			1	1
4	Very fast & relatively easy	Not quite as intuitive/attractive as SPURL	-1	-1		1
5	Search feature appears to be more dynamic than SPURL	Colors. Layout. Appears clunky. I don't like how it is "stand alone"	-1			1
6	Can associate two fragments and stories	Interface hard to figure out and not intuitive. Didn't really get it. Color scheme needs improvement	-1	-1		1
7	simple and easy			1	1	
8	FW has the very useful matching ability, and it also has good navigating ability for the related fragments.	The interface of FW has more room to improve.	-1	1	-1	1
9	shows selection in the context, good for people first learning to find fragments	not enough options, seems very basic			1	
10	Found my searches fast & efficiently. Seeing my searches side by side	N/A			1	1
11	Everything was right there in front of you	That it did not put all the articles you liked in one easy folder for you			-1	1

	FW likes	FW dislikes	FW Attractive	FW Intuitive	FW Display	FW Retrieval
12	All-in-one interface, no unnecessary windows, extremely easy to go back to find associations	Some links that I expected to take me to the article actually took me to an extra information page that was not particularly useful			-1	1

Table 42 – FW (Test Tool) Likes and Dislikes

APPENDIX E

PRECISION, RECALL AND F-MEASURE

	Retrieved		Retrieved Relevant		Relevant
Query Number	Baseline	Test Tool	Baseline	Test Tool	Both Tools
1	1	2	1	2	17
2	1	2	1	2	2
3	1	2	1	2	33
4	1	2	1	2	33
5	1	2	1	2	2
6	1	2	1	2	6
7	1	2	1	2	2
8	1	2	1	2	10
9	1	2	1	2	2
10	1	2	1	2	20
11	1	2	1	2	20
12	1	2	1	2	2
13	2	2	2	2	2
14	1	2	1	2	33
15	1	2	1	2	2
16	5	2	5	2	20
17	1	2	1	2	17
18	1	2	1	2	10
19	1	2	1	2	33
20	1	2	1	2	33
21	1	2	1	2	2
22	1	2	1	2	2
23	1	2	1	2	10

	Retrieved		Retrieved Relevant		Relevant
Query Number	Baseline	Test Tool	Baseline	Test Tool	Both Tools
24	1	2	1	2	2
25	1	2	1	2	10
26	1	2	1	2	2
27	1	2	1	2	33
28	1	2	1	2	20
29	1	2	1	2	2
30	1	2	1	2	2
31	1	2	1	2	6
32	1	2	1	2	33
33	1	2	1	2	33
34	1	2	1	2	18
35	1	2	1	2	2
36	1	2	1	2	2
37	1	2	1	2	2
38	1	2	1	2	17
39	1	2	1	2	2
40	1	2	1	2	2
41	1	2	1	2	2
42	1	2	1	2	2
43	1	2	1	2	33
44	1	2	1	2	2
45	1	2	1	2	2
46	1	2	1	2	20
47	1	2	1	2	2
48	1	2	1	2	2
49	1	2	1	2	16
50	2	2	2	2	2
51	1	2	1	2	33
52	1	2	1	2	2
53	1	2	1	2	20
54	0	2	0	2	2
55	1	2	1	2	33
56	1	2	1	2	2
57	1	2	1	2	18
58	1	2	1	2	2
59	1	2	1	2	2
60	1	2	1	2	17
61	1	2	1	2	17

	Retrieved		Retrieved Relevant		Relevant
Query Number	Baseline	Test Tool	Baseline	Test Tool	Both Tools
62	1	2	1	2	2
63	1	2	1	2	2
64	1	2	1	2	33
65	1	2	1	2	2
66	1	2	1	2	2
67	1	2	1	2	6
68	1	2	1	2	2
69	1	2	1	2	2
70	1	2	1	2	2
71	1	2	1	2	2
72	1	2	1	2	2
73	1	2	1	2	33
74	1	2	1	2	33
75	1	2	1	2	20
76	1	2	1	2	20
77	1	2	1	2	2
78	1	2	1	2	2
79	1	2	1	2	17
80	1	2	1	2	2
81	1	2	1	2	2
82	1	2	1	2	7
83	1	2	1	2	2
84	1	2	1	2	2
85	1	2	1	2	2
86	1	2	1	2	2
87	1	2	1	2	2
88	1	2	1	2	2
89	1	2	1	2	2
90	1	2	1	2	2
91	1	2	1	2	10
92	1	2	1	2	33
93	1	2	1	2	18
94	1	2	1	2	5
95	1	2	1	2	2
96	1	2	1	2	2
97	0	2	0	2	17
98	1	2	1	2	2
99	1	2	1	2	2

	Retrieved		Retrieved Relevant		Relevant
Query Number	Baseline	Test Tool	Baseline	Test Tool	Both Tools
100	1	2	1	2	7
101	1	2	1	2	4
102	0	2	0	2	2
103	1	2	1	2	2
104	1	2	1	2	2
105	1	2	1	2	17
106	1	2	1	2	2
107	1	2	1	2	18
108	1	2	1	2	2
109	1	2	1	2	3
110	1	2	1	2	2
111	1	2	1	2	10
112	1	2	1	2	10
113	1	2	1	2	33
114	1	2	1	2	2
115	1	2	1	2	2
116	1	2	1	2	7
117	1	2	1	2	20
118	1	2	1	2	2
119	1	2	1	2	2
120	1	2	1	2	4
121	1	2	1	2	33
122	1	2	1	2	3
123	1	2	1	2	2
124	1	2	1	2	2
125	1	2	1	2	2
126	1	2	1	2	2
127	1	2	1	2	33
128	1	2	1	2	2
129	1	2	1	2	20
130	1	2	1	2	5
131	1	2	1	2	2
132	1	2	1	2	2
133	1	2	1	2	2
134	1	2	1	2	10
135	1	2	1	2	20
136	1	2	1	2	2
137	1	2	1	2	33

	Retrieved		Retrieved Relevant		Relevant
Query Number	Baseline	Test Tool	Baseline	Test Tool	Both Tools
138	1	2	1	2	33
139	1	2	1	2	18
140	1	2	1	2	2
141	1	2	1	2	2
142	1	2	1	2	2
143	1	2	1	2	2

Table 43 - Retrieved Information Fragments

	Precision		Recall		F-Measure	
Query Number	Baseline	Test Tool	Baseline	Test Tool	Baseline	Test Tool
1	1	1	0.058823529	0.117647059	0.111111	0.210526
2	1	1	0.5	1	0.666667	1
3	1	1	0.03030303	0.060606061	0.058824	0.114286
4	1	1	0.03030303	0.060606061	0.058824	0.114286
5	1	1	0.5	1	0.666667	1
6	1	1	0.166666667	0.333333333	0.285714	0.5
7	1	1	0.5	1	0.666667	1
8	1	1	0.1	0.2	0.181818	0.333333
9	1	1	0.5	1	0.666667	1
10	1	1	0.05	0.1	0.095238	0.181818
11	1	1	0.05	0.1	0.095238	0.181818
12	1	1	0.5	1	0.666667	1
13	1	1	1	1	1	1
14	1	1	0.03030303	0.060606061	0.058824	0.114286
15	1	1	0.5	1	0.666667	1
16	1	1	0.25	0.1	0.4	0.181818
17	1	1	0.058823529	0.117647059	0.111111	0.210526
18	1	1	0.1	0.2	0.181818	0.333333
19	1	1	0.03030303	0.060606061	0.058824	0.114286
20	1	1	0.03030303	0.060606061	0.058824	0.114286
21	1	1	0.5	1	0.666667	1
22	1	1	0.5	1	0.666667	1
23	1	1	0.1	0.2	0.181818	0.333333
24	1	1	0.5	1	0.666667	1
25	1	1	0.1	0.2	0.181818	0.333333
26	1	1	0.5	1	0.666667	1

	Precision		Recall		F-Measure	
Query Number	Baseline	Test Tool	Baseline	Test Tool	Baseline	Test Tool
27	1	1	0.03030303	0.060606061	0.058824	0.114286
28	1	1	0.05	0.1	0.095238	0.181818
29	1	1	0.5	1	0.666667	1
30	1	1	0.5	1	0.666667	1
31	1	1	0.166666667	0.333333333	0.285714	0.5
32	1	1	0.03030303	0.060606061	0.058824	0.114286
33	1	1	0.03030303	0.060606061	0.058824	0.114286
34	1	1	0.055555556	0.111111111	0.105263	0.2
35	1	1	0.5	1	0.666667	1
36	1	1	0.5	1	0.666667	1
37	1	1	0.5	1	0.666667	1
38	1	1	0.058823529	0.117647059	0.111111	0.210526
39	1	1	0.5	1	0.666667	1
40	1	1	0.5	1	0.666667	1
41	1	1	0.5	1	0.666667	1
42	1	1	0.5	1	0.666667	1
43	1	1	0.03030303	0.060606061	0.058824	0.114286
44	1	1	0.5	1	0.666667	1
45	1	1	0.5	1	0.666667	1
46	1	1	0.05	0.1	0.095238	0.181818
47	1	1	0.5	1	0.666667	1
48	1	1	0.5	1	0.666667	1
49	1	1	0.0625	0.125	0.117647	0.222222
50	1	1	1	1	1	1
51	1	1	0.03030303	0.060606061	0.058824	0.114286
52	1	1	0.5	1	0.666667	1
53	1	1	0.05	0.1	0.095238	0.181818
54	0	1	0	1	0	1
55	1	1	0.03030303	0.060606061	0.058824	0.114286
56	1	1	0.5	1	0.666667	1
57	1	1	0.055555556	0.111111111	0.105263	0.2
58	1	1	0.5	1	0.666667	1
59	1	1	0.5	1	0.666667	1
60	1	1	0.058823529	0.117647059	0.111111	0.210526
61	1	1	0.058823529	0.117647059	0.111111	0.210526
62	1	1	0.5	1	0.666667	1
63	1	1	0.5	1	0.666667	1
64	1	1	0.03030303	0.060606061	0.058824	0.114286

