Communications of the Association for Information Systems

Volume 13 Article 22

2-25-2004

Visualizing Research Digital Libraries with Open Standards

Mark Ginsburg
University of Arizona, mginsburg@eller.arizona.edu

Recommended Citation
Ginsburg, Mark (2004) "Visualizing Research Digital Libraries with Open Standards," *Communications of the Association for Information Systems*: Vol. 13 , Article 22.
DOI: 10.17705/1CAIS.01322
Available at: https://aisel.aisnet.org/cais/vol13/iss1/22

Communications of the Association for Information Systems

VISUALIZING RESEARCH DIGITAL LIBRARIES WITH OPEN STANDARDS

Mark Ginsburg
Management Information Systems Department
University of Arizona
mginsbur@eller.arizona.edu

ABSTRACT

Large scale research Digital Libraries (DLs) have a large array of potentially useful metadata. Yet, many popular DLs do not provide a convenient way to navigate the metadata or to visualize classification schema in the user session. For example, in the broad world of Management Information Systems (MIS) research, a high-level overview of MIS topics and their inter-relationships would be useful to navigate a MIS DL before zooming in on a specific article. To address this obstacle, this paper describes a prototype, the Technical Report Visualizer System (TRV), which uses a wide variety of open standards to expose DL classification metadata in the navigation interface. The system captures MIS article metadata from the Open Archives Initiative (OAI) compliant arXiv e-Print archive at Cornell University. The OAI Protocol for Metadata Harvesting (OAI-PMH) is used to collect the topic metadata; the articles' Association for Computing Machinery's (ACM) Computing Classification System codes. We display the topic metadata in a Java hyperbolic tree and make use of XML conceptual product and implementation product standards and specifications, such as the Dublin Core and BiblioML bibliographic metadata sets, XML Topic Maps, Xalan and Xerces, to link user navigation activity to the abstracts and full text contents of the articles. We discuss the flexibility and convenience of XML standards and link this effort to related digital library visualization approaches.

Keywords: digital library visualization, MIS Classification, XML, XML topic maps, metadata, OAI-PMH, ARXIV, ACM CCS, ACM/IEEE CC2001, INSPEC

I. INTRODUCTION

Several limitations exist in accessing and retrieving science and engineering documents from the various large-scale digital libraries (DLs) that researchers enjoy today. Determining trends or locating topical articles in specialized areas involves searching disparate web and non-web based sources, usually with full-text or fielded search. Information foraging in this manner becomes cumbersome with no easy way to compare and classify articles into well-defined topics [Tennant 2000]. Relating keywords and subject matter to specific topics in attempt to compare articles is a difficult and time-consuming part of this process.

In effort to alleviate the problems associated with synthesizing information from different sources, several organizations developed large-scale research article repositories that are classified to varying degrees according to predetermined schemes. For example, the arXiv.org e-Print archive at Cornell University, which houses over 200,000 scientific papers on physics, mathematics, and computer science organizes the topics according to standard classification schemes. Computer science papers submitted to the arXiv contain author-supplied codes based on the Association for Computing Machinery's (ACM) 1998 Computing Classification System (CCS1998) [Coulter, French et al. 1998].

However, most repository systems such as arXiv or the ACM Digital Library (http://www.acm.org/dl) do not present classification metadata in a visual interface. Thus, the user is burdened to construct various fielded searches to locate and compare articles by code. For example, the ACM classification search is not part of the top-level documentation. A researcher must consult the classification scheme and then learn that a fielded search to locate, e.g., heterogenous database research papers (CCS1998 code: H.2.5) must take the form "+primaryccs:H.2.5". The interface burden is an unnecessary obstacle between the information (data and metadata) contained in the archive and the prospective readership. The high-level solution is an overlay of the classification metadata into an interface that allows the user a simple way to explore and navigate the classification scheme.

Fortunately, open standards (mostly falling under the XML umbrella) exist to simplify the intertwining of classification schema in the researcher's digital library navigation experience. Before exploring this point further, we first describe briefly related work in the area.

RELATED WORK

The activity in digital library visualization is significant. The following are examples:

- To accommodate the general library user in an Internet web browsing setting, the Scholastica Project, overlays the Library of Congress Classification (LCC) captions on a visual representation of library shelves[Beagle 2003]. The system provides drill-down capabilities. The user interface, based on the Visual Net XML platform, is intuitive and resembles books grouped by LCC captions on a traditional library shelf.
- If the search space and the potential audience are both broad, algorithms such as a Kohonen self-organizing map (SOM) technique may be used to generate two dimension concept maps. SOM maps are widely used in conjunction with various visualization strategies; for example the comparison of fisheye and fractal views in a broad Internet search task [Yang, Chen et al. 2003]. These authors report that the concept subspaces may overload the user (excessive 'visual load') and visualization strategy is an important consideration.
- Concept maps may be generated with Latent Semantic Indexing [Deerwester, Dumais et al. 1990]. This technique has been applied in numerous specialized domains, for example in biomedicine [Chute, Yang et al. 1991].
- The Alexandria Digital Earth project visualizes geospatial data [Hill, Janée et al. 1999] using the ADEPT digital library architecture [Janée and Frew 2002] which provides logical "buckets" for metadata collection descriptions.
- The GenNav project [Bodenreider 2002] provides an interface to link a large glossary of genetic terms to the visualization of gene ontology pathways for biomedical researchers.

In this paper, we narrow the consideration to a Digital Library with one or more well established *a priori* classification structures and a domain-specific audience, for example MIS and CS researchers. The rest of the paper is devoted to answering the research question "Can we make use of open standards to link classification structures to the digital library user's navigation session in a convenient and flexible manner?" in the affirmative by discussing the design and implementation of two prototypes, the Technical Report Visualizer (TRV) and an extension to accommodate multiple classification views, TRV2. During this discussion we will show how the XML suite of standards are advantageous, in particular XML Topic Maps (XTM).

We first discuss XML conceptual and implementation product standards, introduce process standards, and diagram their interplay in Section II. In Section III, we provide a background on MIS classification schemes domains..In Section IV, we present the TRV and TRV2 system architectures to understand how the XML components are combined to provide multiple classification overlays to the digital library user. We conclude in Section V with a discussion of the logical next step in DL Visualization process standard development; an expansion of digital library visualization to include articles that do not have *a priori* classification.

II. THE FLEXIBILITY OF XML STANDARDS

The basic XML 1.0 specification first released in 1998 and now in its third edition [Bray, Paoli et al., 2004] focuses on a well-defined service: the interchange of extensible documents. This interchange evolves the HTTP standard, which allows the networked sharing of hypermedia resources with a predefined tag set. XML allows self-authored tags thus permitting extensible documents. From this core standard, numerous additional standards were developed that depend on the original specification.