	Precision		Recall		F-Measure	
Query Number	Baseline	Test Tool	Baseline	Test Tool	Baseline	Test Tool
65	1	1	0.5	1	0.666667	1
66	1	1	0.5	1	0.666667	1
67	1	1	0.166666667	0.333333333	0.285714	0.5
68	1	1	0.5	1	0.666667	1
69	1	1	0.5	1	0.666667	1
70	1	1	0.5	1	0.666667	1
71	1	1	0.5	1	0.666667	1
72	1	1	0.5	1	0.666667	1
73	1	1	0.03030303	0.060606061	0.058824	0.114286
74	1	1	0.03030303	0.060606061	0.058824	0.114286
75	1	1	0.05	0.1	0.095238	0.181818
76	1	1	0.05	0.1	0.095238	0.181818
77	1	1	0.5	1	0.666667	1
78	1	1	0.5	1	0.666667	1
79	1	1	0.058823529	0.117647059	0.111111	0.210526
80	1	1	0.5	1	0.666667	1
81	1	1	0.5	1	0.666667	1
82	1	1	0.142857143	0.285714286	0.25	0.444444
83	1	1	0.5	1	0.666667	1
84	1	1	0.5	1	0.666667	1
85	1	1	0.5	1	0.666667	1
86	1	1	0.5	1	0.666667	1
87	1	1	0.5	1	0.666667	1
88	1	1	0.5	1	0.666667	1
89	1	1	0.5	1	0.666667	1
90	1	1	0.5	1	0.666667	1
91	1	1	0.1	0.2	0.181818	0.333333
92	1	1	0.03030303	0.060606061	0.058824	0.114286
93	1	1	0.055555556	0.111111111	0.105263	0.2
94	1	1	0.2	0.4	0.333333	0.571429
95	1	1	0.5	1	0.666667	1
96	1	1	0.5	1	0.666667	1
97	0	1	0	0.117647059	0	0.210526
98	1	1	0.5	1	0.666667	1
99	1	1	0.5	1	0.666667	1
100	1	1	0.142857143	0.285714286	0.25	0.444444
101	1	1	0.25	0.5	0.4	0.666667
102	0	1	0	1	0	1

	Precision		Recall		F-Measure	
Query Number	Baseline	Test Tool	Baseline	Test Tool	Baseline	Test Tool
103	1	1	0.5	1	0.666667	1
104	1	1	0.5	1	0.666667	1
105	1	1	0.058823529	0.117647059	0.111111	0.210526
106	1	1	0.5	1	0.666667	1
107	1	1	0.055555556	0.111111111	0.105263	0.2
108	1	1	0.5	1	0.666667	1
109	1	1	0.333333333	0.666666667	0.5	0.8
110	1	1	0.5	1	0.666667	1
111	1	1	0.1	0.2	0.181818	0.333333
112	1	1	0.1	0.2	0.181818	0.333333
113	1	1	0.03030303	0.060606061	0.058824	0.114286
114	1	1	0.5	1	0.666667	1
115	1	1	0.5	1	0.666667	1
116	1	1	0.142857143	0.285714286	0.25	0.444444
117	1	1	0.05	0.1	0.095238	0.181818
118	1	1	0.5	1	0.666667	1
119	1	1	0.5	1	0.666667	1
120	1	1	0.25	0.5	0.4	0.666667
121	1	1	0.03030303	0.060606061	0.058824	0.114286
122	1	1	0.333333333	0.666666667	0.5	0.8
123	1	1	0.5	1	0.666667	1
124	1	1	0.5	1	0.666667	1
125	1	1	0.5	1	0.666667	1
126	1	1	0.5	1	0.666667	1
127	1	1	0.03030303	0.060606061	0.058824	0.114286
128	1	1	0.5	1	0.666667	1
129	1	1	0.05	0.1	0.095238	0.181818
130	1	1	0.2	0.4	0.333333	0.571429
131	1	1	0.5	1	0.666667	1
132	1	1	0.5	1	0.666667	1
133	1	1	0.5	1	0.666667	1
134	1	1	0.1	0.2	0.181818	0.333333
135	1	1	0.05	0.1	0.095238	0.181818
136	1	1	0.5	1	0.666667	1
137	1	1	0.03030303	0.060606061	0.058824	0.114286
138	1	1	0.03030303	0.060606061	0.058824	0.114286
139	1	1	0.055555556	0.111111111	0.105263	0.2
140	1	1	0.5	1	0.666667	1

	Precision		Recall		F-Measure	
Query Number	Baseline	Test Tool	Baseline	Test Tool	Baseline	Test Tool
141	1	1	0.5	1	0.666667	1
142	1	1	0.5	1	0.666667	1
143	1	1	0.5	1	0.666667	1

Table 44 - Precision, Recall and F-Measure

APPENDIX F

TASK 1 – PROCEDURE INSTRUCTIONS

Task 1 - Procedure

You need to have the following open:

- An IE session for SPURL
- A test tool (FW) session for fragment association creation

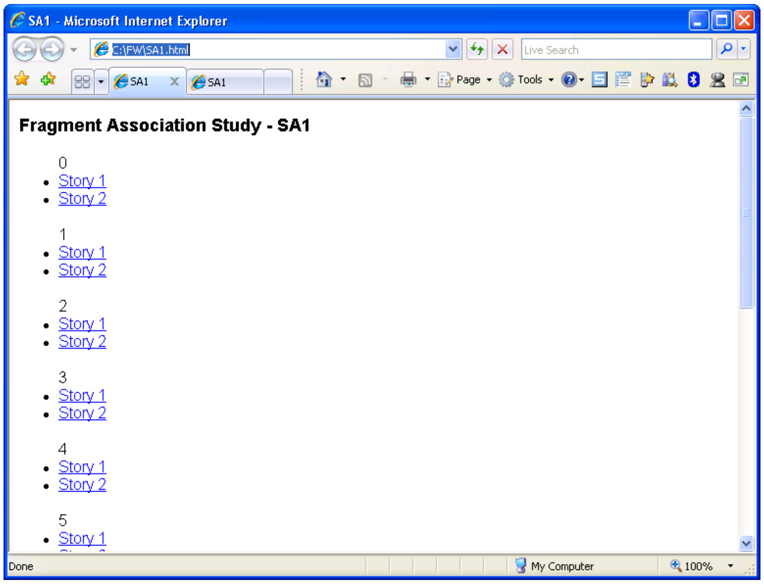
◈Repeat the following procedure 6 times, for the first 6 story pairs:

- **Scan the two Stories**

- In the SPURL IE session make sure that two tabs are open

- In the first tab click on Home  in the SPURL IE session

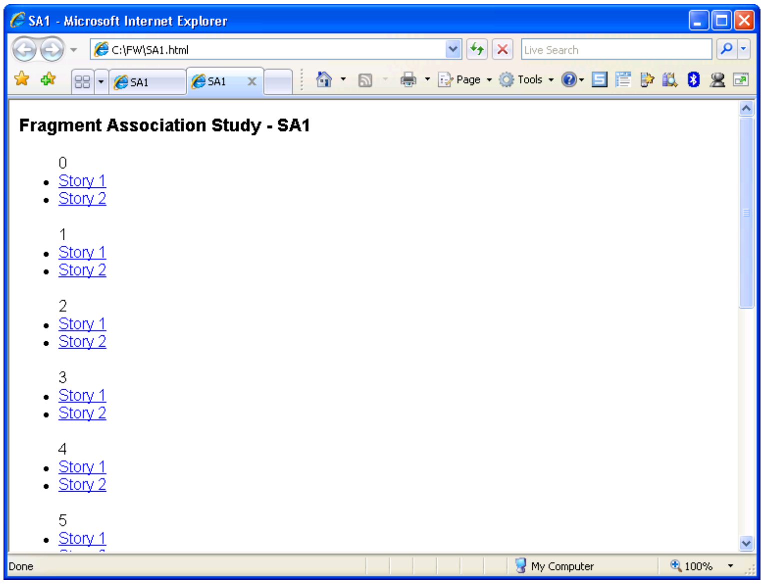
  A page with a list of link pairs is displayed:



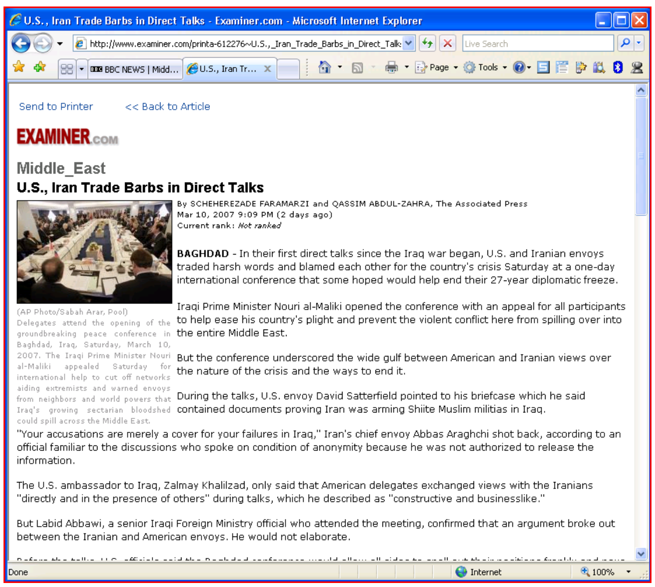
- Click on the first story. The story is displayed in the first tab:

- In the second tab click on Home  in the SPURL IE session

- The same page with the list of link pairs is displayed:



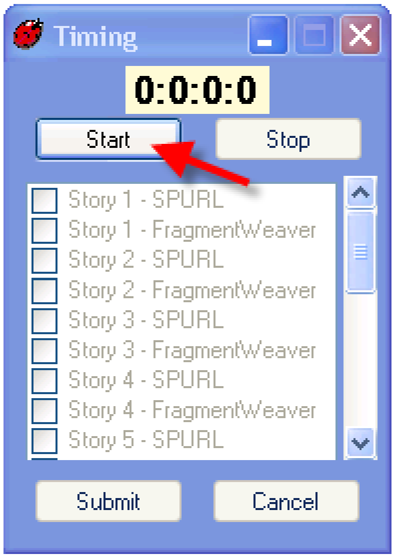
- Click on the second story. The story is displayed in the second tab:



- Scan the two stories and identify two related fragments

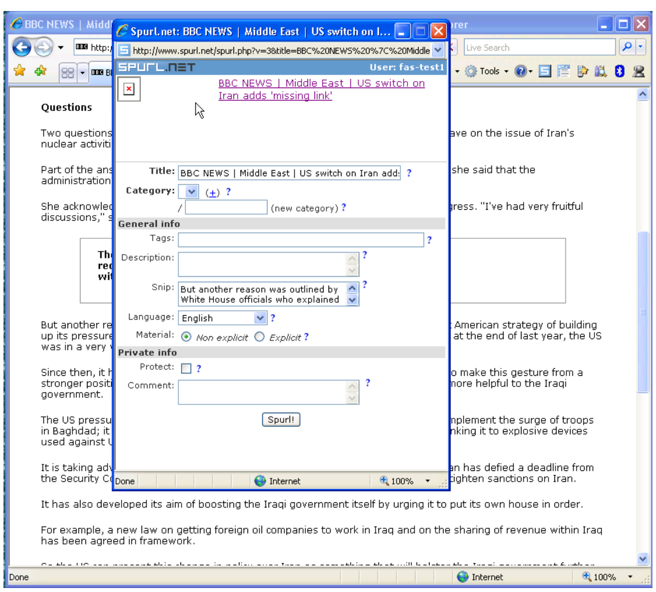
• SPURL

- Start timing ⌚ SPURL



- Highlight the first fragment and click on the SPURL button 

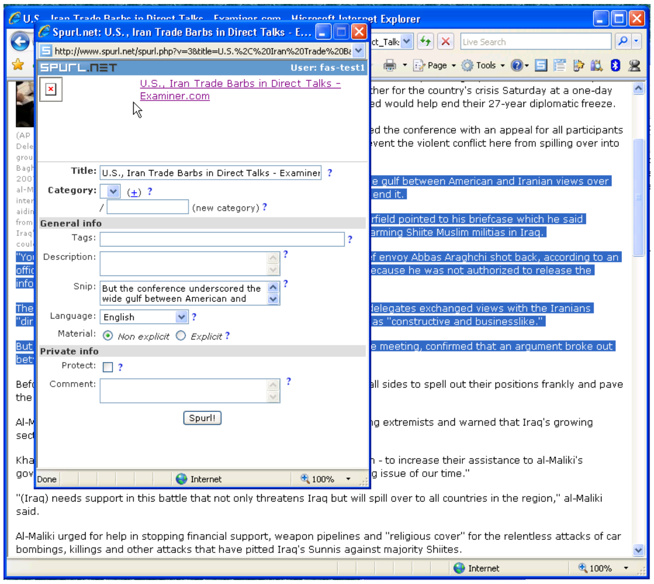
  The SPURL pop-up window will come up containing the fragment you have highlighted:



- Click on the  button to submit your fragment

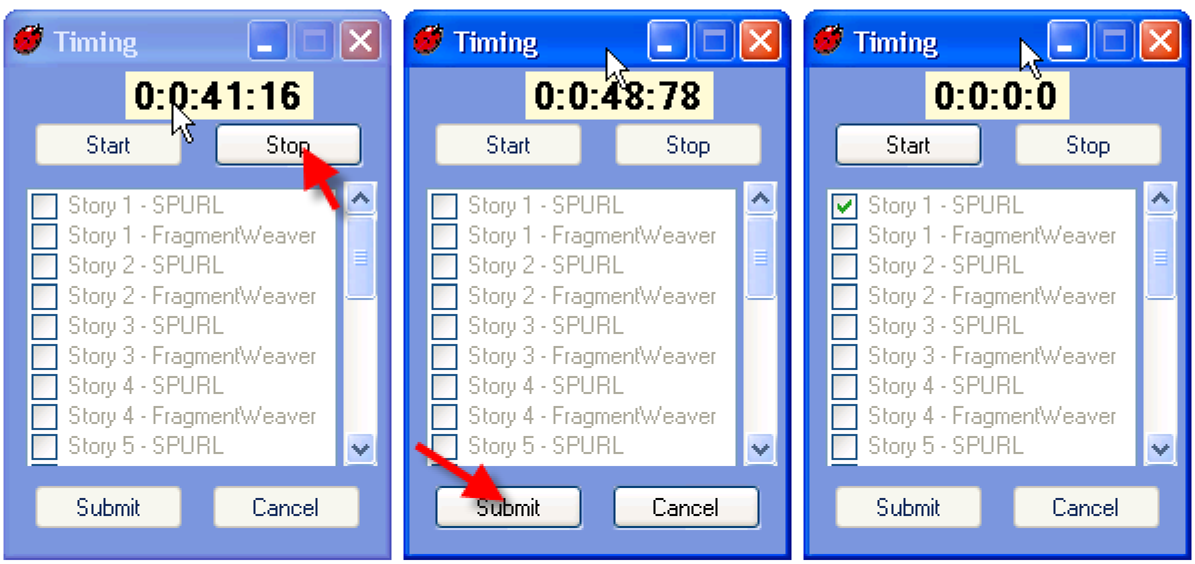
- Highlight the second fragment and click on the SPURL button 

  The SPURL pop-up window will come up containing the fragment you have highlighted:



- Click on the  button to submit your fragment

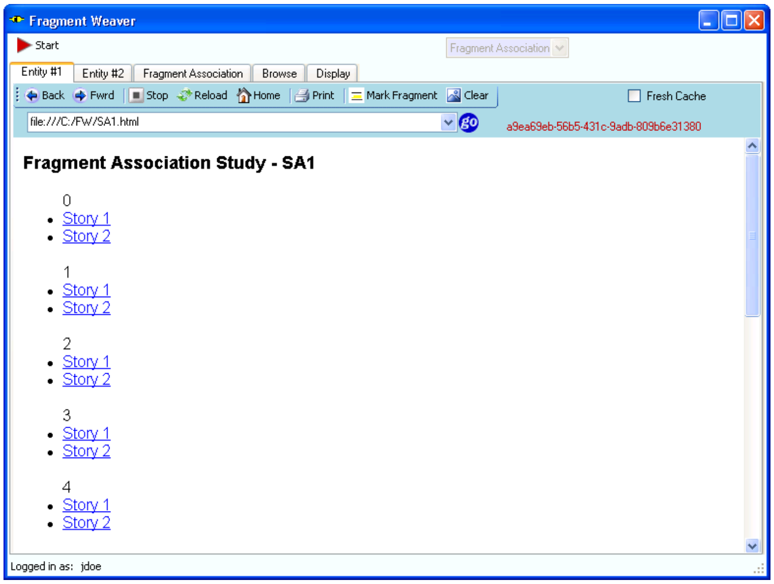
End timing ⌚ SPURL and click on "Submit". The time is recorded and the timer returns to 0



- **<u>FW</u>**

- In the first tab click on Home 

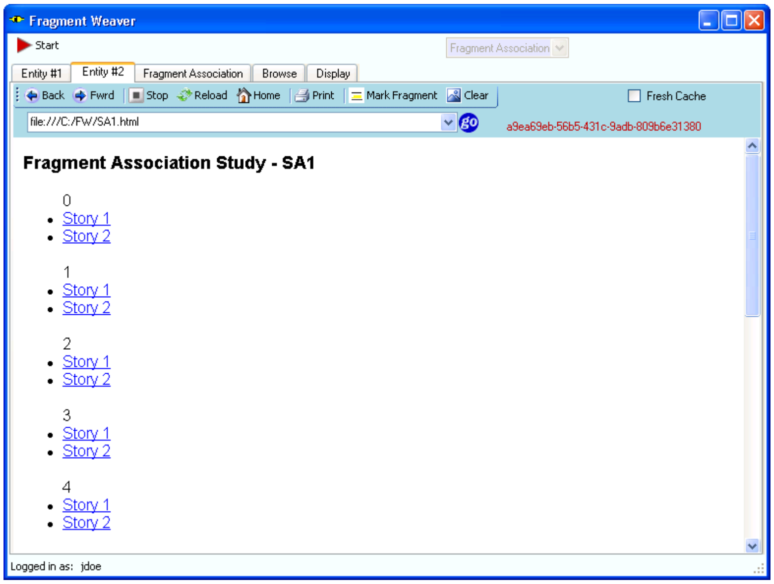
in FW
- The same page with the list of link pairs is displayed:



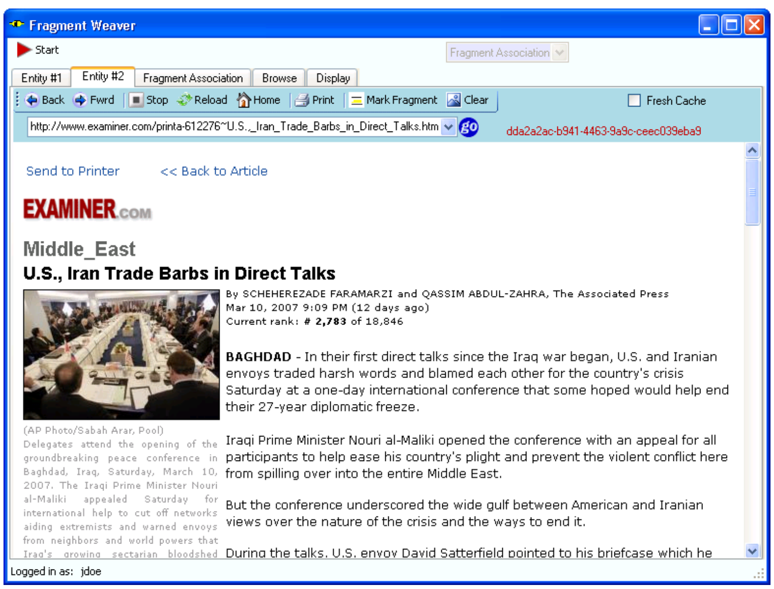
- Click on the first story. The story is displayed in the first tab.

- In the second tab click on Home  in FW

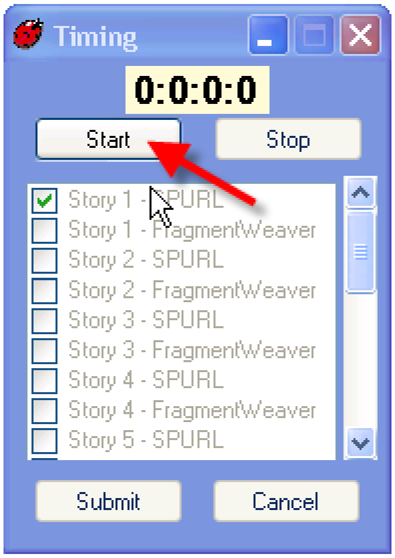
- The same page with the list of link pairs is displayed:



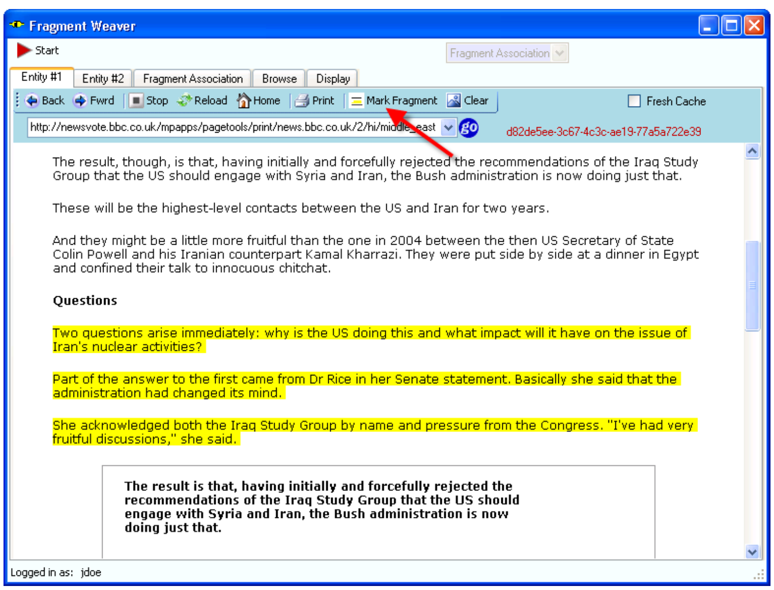
- Click on the second story. The story is displayed in the second tab.