To illustrate this point, we can sketch a path of some of the iterative XML work, with the World Wide Web Consortium (W3C) Recommendation date in parentheses. W3C working groups issue drafts in each XML area with the goal of W3C recommendation status.

For example, XML Namespaces (1/14/99) [Bray, Hollander et al., 1999] bind document element prefixes to URIs. XMLBase, (6/27/01) [Marsh, 2001], provides URI Services to XLINK. XLINK, introduced at the same time, allows an XML document to reference another. The XPATH language (11/6/99) [Clark and DeRose, 1999] allows more than a simple cross-reference: for the first time, it specifies ways to reference internal portions of other XML documents. Around the same time, XSLT (11/16/99) [Clark, 1999] specified a transformation language to transform one XML tree to another. XPointer, based on the XPATH language, (3/25/03) [Grosso, Maler et al., 2003], is another internal document reference set of techniques. And one of the newest, XEvents (8/4/03) [McCarron, Pemberton et al., 2003], defines event listeners and event handlers with DOM Level 2 event interfaces. "The result is to provide … an interoperable way of associating behaviors with document-level markup" [Anonymous 2001]. The focus is shifting from the mere transfer of extensible documents to the new question, "what can we make these documents do?" Part of this new emphasis is DOM, the Document Object Model. DOM is a platform- and language-neutral interface that will allow programs and scripts to access and update the content, structure and style of documents dynamically; the area of another W3C Working Group (http://www.w3c.org/DOM/)

The theme is a new generation of standards building on those in the recent past. They give rise to implementation product standards [Cargill 1989] - tangible software products. Using the W3C XML Recommendations, many implementation standards come from the Apache XML Project (http://xml.apache.org). For example, The Apache website, http://xml.apache.org/, states, "The goals of the *Apache XML Project* (part of The Apache Software Foundation) are:

- to provide commercial-quality standards-based XML solutions that are developed in an open and cooperative fashion,
- to provide feedback to standards bodies (such as IETF and W3C) from an implementation perspective, and
- to be a focus for XML-related activities within Apache projects".

To give two examples of implementation product standards from Apache, Xerces "provides world-class XML parsing and generation. Fully-validating parsers are available for both Java and C++, implementing the W3C XML and DOM (Level 1 and 2) standards … The parsers are highly modular and configurable. Initial support for XML Schema (draft W3C standard) is also provided." Xalan "provides high-performance XSLT stylesheet processing. Xalan fully implements the W3C XSLT and XPath recommendations[1]"

The overall goal of the TRV system, digital library classification visualization, is made possible by the judicious use of a number of the aforementioned XML standards. This type of standard "is inherently mutable . . . a standard that describes an expected outcome for a future need . . . the process, by definition, must be somewhat vague, to allow technology and the market to change without obsoleting the standard, and the future orientation of the standard means that its description of a process must be even more hazy" [Cargill 1989]. Thus, while underlying document exchange mechanisms may change, but the overall result, flexible DL traversal, remains the same.

In today's computing environment, digital libraries are still in their infancy. One of the usability problems is that DL metadata is often obscured from the user to some degree. In the ACM DL, the existing query system makes it difficult to utilize the classification metadata. To simplify our development, we make use of a convenient framework for metadata exchange, the Open Archives Initiative Protocol for Metadata Handling (OAI-PMH) [Lagoze and Van de Sompel 2001]. This protocol allows service providers to capture metadata from conformant DLs (the metadata is "harvested" in OAI-PMH terminology and comes across the network as an XML byte stream). Using OAI-PMH, Java servlets and hyperbolic tree applets, and XML standards from the W3C and Apache, a process standard of DL classification visibility can be shaped. An important property of the standard is flexibility. No single visualization approach can satisfy all audiences and similarly, no single classification tree can adequately describe a given DL. Hence it is important to build adaptability into the system. The remainder of the paper describes two reference implementations of this process standard, the Technical Report Visualizer System (TRV) and an extension, TRV2.

The interplay of OAI-PMH, Java, and various XML conceptual, implementation and process standards in digital library visualization is shown in Figure 1.

The system presented in this paper, the Technical Report Visualizer, is a reference implementation addressing Process Standard #1, "Overlay of Classification Metadata to Facilitate Digital Library Navigation", in Figure 1.

III. CLASSIFYING THE TOPICS OF MANAGEMENT INFORMATION SYSTEMS RESEARCH

The Management Information Systems (MIS) field is difficult to define independently from reference disciplines. Attention has been given to its problematic evolution as a separate scientific field, focusing on its lack of paradigmatic framework and fragmentation of research topics. which can be as diverse as economics, computer science, management theory, and behavioral science. In addition, research in the MIS field tends to shift focus rapidly and lacks overall coordination among its members [Banville, Landry et al. 1989]. As a result of its fluid, multidisciplinary nature, understanding the current state of MIS research is no easy matter.

BUILDING A CLASSIFICATION MODEL FOR MIS

Several attempts were made in the last 20 years to classify and map the development of MIS as a separate academic field. For example, Culnan [Culnan and Swanson 1986; Culnan 1987] demonstrated in her Co-Citation Analyses that distinct clusters of research were emerging within

---

[1] http://xml.apache.org/

the field during the late 1980's. Source articles were gathered from a wide range of journals including MIS Quarterly, Communications of the ACM, Management Science, and several IEEE