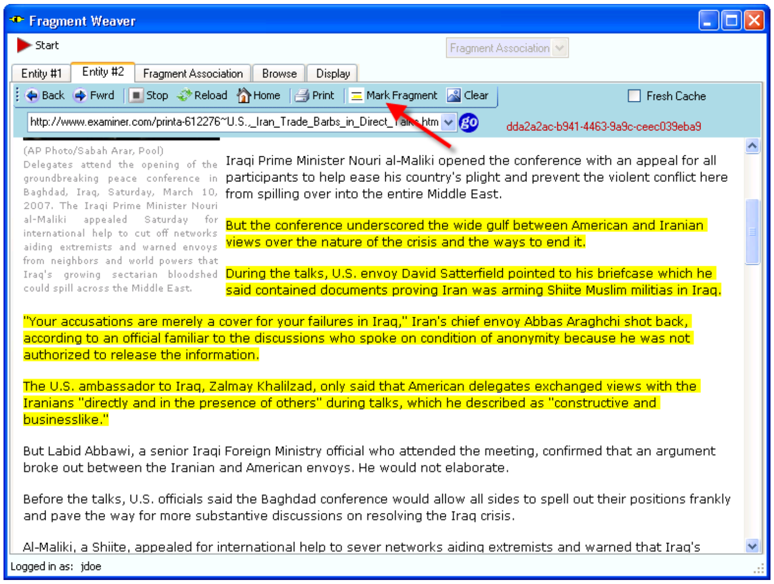
- Start timing ⌚ FW



- Highlight the first fragment and click on Mark Fragment



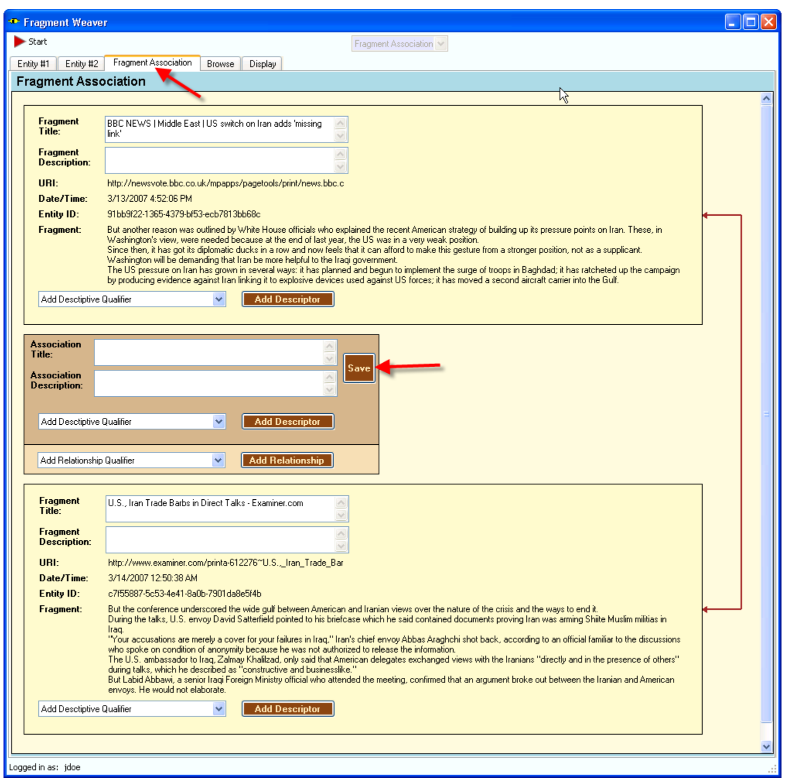
- Highlight the second fragment and click on 



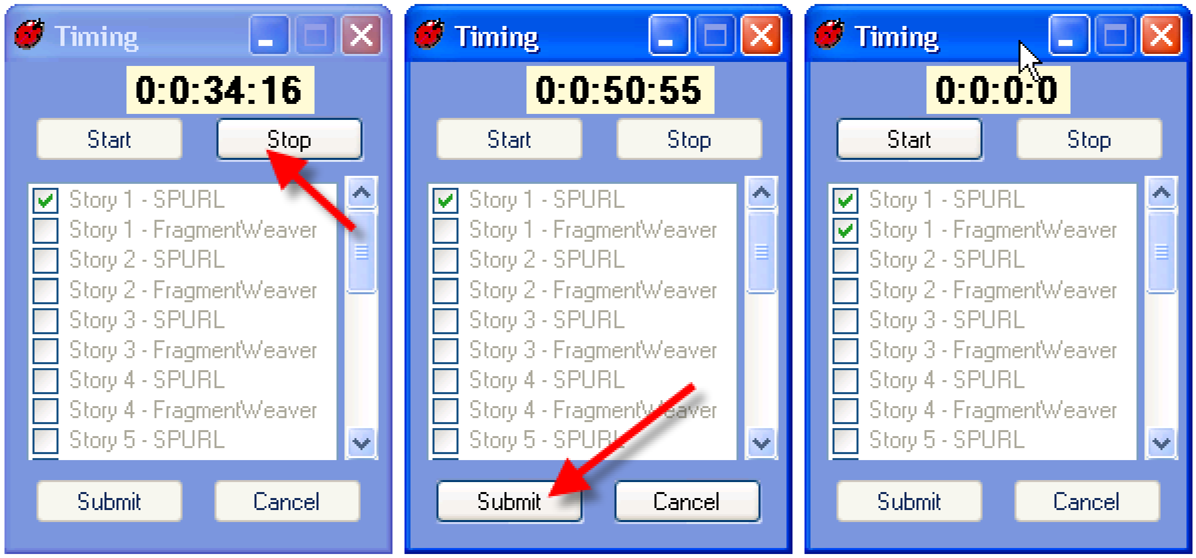
- Go to the third tab (Fragment Association)

  The two fragments are being displayed

  Click on the  button



End timing ◷ FW click on “Submit”. The time is recorded and the timer returns to 0



◈Please repeat the above procedure for 6 more stories, performing the FW section before the SPURL section

◈Please answer the following questions:

- How difficult was it to define a fragment in SPURL?

Very Difficult	❑ 0	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	**Very Easy**

- How difficult was it to define a fragment in FW?

Very Difficult	❑ 0	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	**Very Easy**

- How useful is the FW feature allowing the connection of two fragments?

Not Useful	❑ 0	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	**Very Useful**

- How helpful was SPURL in accomplishing the task of defining fragments?

Not Useful	❑ 0	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	**Very Useful**

- How helpful was FW in accomplishing the task of defining fragments?

Not Useful	❑ 0	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	**Very Useful**

- How enjoyable was using SPURL for accomplishing this task?

Not Enjoyable	❑ 0	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	**Very Enjoyable**

- How enjoyable was using FW for accomplishing this task?

Not Enjoyable	❑ 0	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	**Very Enjoyable**

APPENDIX G

TASK 2 – FIRST SUBGROUP – PROCEDURE INSTRUCTIONS

Task 2 – First Subgroup

- You will be asked to find the matching fragments for the following sets of URLs:
    - A set consisting of 6 of the stories you have bookmarked yourself
    - A set consisting of 6 stories bookmarked by others

- You will have the following open:
    - An IE session for your Task List
    - An IE session for SPURL
    - A test tool (FW) session

First Set

- Please repeat the following procedure for the first 3 URLs of the first set
    - start timing
    - use SPURL to find the fragment
        - bring up SPURL in the first tab
        - paste the URL in the SPURL "Search" box and submit search
        - can you see the fragment and the matching number?
        - identify the fragment
    - use SPURL to find the matching fragment
        - bring up SPURL in the second tab
        - use information from first fragment to search for the matching fragment
        - identify the matching fragment
    - end timing
- Please answer the following questions:
    - Using SPURL, how easy was it to find the fragments you had previously defined?

Very Difficult	❑ 0	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	**Very Easy**

    - How effective was SPURL in accomplishing the task of retrieving the fragments?

Not Useful	❑ 0	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	**Very Useful**

    - How enjoyable was using SPURL for accomplishing this task?

Not Enjoyable	❑ 0	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	**Very Enjoyable**

- Please repeat the following procedure for the remaining 3 URLs of the first set
    - start timing ⏲
    - use FW to find the fragment
        - go to the "Browse" tab
        - paste the URL in the "Search by URI' box and submit search
        - can you see the fragment and the matching number?
        - identify the fragment and matching fragment
    - end timing ⏲
- Please answer the following questions:
    - Using FW, how easy was it to find the fragments you had previously defined?

Very Difficult	❑ 0	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	**Very Easy**

    - How effective was FW in accomplishing the task of retrieving the fragments?

Not Useful	❑ 0	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	**Very Useful**

    - How enjoyable was using FW for accomplishing this task?

Not Enjoyable	❑ 0	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	**Very Enjoyable**

    - How useful is it to you to be able to see two related fragments side by side?

Not Useful	❑ 0	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	**Very Useful**

    - How useful is it to you to have the ability to navigate from one fragment to the other?

Not Useful	❑ 0	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	**Very Useful**

Second Set

- Please repeat the following procedure for the first 3 URLs of the second set
    - start timing 🕑
    - use SPURL to find the fragment
        - bring up SPURL in the first tab
        - paste the URL in the SPURL "Search" box and submit search
        - can you see the fragment and the matching number?
        - identify the fragment
    - use SPURL to find the matching fragment
        - bring up SPURL in the second tab
        - use information from first fragment to search for the matching fragment
        - identify the matching fragment
    - end timing 🕑
- Please answer the following questions:
        - Using SPURL, how easy was it to find the fragments others had defined?

Very Difficult ❑$_0$ ❑$_1$ ❑$_2$ ❑$_3$ ❑$_4$ ❑$_5$ ❑$_6$ **Very Easy**

        - How effective was SPURL in accomplishing the task of retrieving the fragments?

Not Useful ❑$_0$ ❑$_1$ ❑$_2$ ❑$_3$ ❑$_4$ ❑$_5$ ❑$_6$ **Very Useful**

        - How enjoyable was using SPURL for accomplishing this task?

Not Enjoyable ❑$_0$ ❑$_1$ ❑$_2$ ❑$_3$ ❑$_4$ ❑$_5$ ❑$_6$ **Very Enjoyable**

- Please repeat the following procedure for the remaining 3 URLs of the second set
    - start timing 🕑
    - use FW to find the fragment
        - go to the “Browse” tab
        - paste the URL in the “Search by URI’ box and submit search
        - can you see the fragment and the matching number?
        - identify the fragment and matching fragment
    - end timing 🕑
- Please answer the following questions:
    - Using FW, how easy was it to find the fragments others had defined?

Very Difficult	❑ 0	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	**Very Easy**

    - How effective was FW in accomplishing the task of retrieving the fragments?

Not Useful	❑ 0	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	**Very Useful**

    - How enjoyable was using FW for accomplishing this task?

Not Enjoyable	❑ 0	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	**Very Enjoyable**

    - How useful is it to you to be able to see two related fragments side by side?

Not Useful	❑ 0	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	**Very Useful**

    - How useful is it to you to have the ability to navigate from one fragment to the other?

Not Useful	❑ 0	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	**Very Useful**

Final Questions

- Please answer the final questions:
  - FW provides the capability of having a unique identifier for each fragment for direct reference. How useful is this to you?

Not Useful	❑0	❑1	❑2	❑3	❑4	❑5	❑6	**Very Useful**

  - What did you like about SPURL?

    ____

    ____

    ____

    ____

  - What did you dislike about SPURL?

    ____

    ____

    ____

    ____

  - What did you like about FW?

    ____

    ____

    ____

    ____

  - What did you dislike about FW?

    ____

    ____

    ____

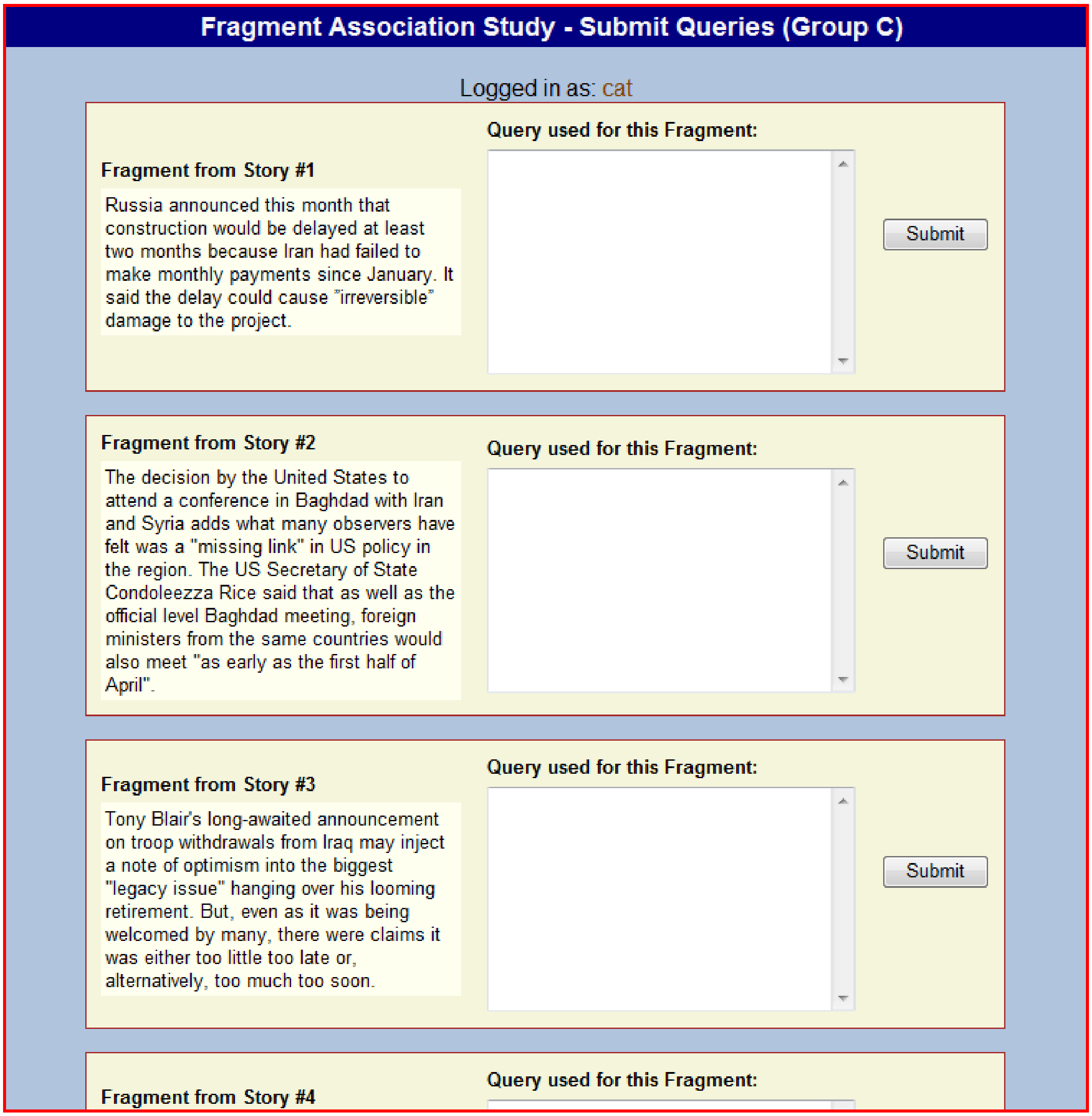
    ____

APPENDIX H

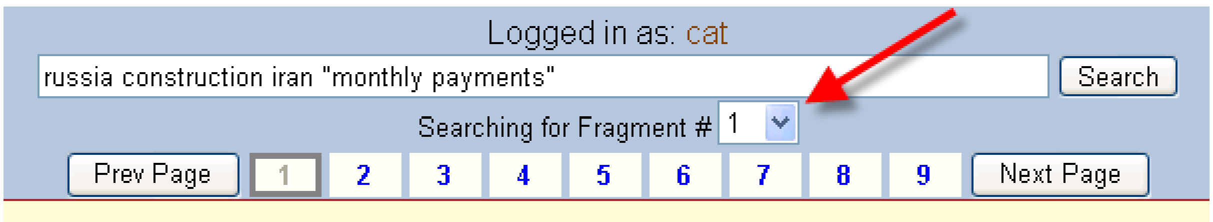
TASK 3 – PROCEDURE INSTRUCTIONS

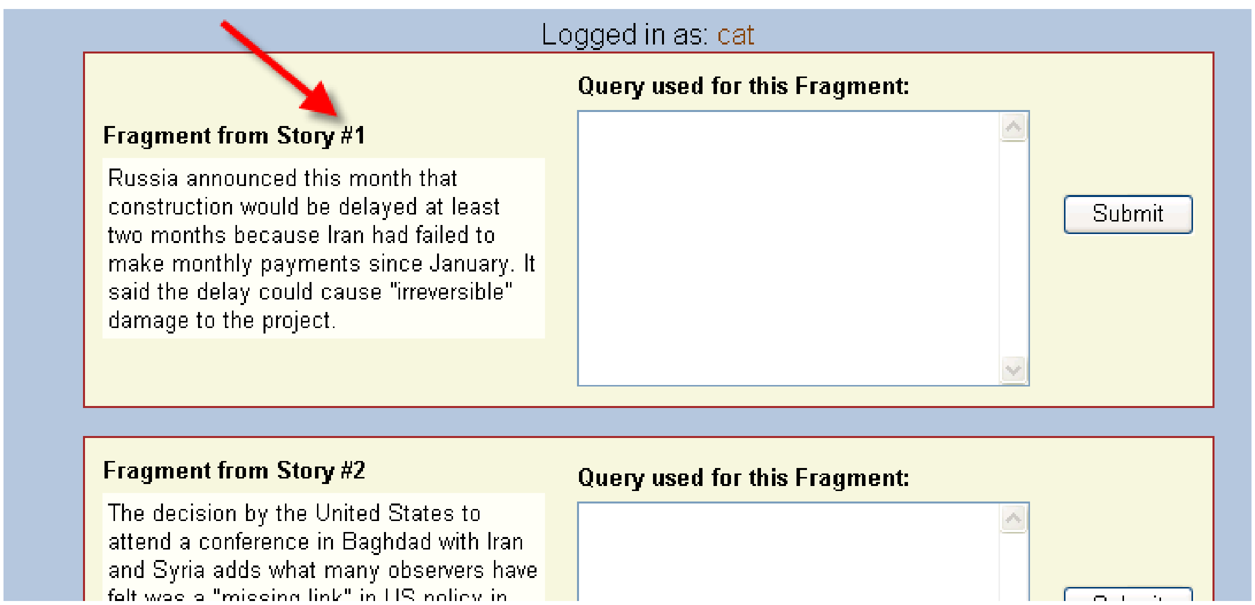
<u>Fragment Association Study -Searching</u>

- Go to http://fastudy.demetrios.info/FAStudySubmitQueries
- Log in with the username and password provided
- You will see a listing of fragments from news stories



- Go to http://fastudy.demetrios.info/FAStudySearch
- Log in with the same username and password
- Using the information in the fragment, devise a search query which successfully retrieves a story containing the entire fragment
- Please make sure to specify the number of the fragment for which you are searching in the search interface.





- Please note that the same news story may be retrieved from several news sites. Some sites may have variants of the story. All you need to do is to ensure that the fragment is contained in one of the retrieved stories exactly as quoted in the Query Submission page. In this example, we found an exact match:

Fragment from Story #1

Russia announced this month that construction would be delayed at least two months because Iran had failed to make monthly payments since January. It said the delay could cause "irreversible" damage to the project.