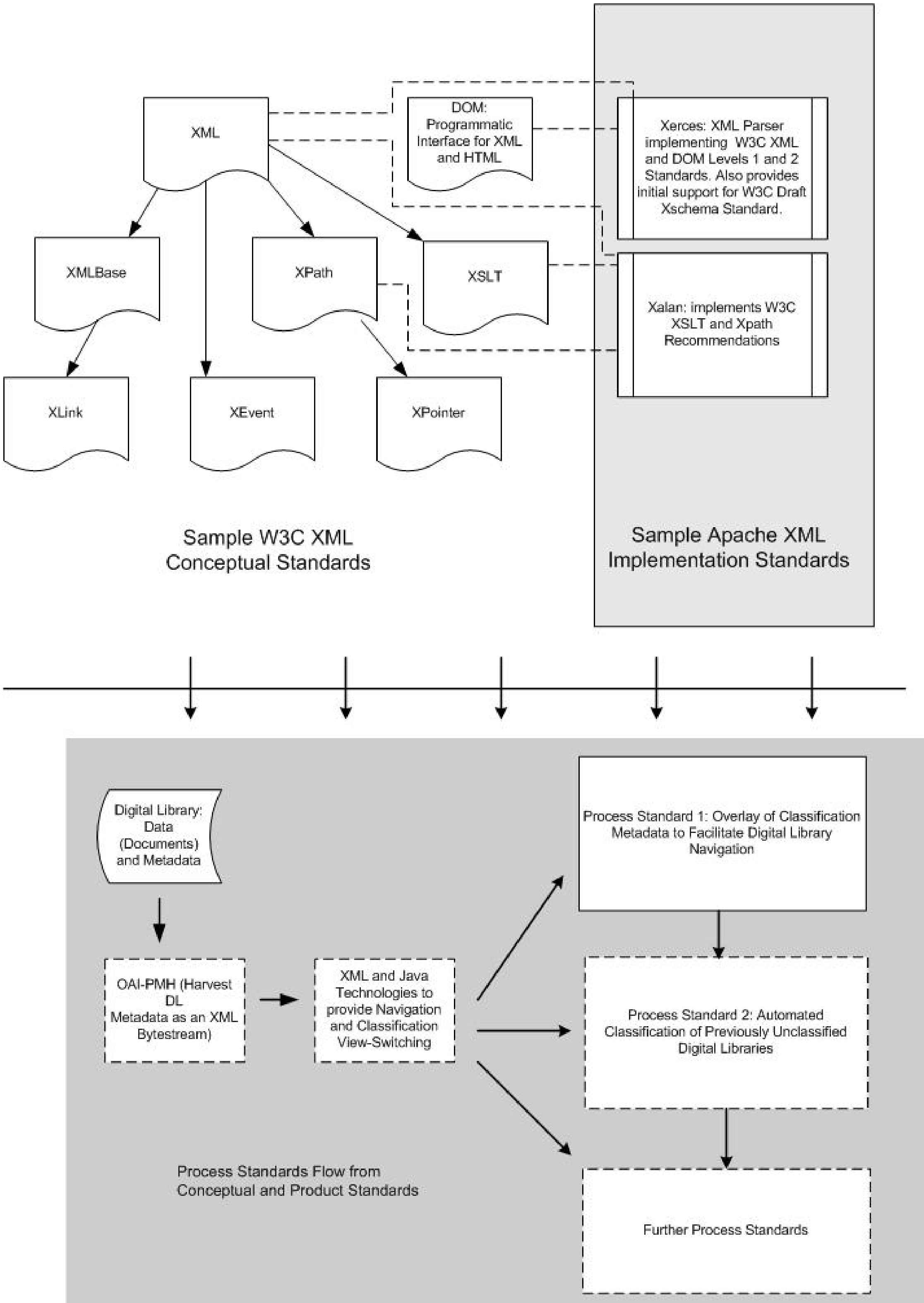


Figure 1. A Family of XML Standards Feed Digital Library Process Standards.

periodicals. Bibliographies were then compiled that contained all the articles that cited at least two prominent authors within each research cluster. Culnan determined that, despite the emergence of research clusters, the MIS field remained fragmented based on the scattered distribution of research publications.

While Culnan's efforts focused on developing a map of the intellectual structure of MIS based on research patterns, others such as Barki, et al attempted to develop a keyword classification system that would codify the entire set of topics that comprise the MIS discipline [Barki, Rivard et al. 1988; Barki, Rivard et al. 1993]. In addition, Barki et al. intended their IS Keyword Classification Scheme to facilitate the development of computerized information retrieval systems, customized for the IS field.

In computer science and MIS Digital Libraries, several classification schemes are possible: the ACM's Computing Classification System (CCS)[2] [Coulter, French et al. 1998] for Communications of the ACM and arXiv, the CoRR system, the joint ACM/IEEE CC2001 curriculum[3] and the INSPEC system[4] for IEE and IEEE journals. For example, in the CCS hierarchy, most MIS articles are classified under CCS Node "H" (Information Systems) and/or "I" (Computing Methodologies) whereas with INSPEC, MIS falls under C7100 and D2010. In CC2001, MIS articles might fall under any number of "IS" (Intelligent Systems), "IM" (Information Management), "HC" (Human-Computer Interaction) or "SE" (Software Engineering) nodes.

RESEARCH DIGITAL LIBRARIES AND THEIR METADATA

In the mid-1990s, the rapid ascent of the World Wide Web (WWW) made it much easier for technical report authors and conference organizers to post MIS articles online. Some MIS journals soon followed suit. Digital libraries encompassing MIS themes proliferated; some with classification data and some without. These libraries include NEC's CiteSeer, and the Los Alamos National Lab's Computing Research Repository (CoRR) under the aegis of Cornell University's e-Print Archive. Professional societies, such as the ACM and IEEE, offer collated digital libraries of their journal and conference offerings. There is also an effort by Georgia State University researchers to collate key MIS journals in the ISBib project [Chua, Cao et al. 2002]. Another meta-level effort is the Computing and Information Technology Interactive Digital Educational Library (CITIDEL; http://www.citidel.org/) that uses the OAI-PMH to harvest records from conformant member collections, such as the ACM DL, ACM CHI, and the DBLP Computer Science Bibliography. Note that the existence of a classification scheme for a set of publications does not automatically mean the corresponding digital library uses it. For example, as of this writing the IEEE XPlore DL does not offer the user the chance to use the INSPEC scheme to search the archive. Table I shows a selection of research digital libraries and the classification schema available from their UI search features.

ACCESSING DIGITAL LIBRARY METADATA WITH THE OPEN ARCHIVES INITIATIVE (OAI) STANDARD

Networked digital libraries are of limited usefulness if they stand isolated, with no way to interchange data about their collection; i.e. their metadata. To build the vision of a global digital library [Fox and Marchionini 1998] standards are called for to facilitate networked transfer of information about DL collections and subcollections. The standards should specify the syntax and

---

[2] http://www.acm.org/class/1998/overview.html provides the high-level overview of the ACM CCS system.

[3] http://www.computer.org/education/cc2001/

[4] IEE information on the INSPEC classification scheme is available at: http://www.iee.org/Publish/Support/INSPEC/Document/Class/index.cfm; the INSPEC Thesaurus, based on the INSPEC classification, is available at: http://ieeexplore.ieee.org/lpdocs/epic03/xplorehelp/thesaurus.htm.

network protocol(s) to construct such queries. Protocols under the umbrella of the Open Archives Initiative (OAI) [Breeding 2002; Cole, Habing et al. 2002] such as the OAI Protocol for Metadata Harvesting (OAI-PMH) answer precisely this need[5]. OAI was designed to be a catalyst to the establishing of a low-entry and well-defined interoperability framework applicable across domains [Lagoze and Van de Sompel 2001]. OAI-PMH allows users to make date-based harvesting requests on the archive.

Table 1: A Selection of Research Digital Libraries and their URLs.