Nuclear plant travails strengthen Iran's resolve

Trouble with international partners used to justify domestic efforts

Associated Press

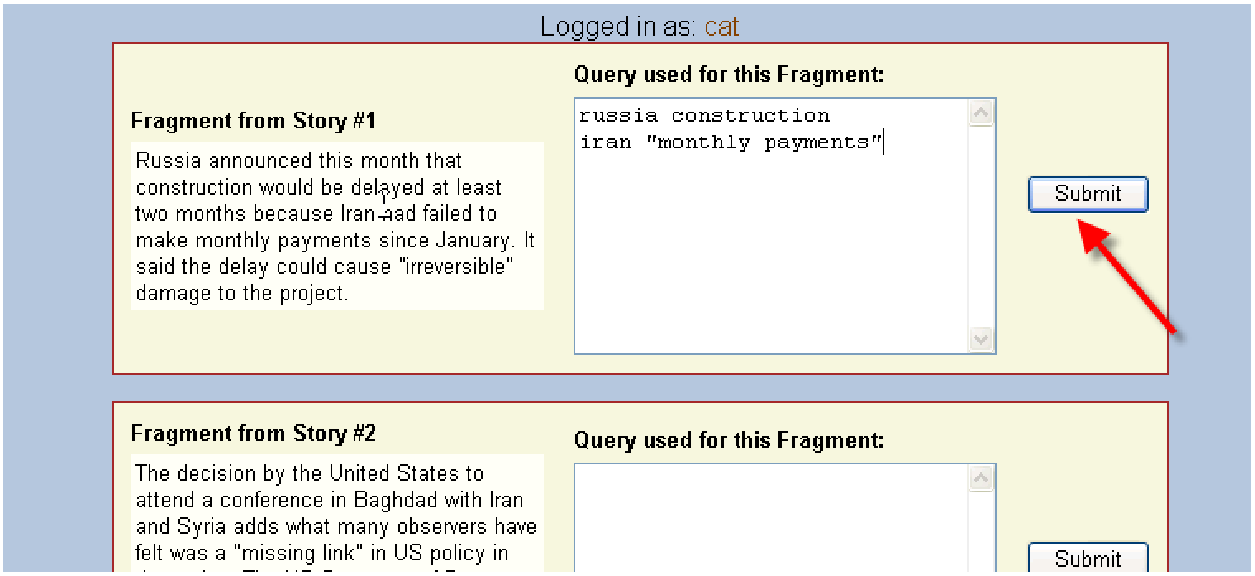
Updated: 10:04 p.m. ET March 17, 2007

BUSHEHR, Iran - For Iranians across the political spectrum, delays in construction of the country's first nuclear reactor have strengthened their resolve to master their own nuclear technology and resist U.N. efforts to stop them.

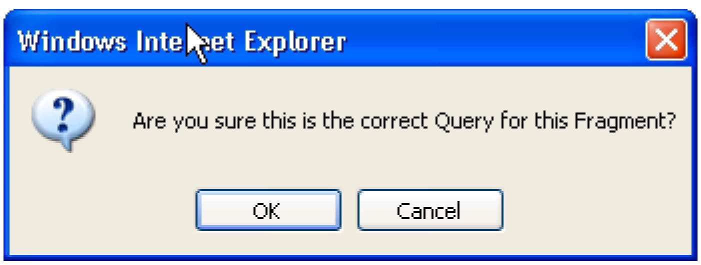
The reactor, already eight years behind schedule, is snagged on what Iran calls a politically motivated business dispute with longtime ally Russia.

Russia announced this month that construction would be delayed at least two months because Iran had failed to make monthly payments since January. It said the delay could cause "irreversible" damage to the project.

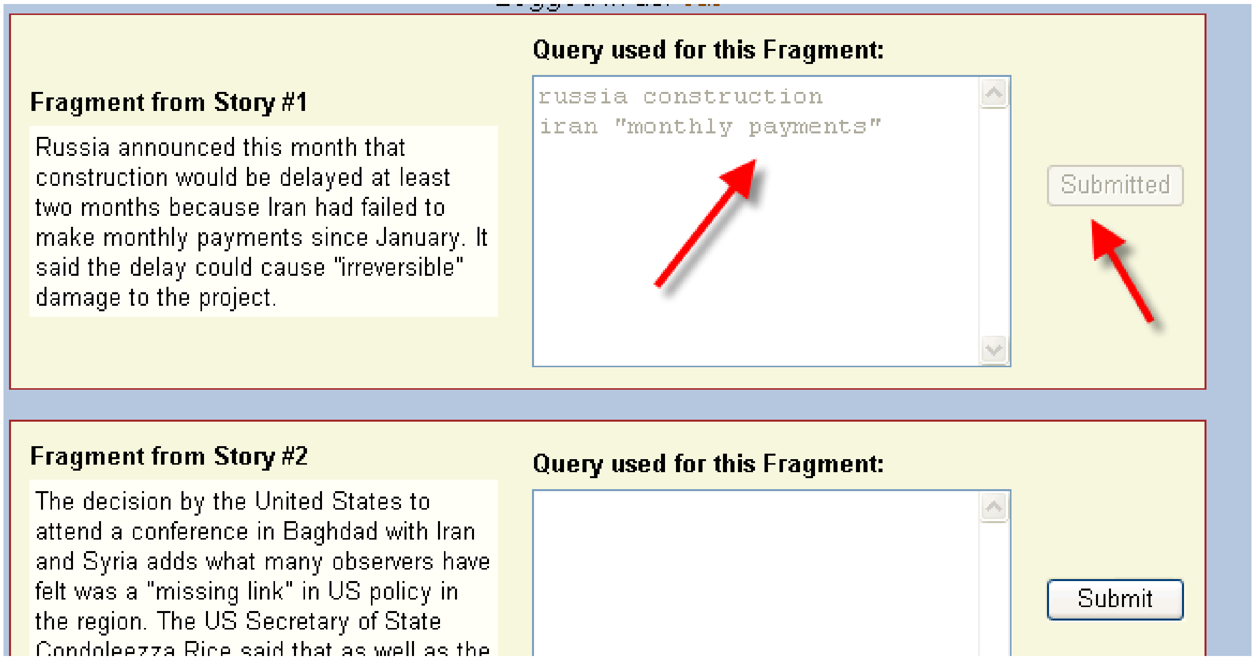
- Once you have successfully found a story you can copy the successful query string, paste it in the appropriate box in the Query Submission page (http://fastudy.demetrios.info/FAStudySubmitQueries) and submit it.



- You will be prompted. If you are sure that this is the query you would like to submit you can press "OK"



- Once you submitted your query, the display will change specifying that the query has been submitted


Fragment from Story #1
Russia announced this month that construction would be delayed at least two months because Iran had failed to make monthly payments since January. It said the delay could cause "irreversible" damage to the project.
Query used for this Fragment:
russia construction
iran "monthly payments"
Submitted
Fragment from Story #2
The decision by the United States to attend a conference in Baghdad with Iran and Syria adds what many observers have felt was a "missing link" in US policy in the region. The US Secretary of State
Query used for this Fragment:
Submit

- The following are techniques useful for searching:

Symbol	Function	Example
+	Finds pages that contain ALL the terms preceded by the + symbol, and allows inclusion of terms that are usually ignored	cat +in +the hat
"	Finds the exact words as quoted	"cirque du soleil"
( )	Finds or excludes pages that contain a group of words	(mobile device) OR (wireless device) (mobile device) NOT(wireless device)
AND	Finds pages that contain ALL the terms or phrases	dog AND cat
NOT	Excludes pages that contain a term or phrase	(mobile device) NOT (wireless device)
OR	Finds pages that contain either term or phrase	(mobile device) OR (wireless device)

Notes

- Stop words, as well as all punctuation marks (except for the symbols noted above), are ignored unless surrounded by quotation marks or preceded by the **+** symbol.
- Only the first 10 terms are used in obtaining search results.
- Search words for basic searches aren't case-sensitive.
- You don't have to type the word **AND** between your search words. By default, all searches are **AND** searches.
- You can type up to 150 characters in the search box, including spaces.

Since the searches performed through this interface are recorded in support of this specific study, please use this search interface only for searches related to this study.

BIBLIOGRAPHY

[1] J. Marsh, "XML Base W3C Recommendation," June 27, 2001 2001. http://www.w3.org/TR/xmlbase/.

[2] S. DeRose, E. Maler, and D. Orchard, "XML Linking Language (XLink) Version 1.0 W3C Recommendation " World Wide Web Consortium June 27, 2001 2001. http://www.w3.org/TR/xlink/.

[3] P. Grosso, E. Maler, J. Marsh, and N. Walsh, "XPointer Framework W3C Recommendation," World Wide Web Consortium March 25, 2003 2003. http://www.w3.org/TR/xptr-framework/.

[4] P. Ciancarini, F. Folli, D. Rossi, and F. Vitali, "XLinkProxy: external linkbases with XLink," in *2002 ACM symposium on Document engineering*, McLean, Virginia, USA, 2002, pp. 57 - 65. http://doi.acm.org/10.1145/585058.585070.

[5] T. H. Nelson, *Literary Machines: The Report on, and of, Project Xanadu Concerning Word Processing, Electronic Publishing, Hypertext, Thinkertoys, Tomorrow's Intellectual Revolution, and Certain Other Topics, Including Knowledge, Education and Freedom*, 87.1 ed.: Published by the Author, 1987.

[6] P. Grosso, E. Maler, J. Marsh, and N. Walsh, "XPointer Framework W3C Recommendation," World Wide Web Consortium March 25, 2003. http://www.w3.org/TR/xptr-framework/.

[7] J. Marsh, "XML Base W3C Recommendation," World Wide Web Consortium June 27, 2001. http://www.w3.org/TR/xmlbase/.

[8] S. DeRose, E. Maler, and D. Orchard, "XML Linking Language (XLink) Version 1.0 W3C Recommendation " World Wide Web Consortium June 27, 2001. http://www.w3.org/TR/xlink/.

[9] F. M. I. Shipman and R. McCall, "Incremental formalization with the hyper-object substrate," *ACM Transactions on Information Systems (TOIS)*, vol. 17, pp. 199 - 227, April 1999 1999. http://doi.acm.org/10.1145/306686.306690.

[10] F. G. Halasz, "Reflections on NoteCards: seven issues for the next generation of hypermedia systems," *Communications of the ACM*, vol. 31, pp. 836 - 852, July 1988 1988. http://doi.acm.org/10.1145/48511.48514.

[11] F. Miksa, "The Universe of Information Entities (Chapter 1)," in *Information Entity Access Control and Information Retrieval*, 2000. http://www.gslis.utexas.edu/%7Eitesm/readings/CIES-c1-000801.pdf.

[12] L. Schamber, "What is a document? Rethinking the concept in uneasy times," *Journal of the American Society for Information Science*, vol. 47, pp. 669-671, 1996. http://www3.interscience.wiley.com/cgi-bin/abstract/57816/ABSTRACT.