Research Digital Library	URL	Supports OAI-PMH?	Classification Schemes available in the User Interface
NEC's CiteSeer	http://citeseer.nj.nec.com/	No	None
The Los Alamos National Lab's Computing Research Repository (CoRR)	http://xxx.lanl.gov/archive/cs/intro.html	No	ACM CCS and CORR/arXiv (internal)[6]
Association for Information Systems (AIS): CAIS, JAIS, MISQ, eMISQ	http://aisnet.org/	No	None
ACM DL	http://www.acm.org/dl/	No	ACM CCS
IEEE XPlore DL	http://ieeexplore.ieee.org/Xplore/DynWel.jsp	No	None
Georgia State University's ISBib	http://readable.eci.gsu.edu:8080/examples/servlet/isbib	No	None
Cornell University's e-Print Archive	http://www.arxiv.org/	Yes	For CS Archive, ACM CCS plus CORR/arXiv
Virginia Tech's CITIDEL	http://www.citidel.org/	Yes	ACM CCS, CC2001, CORR/arXiv, and Mathematical Subject Classification 2000[7]

IV. THE TRV PROTOTYPE: VISUALIZING THE ARXIV CITATION TOPICS

The stage is now set to present a reference implementation of a classification overlay using XML standards: the Technical Report Visualizer (TRV). TRV provides an intuitive navigation interface to the Management Information Systems subset of the arXiv computer science papers[8] via a hyperbolic-tree- based visual map of the ACM code metadata with hyperlinks to the full text. The system uses a wide variety of XML conceptual product and implementation product standards from the World Wide Web Consortium (W3C) and the Apache XML Project. Of particular interest is the system's use of XML Topic Maps (XTM) [Auillans 2002]. The XTM standard facilitates locating and managing information through topic organization and relationship. Within an XTM,

---

[5] http://www.openarchives.org/ has information on internationally funded OAI projects as well as OAI software and industry news.

[6] http://arxiv.org/new/cs.html

[7] http://www.ams.org/msc/

[8] The TRV prototype is accessible at: http://louvain.bpa.arizona.edu/trv/tree.html - to run it, the Java 2 plug-in is required.

topics are linked to form a semantic network of information. This organized view of related topics lends itself well to visualization [Auillans 2002]. Later in this section, we present TRV2, an extension to provide multiple classification views.

Using OAI-PMH, the Technical Report Visualizer (TRV) system harvests classification metadata from the Cornell arXiv e-Print Technical Report archive. The arXiv collection gathers ACM CCS1998 classification codes [Coulter, French et al. 1998] from the author(s) upon submission; the TRV system harvests them and shows the ACM scheme in a hyperbolic tree representation for the researchers. Using the tree interface, the original TRV system allowed the users navigate the ACM classification scheme. The system requires XML Topic Maps [XTM], one XTM file per classification scheme, and BiblioML citation records. BiblioML is an XML specification (not a standard) based on the International Federation of Library Associations and Institutions (IFLA) UNIMARC classification format [Anonymous 1994; Sévigny and Bottin 2000; Cover 2001]. In related work on citation management, BiblioML was used to provide a middle ground, or 'hub', between the 'spokes' EndNote (Windows) and BibTeX (Unix) [Ginsburg 2004]. In the TRV system, Dublin Core citation metadata is fetched from arXiv and transformed to BiblioML using XSLT. There is no intrinsic advantage of BiblioML (it was a legacy decision made in the Open Citation System [Ginsburg 2004]) and the native arXiv Dublin Core metadata set would serve just as well

The system encodes a Classification Tree (for example, ACM CCS1998) manually as an XTM file [Mason and al. 2000] file. XTM is an official ISO/IEC standard that allows representing the parent/child relationships. Metadata is fetched and converted to BiblioML format. These records include a hyperlink pointing back to the full content. We use the BiblioML "Note" element of type "SubjectAccess" to store the ACM Classification of each citation using an Xlink reference (see Appendices I and II for sample BiblioML records). Multiple classifications can be handled here (cf. Appendix II); our "Note" format uses the prefix as the name of the tree (e.g. ACM CCS) and the suffix as the node label, e.g. "H.2.1". A pound sign ("#") is the delimiter. More detail on the technical details of the implementation are given in [Keippel and Watson 2002]. Appendices I, II, and III for sample listings of BiblioML and XML Topic Map files used by the TRV system

THE TRV SYSTEM IS BUILT ON XML CONCEPTUAL AND IMPLEMENTATION PRODUCT STANDARDS

Since the arXiv DL stores its metadata records in Dublin Core XML format, and the OAI-PMH protocol streams results from metadata harvests in XML as well, it is natural for our system to use XML technologies to process OAI-PMH harvest records, match them to our XML Topic Map(s), display them with the applet, and transform them with the servlet.

Appendix IV summarizes the XML technologies used in the TRV prototype. Most of the conceptual product standards are under the aegis of the W3C and most of the implementation product standards come from the open source Apache XML project (Figure 1). The high availability and rapid development of these standards means the TRV modular system is able to take advantage of advances in any of its underlying components.

COMMENTS ON OUR INFORMATION VISUALIZATION STRATEGY

In the large field of information visualization, some key summary lessons can be drawn [Hearst 1999]. One of them is to provide the user with a way to jump around quickly; what the user sees on the screen can trigger a secondary goal and the user should be able to jump quickly to that goal – the idea of support for "berrypicking" [Bates 1989]. Another goal of the interface should be to support learning. The user, while interacting with the interface over time, can learn more about the underlying structure of the document collection. This idea is supported by the hyperbolic interface [Pirolli, Card et al. 2000; Pirolli, Card et al. 2001] that we use in our TRV prototype, we use the effective hyperbolic interface in the top frame, while providing detailed results using XML and XSL in the bottom frame.

In Figures 2 through 4 in the next subsection, the TRV system shows an example of "Focus + Context" information visualization [Green, Marchionini et al. 1997] [Leung and Apperley 1994] [Plaisant, Carr et al. 1995]. This type of visual organization was shown to lead to faster and more intuitive user navigation [Pirolli, Card et al. 2000; Pirolli, Card et al. 2001; Börner and Chen 2002]. The Focus + Context frames give the users clues in both windows to help locate information more quickly to "reduce the cost structure of information" [Card, Mackinlay et al. 1999]. This approach is consistent with the general knowledge management literature [Davenport and Prusak 1998] [Nonaka 1994] [Ginsburg 1998; Ginsburg 1999]; for example, inside a company the professional authors spend time and effort creating documents. Hence if interested professional readers cannot locate the authors' work products in a timely and effective manner the knowledge management practice is poor and can stand improvement.

However, a hyperbolic interface is not a panacea that will always meet the users' requirements. If the tree structure is overly dense, not enough screen real estate is available to navigate the structure easily. This problem points out an advantage of our modular approach, which separates the visualization strategy from the XML-based processing. The system can easily be modified to suit other navigation approaches and underlying structures. Note that the visualization strategy (realized with a set of Java components) is completely decoupled from the metadata harvesting and transformation (realized with a set of XML components). Thus the process standard is mutable simply by swapping or extending components. New implementation standards can be developed easily. The next subsection details such a process evolution; the TRV2 system.