[13] M. K. Buckland, "What is a document?," *Journal of the American Society for Information Science*, vol. 48, pp. 804 - 809, September 1997 1997.
[14] "Prof. Nelson Talk Analyzes P.R.I.D.E.," in *Vassar Miscennany News* Poughkeepsie, NY, 1965, pp. 3-4.
[15] T. H. Nelson, "Xanalogical Structure, Needed Now More than Ever: Parallel Documents, Deep Links to Content, Deep Versioning, and Deep Re-Use," *ACM Computing Surveys*, vol. 31, p. Theodor Holm Nelson, December 1999 2000.
[16] T. H. Nelson, "The heart of connection: hypermedia unified by transclusion," *Communications of the ACM*, vol. 38, pp. 31-35, August 1995 1995. http://doi.acm.org/10.1145/208344.208353.
[17] M. Kay, "XSL Transformations (XSLT) Version 2.0: W3C Candidate Recommendation " World Wide Web Consortium November 3, 2005. http://www.w3.org/TR/xslt20/.
[18] S. Adler, A. Berglund, J. Caruso, S. Deach, T. Graham, P. Grosso, E. Gutentag, A. Milowski, S. Parnell, J. Richman, and S. Zilles, "Extensible Stylesheet Language (XSL) Version 1.0: W3C Recommendation," World Wide Web Consortium October 15, 2001. http://www.w3.org/TR/xsl/
[19] S. Boag, D. Chamberlin, M. F. Fernández, D. Florescu, J. Robie, and J. Siméon, "XQuery 1.0: An XML Query Language," World Wide Web Consortium January 23, 2007 2007. http://www.w3.org/TR/xquery/
[20] O. Lassila and R. R. Swick, "Resource Description Framework (RDF) Model and Syntax Specification," World Wide Web Consortium, 1999. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/
[21] J. Conklin, "Hypertext: an introduction and survey," *Computer*, vol. 20, pp. 17 - 41, September 1987 1987.
[22] R. Trigg, "A Network-Based Approach to Text Handling for the Online Scientific Comm," in *Department of Computer Science*. vol. PhD College Park MD 20742: University of Maryland, 1983.
[23] S. J. DeRose, "Expanding the notion of links," in *Proceedings of the second annual ACM conference on Hypertext*, Pittsburgh, Pennsylvania, United States, 1989, pp. 249-257. http://doi.acm.org/10.1145/74224.74245.
[24] H. C. Davis, "Referential integrity of links in open hypermedia systems," in *Ninth ACM conference on Hypertext and hypermedia : links, objects, time and space---structure in hypermedia systems: links, objects, time and space---structure in hypermedia systems*, Pittsburgh, Pennsylvania, 1998, pp. 207 - 216.
[25] K. Grønbæk and R. H. Trigg, "Toward a Dexter-based model for open hypermedia: unifying embedded references and link objects," in *Proceedings of the the seventh ACM conference on Hypertext*, Bethesda, Maryland, United States, 1996, pp. 149-160. http://doi.acm.org/10.1145/234828.234843.
[26] F. Halasz and M. Schwartz, "The Dexter hypertext reference model," *Communications of the ACM*, vol. 37, pp. 30 - 39, 1994. http://doi.acm.org/10.1145/175235.175237.
[27] R. B. Neto, C. A. Izeki, P. Maria da Graca, R. P. Fortes, and K. N. Truong, "An open linking service supporting the authoring of web documents," in *Proceedings of the 2002 ACM symposium on Document engineering*, McLean, Virginia, USA, 2002, pp. 66-73. http://doi.acm.org/10.1145/585058.585071.

[28] R. W. Kopak, "Functional link typing in hypertext," *ACM Computing Surveys (CSUR),* vol. 31, p. Article No. 16, December 1999 1993. http://doi.acm.org/10.1145/345966.346005.
[29] J. Nanard and M. Nanard, "Should anchors be typed too?: an experiment with MacWeb," in *Fifth ACM Conference on Hypertext and Hypermedia* Seattle, Washington, USA, 1993, pp. 51 - 62
[30] R. Conradi and B. Westfechtel, "Version models for software configuration management," *ACM Computing Surveys,* vol. 30, pp. 232-282, 1998. http://doi.acm.org/10.1145/280277.280280
[31] E. J. Whitehead, Jr, "An analysis of the hypertext versioning domain," in *Department of Computer and Information Science* Irvive, CA: University of California, Irvine, 2000, p. 242.
[32] A. Haake and D. Hicks, "VerSE: towards hypertext versioning styles," in *Seventh ACM conference on Hypertext*, Bethesda, Maryland, United States, 1996, pp. 224-234. http://doi.acm.org/10.1145/234828.234850.
[33] F. Vitali, "Versioning hypermedia," *ACM Computing Surveys (CSUR),* vol. 31, pp. 1-7, 1999. http://doi.acm.org/10.1145/345966.346019.
[34] S. Reich, U. K. Wiil, P. J. Nuernberg, H. C. Davis, K. Groenbaek, K. M. Anderson, D. E. Millard, J. M. Haake, and K. Groenbaek, "Addressing Interoperability in Open Hypermedia: the Design of the Open Hypermedia Protocol," *The New Review of Hypermedia and Multimedia* vol. 5, pp. pp. 207-248, 2000.
[35] F. M. Shipman, III and C. C. Marshall, "Spatial hypertext: an alternative to navigational and semantic links," *ACM Computing Surveys (CSUR),* vol. 31, pp. 1-5, December 1999 1999.
[36] C. C. Marshall and I. Frank M. Shipman, "Spatial hypertext: designing for change," *Communications of the ACM*, vol. 38, pp. 88 - 97, 1995. http://doi.acm.org/10.1145/208344.208350.
[37] P. J. Nurnberg, J. J. Leggett, and U. K. Wiil, "An agenda for open hypermedia research," in *Proceedings of the ninth ACM conference on Hypertext and hypermedia : links, objects, time and space---structure in hypermedia systems: links, objects, time and space---structure in hypermedia systems*, Pittsburgh, Pennsylvania, United States, 1998, pp. 198-206. http://doi.acm.org/10.1145/276627.276649.
[38] U. K. Wiil and J. J. Leggett, "The HyperDisco approach to open hypermedia systems," in *Proceedings of the the seventh ACM conference on Hypertext*, Bethesda, Maryland, United States, 1996, pp. 140-148. http://doi.acm.org/10.1145/234828.234842.
[39] L. Carr, W. Hall, H. Davis, and R. Hollom, "The Microcosm Link Service and its Application to the World Wide Web (1994) " in *First World-Wide Web Conference*, CERN, 1994.
[40] W. Hall, G. Hill, and H. Davis, "The microcosm link service," in *Proceedings of the fifth ACM conference on Hypertext*, Seattle, Washington, United States, 1993, pp. 256-259. http://doi.acm.org/10.1145/168750.168842.
[41] H. Davis, W. Hall, I. Heath, G. Hill, and R. Wilkins, "Towards an integrated information environment with open hypermedia systems," in *Proceedings of the ACM conference on Hypertext* Milan, Italy: ACM Press, 1992. http://doi.acm.org/10.1145/168466.168522.

[42] A. M. Fountain, W. Hall, I. Heath, and H. C. Davis, "MICROCOSM: an open model for hypermedia with dynamic linking," in *Hypertext: concepts, systems and applications*: Cambridge University Press, 1992, pp. 298-311.
[43] K. Grønbæk, J. A. Hem, O. L. Madsen, and L. Sloth, "Designing Dexter-based cooperative hypermedia systems," in *Proceedings of the fifth ACM conference on Hypertext* Seattle, Washington, United States: ACM Press, 1993. http://doi.acm.org/10.1145/168750.168764.
[44] K. Grønbæk, N. O. Bouvin, and L. Sloth, "Designing Dexter-based hypermedia services for the World Wide Web," in *Proceedings of the eighth ACM conference on Hypertext* Southampton, United Kingdom: ACM Press, 1997. http://doi.acm.org/10.1145/267437.267453.
[45] K. Grønbæk and R. H. Trigg, *From Web to Workplace : Designing Open Hypermedia Systems*. Cambridge, Mass.: MIT Press, 1999.
[46] K. M. Anderson, R. N. Taylor, and J. E. James Whitehead, "Chimera: hypertext for heterogeneous software environments," in *Proceedings of the 1994 ACM European conference on Hypermedia technology* Edinburgh, Scotland: ACM Press, 1994.
[47] K. M. Anderson, "Integrating open hypermedia systems with the World Wide Web," in *Proceedings of the eighth ACM conference on Hypertext* Southampton, United Kingdom: ACM Press, 1997. http://doi.acm.org/10.1145/267437.267454.
[48] K. Andrews, F. Kappe, and H. Maurer, "Serving information to the Web with Hyper-G," *Computer Networks and ISDN Systems*, vol. 27, pp. 919-926, 1995.
[49] "Hyperwave." http://www.hyperwave.com.
[50] "Hyperwave Functions," Sourceforge. http://quanta.sourceforge.net/docs/php/ref.hyperwave.html.
[51] L. A. Carr, D. De Roure, W. Hall, and G. Hill, "Implementing an open link service for the World Wide Web," *World Wide Web*, vol. 1, pp. 61-71, 1998. http://dx.doi.org/10.1023/A:1019251328413
[52] K. Grønbæk, L. Sloth, and P. Ørbæk, "Webvise: browser and proxy support for open hypermedia structuring mechanisms on the World Wide Web," in *Proceeding of the eighth international conference on World Wide Web* Toronto, Canada: Elsevier North-Holland, Inc., 1999.
[53] K. Grønbæk and R. H. Trigg, "Design issues for a Dexter-based hypermedia system," in *Proceedings of the ACM conference on Hypertext* Milan, Italy: ACM Press, 1992. http://doi.acm.org/10.1145/168466.168525.
[54] A. L. Hors, P. L. Hégaret, L. Wood, G. Nicol, J. Robie, M. Champion, and S. Byrne, "Document Object Model (DOM) Level 3 Core Specification," World Wide Web Consortium April 07, 2004 2004. http://www.w3.org/TR/DOM-Level-3-Core
[55] N. O. Bouvin, "Augmenting the web through open hypermedia," *The New Review of Hypermedia and Multimedia*, vol. 8, pp. 3-25, 2003.
[56] P. T. Zellweger, N. O. Bouvin, H. Jeh, and J. D. Mackinlay, "Fluid annotations in an open world," in *Proceedings of the twelfth ACM conference on Hypertext and Hypermedia* Århus, Denmark: ACM Press, 2001. http://doi.acm.org/10.1145/504216.504224.
[57] J. Kahan and M.-R. Koivunen, "Annotea: An Open RDF Infrastructure for Shared Web Annotations," in *Proceedings of the 10th international conference on World Wide Web* Hong Kong, 2001, pp. 623-632. http://www10.org/cdrom/papers/488/index.html.

[58] M.-R. Koivunen, "Annotea shared bookmarks: Semantic Web at your fingertips," in *International Semantic Web Conference*, Hiroshima, Japan, 2004. http://www.annotea.org/ISWC2004/annoteademo.html.

[59] D. Beckett and B. McBride, "RDF/XML Syntax Specification (Revised) - W3C Recommendation," World Wide Web Consortium, 2004. http://www.w3.org/TR/rdf-syntax-grammar/

[60] N. Karousos, I. Pandis, S. Reich, and M. Tzagarakis, "Offering open hypermedia services to the WWW: a step-by-step approach for developers," in *Proceedings of the 12th international conference on World Wide Web*, Budapest, Hungary, 2003, pp. 482-489. http://doi.acm.org/10.1145/775152.775221.

[61] R. Chinnici, J.-J. Moreau, A. Ryman, and S. Weerawarana, "Web Services Description Language (WSDL) Version 2.0 Part 1: Core Language - W3C Recommendation," World Wide Web Consortium, 2007. http://www.w3.org/TR/wsdl20.

[62] F. Curbera, M. Duftler, R. Khalaf, W. Nagy, N. Mukhi, and S. Weerawarana, "Unraveling the Web Services Web: An Introduction to SOAP, WSDL, and UDDI," *IEEE Internet Computing*, vol. 6, pp. 86 - 93, March 2002 2002.

[63] L. Clément, A. Hately, C. v. Riegen, and T. Rogers, "UDDI Spec Technical Committee Draft, Dated 20041019," OASIS October 19, 2004 2004. http://uddi.org/pubs/uddi_v3.htm.

[64] K.-P. Yee, "CritLink: Advanced Hyperlinks Enable Public Annotation on the Web," in *Computer Supported Cooperative Work (CSCW)*, New Orleans, Louisiana, USA, 2002. http://zesty.ca/crit/yee-crit-cscw2002-demo.pdf.

[65] C. C. Marshall, I. Frank M. Shipman, and J. H. Coombs, "VIKI: spatial hypertext supporting emergent structure," in *1994 ACM European conference on Hypermedia technology*, Edinburgh, Scotland, 1994, pp. 13 - 23. http://doi.acm.org/10.1145/192757.192759.

[66] C. C. Marshall and I. Frank M. Shipman, "Spatial hypertext and the practice of information triage," in *Eighth ACM conference on Hypertext and Hypermedia*, Southampton, United Kingdom, 1997, pp. 124-133. http://doi.acm.org/10.1145/267437.267451.

[67] M. N. Price, B. N. Schilit, and G. Golovchinsky, "XLibris: the active reading machine," in *CHI 98 conference summary on Human factors in computing systems* Los Angeles, California, United States 1998, pp. 22-23. http://doi.acm.org/10.1145/286498.286510.

[68] M. Margolis and D. Resnick, "Third Voice: Vox Populi Vox Dei?," *First Monday*, vol. 4, October 1999 1999. http://firstmonday.org/issues/issue4_10/margolis/index.html.

[69] C. Oakes, "The Web's New Graffiti?," *Wired*, June 9, 1999 1999. http://www.wired.com/print/science/discoveries/news/1999/06/20101.

[70] V. Wielbut, "Third Voice," *Spotlight: Online Newsletter of the Alliance for Community Technology*, June 28, 1999 1999. http://www.communitytechnology.org/newsletter/no2.html#thirdvoice.

[71] L. B. Page, Sergey; Motwani, Rajeev; Winograd, Terry, "The PageRank Citation Ranking: Bringing Order to the Web.," in *Stanford Database Group Publication Server* 1999. http://dbpubs.stanford.edu/pub/1999-66.

[72] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, pp. 604 - 632 1999. http://doi.acm.org/10.1145/324133.324140.

[73] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg, "Automatic resource compilation by analyzing hyperlink structure and associated text," vol. 30, p. 65, 1998. http://www.sciencedirect.com/science/article/B6TYT-3WRC342-2T/2/1b4358ad41d1b1c04093d212781f1f00
[74] T. H. Nelson, "Embedded Markup Considered Harmful," in *xml.com*: O'Reilly, 1997. http://www.xml.com/pub/a/w3j/s3.nelson.html.
[75] N. M. Delisle and M. D. Schwartz, "Collaborative writing with hypertext," *Professional Communication, IEEE Transactions on,* vol. 32, pp. 183-188, 1989.
[76] C. C. Marshall, "Toward an ecology of hypertext annotation," in *Ninth ACM conference on Hypertext and Hypermedia : links, objects, time and space---structure in hypermedia systems: links, objects, time and space---structure in hypermedia systems*, Pittsburgh, Pennsylvania, United States, 1998, pp. 40-49. http://doi.acm.org/10.1145/276627.276632.
[77] G. Evgeniy, D. Susan, and H. Eric, "Newsjunkie: providing personalized newsfeeds via analysis of information novelty," in *Proceedings of the 13th international conference on World Wide Web*, New York, NY, USA, 2004, pp. 482-490. http://doi.acm.org/10.1145/988672.988738.
[78] G. M. DelCorso, A. Gull, and F. Romani, "Ranking a stream of news," in *Proceedings of the 14th international conference on World Wide Web*, Chiba, Japan, 2005, pp. 97-106. http://doi.acm.org/10.1145/1060745.1060764.
[79] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," 2001. http://www.sciam.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.
[80] G. Klyne and J. J. Carroll, "Resource Description Framework (RDF): Concepts and Abstract Syntax W3C Recommendation," World Wide Web Consortium, 2004. http://www.w3.org/TR/rdf-concepts
[81] T. J. Lukka and B. Fallenstein, "Freenet-like GUIDs for implementing xanalogical hypertext," in *Thirteenth ACM conference on Hypertext and Hypermedia*, College Park, Maryland, USA, 2002, pp. 194 - 195.
[82] C. C. Marshall, "Annotation: from paper books to the digital library," in *Second ACM international conference on Digital libraries*, Philadelphia, Pennsylvania, United States, 1997, pp. 131 - 140.
[83] T. A. Phelps and R. Wilensky, "Robust intra-document locations," *Computer Networks,* vol. 33, pp. 105-118, June 2000 2000.
[84] T. A. Phelps and R. Wilensky, "Multivalent Documents."
[85] C. C. Marshall and I. Frank M. Shipman, "Searching for the missing link: discovering implicit structure in spatial hypertext," in *Fifth ACM conference on Hypertext and Hypermedia*, Seattle, Washington, United States, 1993, pp. 217 - 230. http://doi.acm.org/10.1145/168750.168826.
[86] M. M. Louwerse and H. H. Mitchell, "Toward a Taxonomy of a Set of Discourse Markers in Dialog: A Theoretical and Computational Linguistic Account " *Discourse Processes,* vol. 35, pp. 119-239, May 1, 2003 2003.
[87] A. Knott and C. Mellish, "A Feature-Based Account of the Relations Signalled by Sentence and Clause Connective," *Language and Speech,* vol. 39, pp. 143-183, 1996.
[88] M. Gudgin, M. Hadley, N. Mendelsohn, J.-J. Moreau, H. F. Nielsen, A. Karmarkar, and Y. Lafon, "SOAP Version 1.2 Part 1: Messaging Framework - W3C Recommendation," World Wide Web Consortium, 2007. http://www.w3.org/TR/soap12-part1/.

[89] M. Gudgin, M. Hadley, N. Mendelsohn, J.-J. Moreau, H. F. Nielsen, A. Karmarkar, and Y. Lafon, "SOAP Version 1.2 Part 2: Adjuncts - W3C Recommendation," World Wide Web Consortium, 2007. http://www.w3.org/TR/soap12-part2/.
[90] L. Grossman, "Time's Person of the Year: You," in *Time*, 2006. http://www.time.com/time/magazine/article/0,9171,1569514,00.html.
[91] IDF, "The DOI System," The International DOI Foundation (IDF). http://www.doi.org.
[92] P. Leach, M. Mealling, and R. Salz, "A Universally Unique IDentifier (UUID) URN Namespace," The Internet Engineering Task Force. Network Working Group. July 2005 2005. http://www.ietf.org/rfc/rfc4122.txt.
[93] National Information Standards Organization, "Syntax for the Digital Object Identifier," September 30, 2005. http://www.niso.org/standards/resources/Z39-84-2005.pdf?CFID=14430330&CFTOKEN=55341173.
[94] N. I. S. Organization, "Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies," July 25, 2005 2005. http://www.niso.org/standards/resources/Z39-19-2005.pdf?CFID=14430330&CFTOKEN=55341173.
[95] "North American Industry Classification System," United States. Department of Commerce. Bureau of the Census, 2002. http://www.census.gov/epcd/naics02/.
[96] "Thesaurus of ERIC Descriptors," United States. Department of Education. Institute of Education Sciences. Education Resources Information Center. http://www.eric.ed.gov.
[97] "UNESCO Thesaurus," United Nations Educational, Scientific and Cultural Organisation http://databases.unesco.org/thesaurus/.
[98] "CardSpace." http://msdn.microsoft.com/winfx/reference/infocard/default.aspx.
[99] G. L. Rein, "A reference model for designing effective reputation information systems," *Journal of Information Science*, vol. 31, pp. 365-380, October 1, 2005 2005. http://jis.sagepub.com/cgi/content/abstract/31/5/365
[100] J. Sabater and C. Sierra, "Review on Computational Trust and Reputation Models," *Artificial Intelligence Review*, vol. 24, p. 33, 2005. http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s10462-004-0041-5
[101] M. E. Winston, R. Chaffin, and D. Herrmann, "A taxonomy of part-whole relations," *Cognitive Science*, vol. 11, p. 417, 1987. http://www.sciencedirect.com/science/article/B6W48-4F2T79V-N/2/f593f56d2d4f562c5e0a75da4bf2f883
[102] "ThomasNet: Powered by ThomasRegister and ThomasRegional," Thomas Industrial Network. http://www.thomasnet.com/browse.
[103] "OECD Macrothesaurus," Organisation for Economic Co-operation and Development http://info.uibk.ac.at/info/oecd-macroth/en/.
[104] K. Østerbye, "Structural and cognitive problems in providing version control for hypertext," in *ACM conference on Hypertext*, Milan, Italy, 1993, pp. 33-42. http://doi.acm.org/10.1145/168466.168484.
[105] "Internet Archive: Universal Access toHuman Knowledge." http://www.archive.org/about/about.php.
[106] T. H. Nelson, *Literaty Machines: The Report on, and of, Project Xanadu Concerning Word Processing, Electronic Publishing, Hypertext, Thinkertoys, Tomorrow's Intellectual*

Revolution, and Certain Other Topics, Including Knowledge, Education and Freedom, 1980.