TRV2: EXTENDING THE CONCEPTUAL PROCESS STANDARD WITH MULTIPLE VIEWS AND NODE MEMBERSHIP CONTEXT CLUES

There is no compelling reason to box the users into one classification scheme in a visualization interface. Therefore, a key design requirement is to allow dynamic changes of the view inside the user session. Different views can accommodate multiple taxonomies on the same document archive, or different levels of user expertise in the subject area (such as University researchers versus K-12 students). In an extension to the original TRV system, students Luis Chaboya and Christopher Willard implemented dynamic runtime view switching between CCS1998 and CC2001 to build TRV2[9]. In addition, the extension addressed a drawback of the original TRV system: in the initial effort, the tree nodes did not indicate the citation population underlying a given node. Thus the users, when traversing the tree and selecting a node, had no way of knowing ahead of time how big a result set to expect in the bottom frame. This lack of context is a serious deficiency in many tree-based approaches. We approached it in TRV2 with two proof of concept population context strategies, both present in the new interface for the user to choose. The first new population context clue is "Colorization", where nodes with more members are given a higher intensity color (the basic system starts each level in the tree with the same color). The second is "Box Size", where an enlarged border is used for nodes with more members. In both context extension strategies, the number of citations encompassed by the node are also indicated in square brackets as a suffix to the node label.

This extension allows each BiblioML file to store multiple encodings. On the front-end, the user can select a particular view and the Tree Visualizer Servlet and Applet will work with the selected Topic Map. The approach requires preparation of a new Topic Map file for each new view provisioned, and the additional "Note" encoding (Appendix II shows a sample listing) in the BiblioML files. Documents already classified with the CCS1998 scheme were given an extra CC2001 classification manually. Appendix V shows the overall architecture of the TRV2 system.

Figures 2, 3, and 4, taken from the TRV2 system, demonstrate runtime view switching and the colorization and box size population context strategies. In Figure 2, we see an example of

[9] The TRV2 prototype is accessible at: http://louvain.bpa.arizona.edu/trv2/tree.html - to run it, the Java 2 plug-in is required.

Colorized Nodes to give membership context; with the standard ACM topic map being used to form the tree that we saw in the previous discussion.

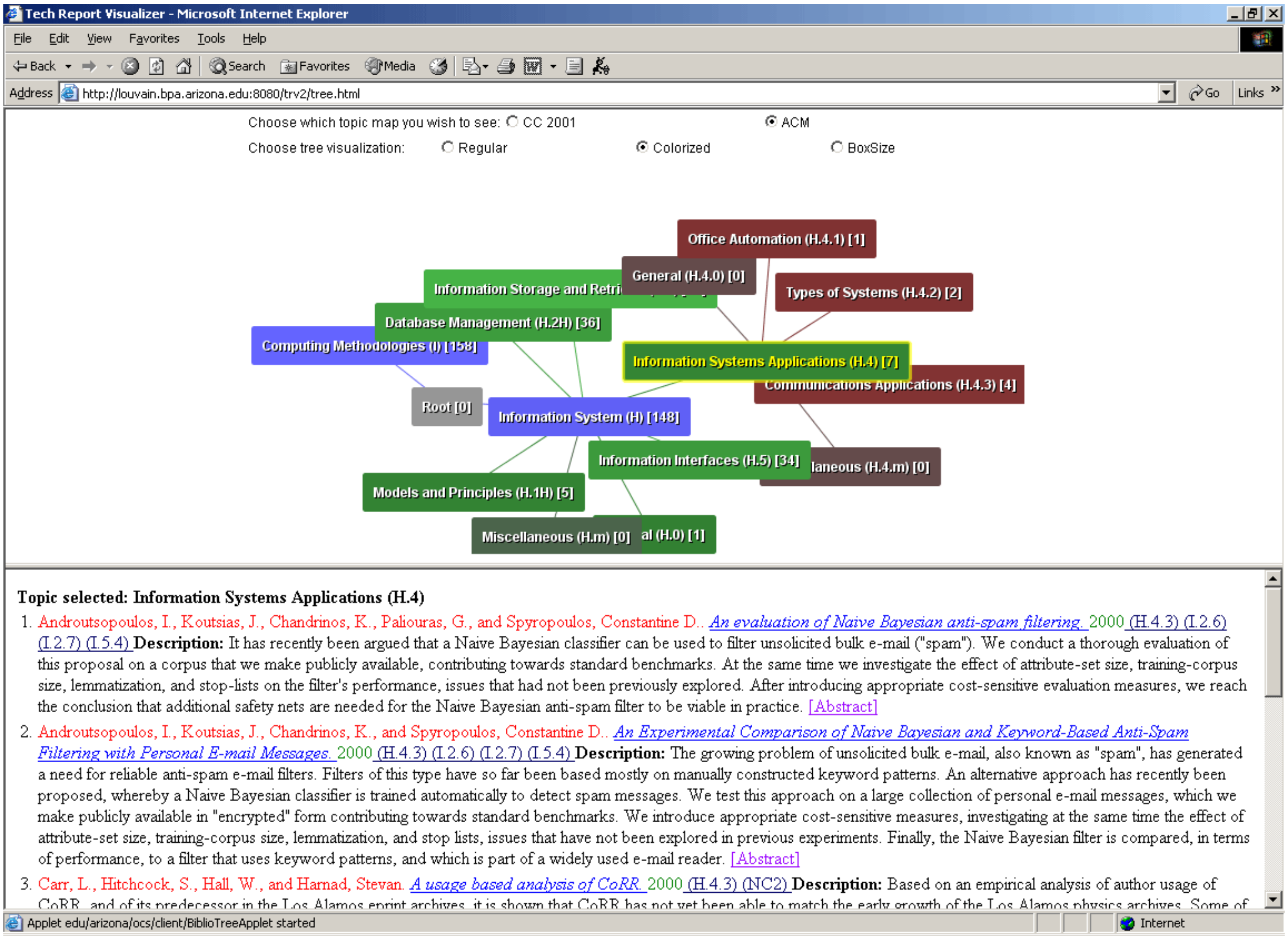


Figure 2. Additional Context: ACM View Selected and Boxes Colorized to Indicate Node Membership

In Figure 3, we again use the ACM View, but this time we select the BoxSize Membership Size context strategy.

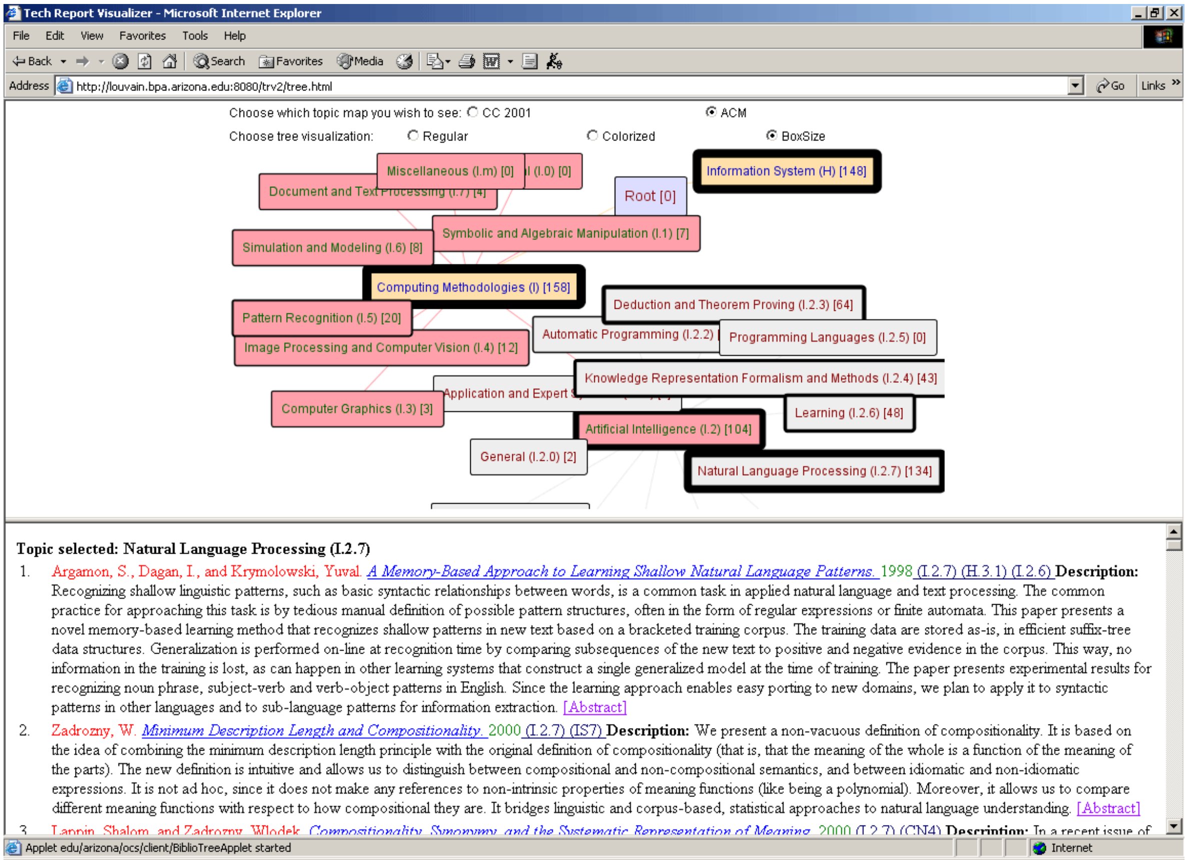


Figure 3: Additional Context: ACM View Selected and Box Border Size Varies to Indicate Node Membership

Figure 4 on the next page presents the user switching the view away from the ACM Topic Map and instead picking the CC2001 tree.

To sum up, the TRV is a reference implementation of a conceptual process standard – to remove the mystery of classification schemes from the user's perspective. The TRV2 extension allows view switching and provides additional membership node information via colorization or border size, to help the users expect a certain number of results before they actually click on a given node. As can be seen in Appendix I, the visualization strategies are decoupled from the metadata fetching and classification strategies for maximum flexibility – an indicated property of conceptual process standards.

RELATED WORK IN SPECIALIZED DIGITAL LIBRARY VISUALIZATION: THE CITIDEL DL

CITIDEL (http://www.citidel.org/?op=browse) also uses OAI-PMH to harvest metadata from conforming collections. CITIDEL passes the metadata to an interface which features tabs that allow view switching, and a tabular representation of each view rather than the TRV hyperbolic tree. A screenshot of the CITIDEL approach is presented in Figure 5. This implementation process standard of the same conceptual process standard uses Perl scripts to handle the user browse requests (e.g. Tab selections) and stores the classification information in a database (the TRV and TRV2 systems do not use a database). Thus different middleware components and different front-end components were used to create a different look and feel.

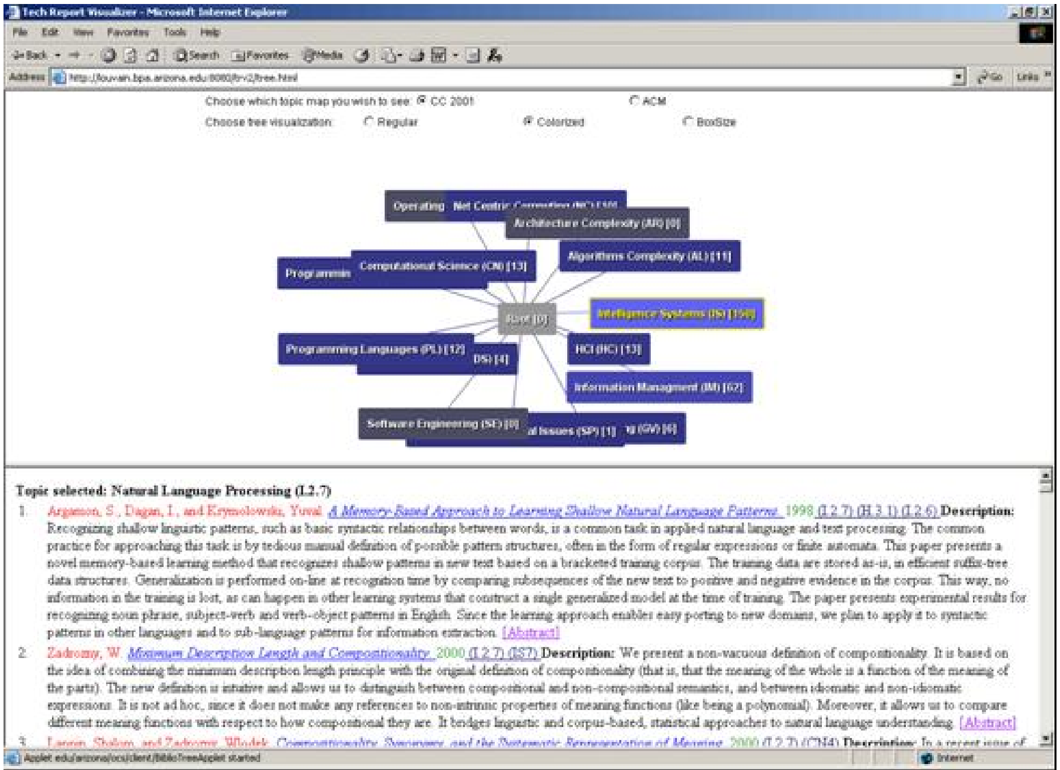


Figure 4: Switching the View: The CC2001 Tree is Visualized with Colorization Membership Context

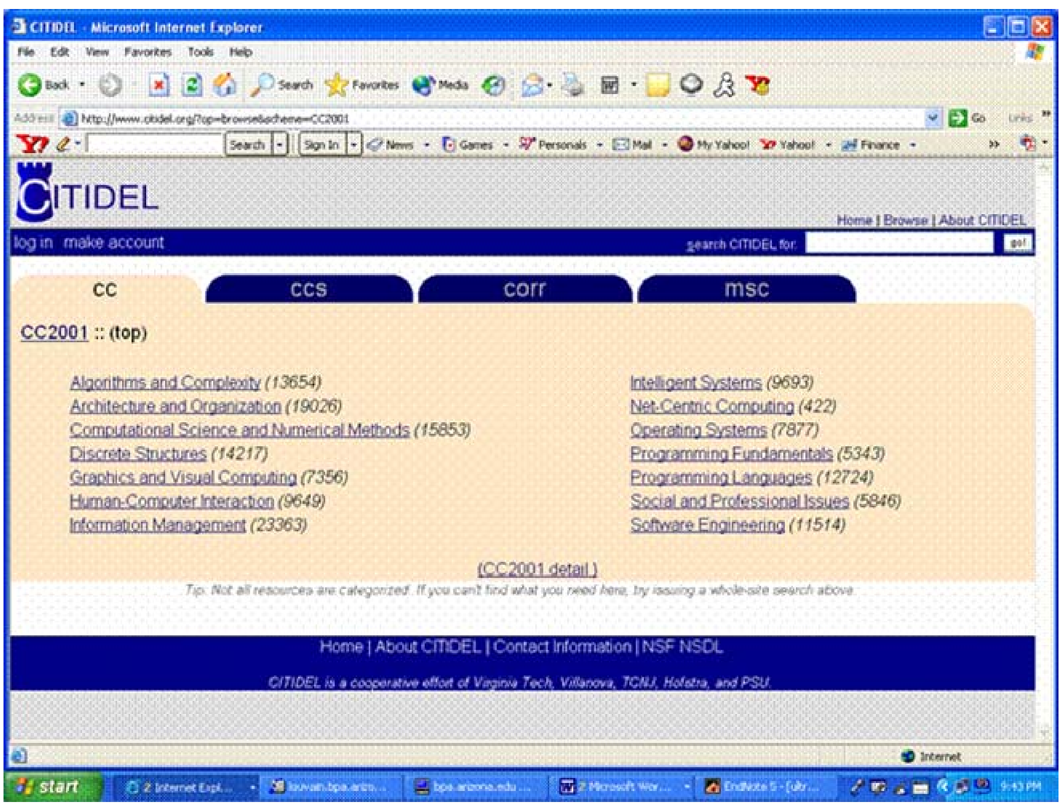


Figure 5: An Alternative Implementation Process Standard: CITIDEL's Metadata Visualization Strategy

V. FUTURE PLANS AND CONCLUDING REMARKS

FUTURE PLANS

Our first step will be to validate the TRV and TRV2 classification interfaces with usability studies. Referring to the bottom of Figure 1, the next step is a process standard to handle the situation of digital libraries that do not use classification metadata *a priori*. The construction of classification metadata (even if none exists) is essential to the realization of this effort. Since many repositories do not use the ACM classification scheme and others do not use any scheme at all, innovative techniques such as running index terms through classification engines will be required to present the articles effectively in an organized manner.

To classify documents into previously established hierarchies, a number of techniques are available. For example Chung and Clarke [2002] used HTML data and metadata to classify Web documents using such diverse algorithms as Naïve Bayes Classifiers [Mitchell 1997], Rocchio Feedback [Joachims 1997] and Support Vector Machines (SVM) [Joachims 2001]. In addition, a linear function categorizer shows promise [Wibowo and Williams 2002].

Categorization of large-scale DLs, such as the ACM DL, may require a relational database or other optimization techniques to ensure adequate performance as the system scales up. Note that we are not restricted to DLs; similar techniques can be applied to collections of articles that sit dormant in file system storage. This effort will require careful usability validation and subsequent iteration due to the danger of system misclassification.

CONCLUSIONS

We can summarize the progress described in this paper in knowledge management terms. Think of DLs as a collection of passive information atoms (articles) that represents potential knowledge flow between the authors and the readers. The reader, if he or she cannot locate the article due to navigation hindrances, will not see the information and knowledge transfer cannot occur. Using open standards to reduce the opacity of DL metadata coupled with the provision of a focus-and-context interface provides a means to move around the DL at the high level of the metadata classification tree while seeing specific results at the same time.

Consider the case of the MIS researcher interacting with a DL visualization system. Two key knowledge transfer advantages are offered.

1. The researcher is exposed to more potential matches per unit time than in a conventional full-text or fielded-search DL, such as the ACM DL, that offers an awkward back-and-forth iteration to try and retry metadata queries.
2. The act of simply seeing the high-level overview of the domain, in this case MIS, already accomplishes knowledge transfer: the researcher gains an appreciation for the inter-relationships of the topics in, for example, the ACM CCS1998 or the CC2001 classification scheme. This interaction reinforces the notion that the interface should provide a mechanism for learning [Pirolli, Card et al. 2000; Pirolli, Card et al. 2001]. Ongoing work with the system helps researchers learn standard MIS classification schema.

We also demonstrated another key property of an open-standards based DL visualization interface: flexibility. Because no single classification scheme can satisfy all audiences, it is important to allow specific audiences to select appropriate classification views. As we move forward to handle unclassified DLs, we should retain these properties and ensure the interface remains a flexible field of learning.

ACKNOWLEDGMENTS

We thank Britton Watson and Jeff Keippel who worked on the TRV prototype under the author's supervision as part of their MS degree requirements in the MIS Department at the University of

Arizona in the Fall Term, 2002. Another MS student, Damien Daspit, was instrumental in integrating XML and Java components in the Open Citation System that formed a useful starting point for the TRV system. MS students Luis Chaboya and Chris Willard worked on the TRV2 extension in the Spring Term 2003

Editor's Note: This article was received on December 3, 2003 and was published on March 30, 2004.

REFERENCES

EDITOR'S NOTE: The following reference list contains the address of World Wide Web pages. Readers who have the ability to access the Web directly from their computer or are reading the paper on the Web, can gain direct access to these references. Readers are warned, however, that

1. these links existed as of the date of publication but are not guaranteed to be working thereafter.

2. the contents of Web pages may change over time. Where version information is provided in the References, different versions may not contain the information or the conclusions referenced.

3. the authors of the Web pages, not CAIS, are responsible for the accuracy of their content.

4. the author of this article, not CAIS, is responsible for the accuracy of the URL and version information.

Anonymous (1994). "UNIMARC Manual : Bibliographic Format 1994". http://www.ifla.org/VI/3/p1996-1/sectn1.htm (current 3/2003).

Anonymous (2001). "XEvents: An Updated Events Syntax for XHTML™ and friends", W3C. http://www.w3.org/TR/2001/WD-xhtml-events-20010607/ Overview. Html (current 6/2003).

Auillans, P. (2002). "Toward Topic Maps Processing and Visualization". Proceedings of the Sixth International Conference on Information Visualization, London, England.

Banville, C., M. Landry, et al. (1989). "Can the Field of MIS be Disciplined?" *Communications of the ACM,* (32)1, pp. 48-60.

Barki, H., S. Rivard, et al. (1988). "An Information Systems Keyword Classification Scheme." *MIS Quarterly* (12)2, pp. 298-311.

Barki, H., S. Rivard, et al. (1993). "A Keyword Classification Scheme for IS Research Literature: An Update." *MIS Quarterly* (17)2, pp. 209-226.

Bates, M. J. (1989). "The Design of Browsing and Berrypicking Techniques for the On-line Search Interface." *Online Review* (13)5, pp. 407-431.

Beagle, D. (2003). "Visualizing Keyword Distribution Across Multidisciplinary C-Space." *D-Lib Magazine* (9)6.

Bodenreider, O. (2002). "Experiences in Visualizing and Navigating Biomedical Ontologies and Knowledge Bases". Intelligent Systems for Molecular BIology 2002, Edmonton, Canada.

Börner, K. and C. Chen (2002). "Visual Interfaces to Digital Libraries". Proceeding of the Second ACM/IEEE-CS Joint Conference on Digital libraries. New York: ACM Press.

Bray, T., J. Paoli et al., "Extensible Markup Language (XML) 1.0 (Third Edition)". http://www.w3.org/TR/2004/REC-xml-20040204/ (current 03/27/2004).

Bray, T., Hollander, D. and A. Layman. "Namespaces in XML". http://www.w3.org/TR/REC-xml-names/ (current 03/27/2004).

Breeding, M. (2002). "The Emergence of the Open Archives Initiative." *Information Today* (19)4, pg. 46.

Card, S. K., J. D. Mackinlay, et al. (1999). *Information Visualization*. San Francisco, CA: Morgan Kaufmann.

Cargill, C. F. (1989). *Information Technology Standardization: Theory, Process, and Organizations*. Bedford, MA: Digital Press.
Chua, C. E. H., L. Cao, et al. (2002). "IS Bibliographic Repository (ISBIB): A Central Repository of Research Information for the IS Community". *Communications of the Association for Information Systems*. (8)27, pp. 392-412.
Chung, C. and C. L. A. Clarke (2002). "Topic-Oriented Collaborative Crawling". Conference on Information and Knowledge Management (CIKM), McLean, VA.
Chute, G. C., Y. Yang, et al. (1991). "Latent Semantic Indexing of Medical Diagnoses Using UMLS Semantic Structures". Proceedings, 15th SCAMC, Washington, DC.
Clark, J. "XSL Transformations (XSLT)". http://www.w3.org/TR/xslt (current 03/27/2004).
Clark, J. and S. DeRose."XML Path Language (XPath)." http://www.w3.org/TR/xpath (current 03/27/2004).
Cole, T. W., T. G. Habing, et al. (2002). "Implementation of a Scholarly Information Portal Using Open Archives Initiative Metadata Harvesting Protocols. Interim Report to the Andrew W. Mellon Foundation", University of Illinois at Urbana-Champaign.
Coulter, N., J. French, et al. (1998). "Computing Classification System 1998: Current Status and Future Maintenance Report of the CCS Update Committee." *Computing Reviews*.
Cover, R. (2001). "BiblioML - XML for UNIMARC Bibliographic Records". http://xml.coverpages.org/biblioML.html (current 04/01/2003).
Culnan, M. J. (1987). "Mapping the Intellectual Structure of MIS, 1980-1985: A Co-Citation Analysis." *MIS Quarterly* (11)3, pp. 341-353.
Culnan, M. J. and E. B. Swanson (1986). "Research in Management Information Systems, 1980-1984: Points of Work and Reference." *MIS Quarterly* (10)3, pp. 289-302.
Davenport, T. H. and L. Prusak (1998). *Working Knowledge: How Organizations Manage What They Know*. Boston, MA: Harvard Business School Press.
Deerwester, S., S. T. Dumais, et al. (1990). "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41(6), pp. 391-407.
Fox, E. A. and G. Marchionini (1998). "Toward a Worldwide Digital Library." *Communications of the ACM* (41)4, pp. 29-32.
Ginsburg, M. (1998). "Annotate: A Knowledge Management Support System for Intranet Document Collections". PhD Dissertation, Information Systems Department, New York, NY: New York University,.
Ginsburg, M. (2004). "Unified Citation Management and Visualization Using Open Standards: The Open Citation System." *International Journal of IT Standards and Standardization Research* (2)1, pp. 23-41.
Ginsburg, M., A. Kambil (1999). "Annotate: A Knowledge Management Support System". HICSS-32, Hawaii, Los Alamitos, CA: IEEE Press.
Green, S., G. Marchionini, et al. (1997). "Previews and overviews in digital libraries: Designing surrogates to support visual information seeking". College Park, MD, University of Maryland CS.
Grosso, P., Maler, E., Marsh, J. and N. Walsh. "Xpointer Framework." http://www.w3.org/TR/xptr-framework/ (current 03/27/2004).
Hearst, M. A. (1999). "User Interfaces and Visualization". *Modern Information Retrieval*. Harlow, UK: Addison Wesley.
Hill, L. L., G. Janée, et al. (1999). "Collection metadata solutions for digital library applications." *Journal of the American Society for Information Science* (50)13, pp. 1169-1181.
Janée, G. and J. Frew (2002). "The ADEPT Digital Library Architecture". JCDL'02, Portland, OR, ACM.
Joachims, T. (1997). "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Classification". 14th International Conference on Machine Learning.
Joachims, T. (2001). "A Statistical Learning Model of Text Classification for Support Vector Machines". 24th Annual Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.
Keippel, J. and B. Watson (2002). "Visualizing the Management Information Systems Field Using Digital Libraries". MS-MIS Thesis, MIS Department, University of Arizona, Tucson, AZ

http://monet.ecom.arizona.edu/msr/PROJECTS/20024/MP-09-02.doc (current 02/01/2004).

Lagoze, C. and H. Van de Sompel (2001). "The Open Archives Initiative: Building a Low-barrier Interoperability Framework". Proceedings of the first ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 54-62, New York: ACM Press.

Leung, Y. K. and M. D. Apperley (1994). "A Review and Taxonomy of Distortion-oriented Presentation Techniques." *ACM Transactions on Computer-Human Interaction*(1)2, pp. 126-160.

Marsh, J. “XML Base“. http://www.w3.org/TR/xmlbase/ (current 03/27/2004).

Mason, J. D. and e. al. (2000). "ISO/IEC 13250 Topic Maps", ISO/IEC JTC 1/SC34.

McCarron, S., Pemberton, S. and T.V. Raman. “XML Events.” http://www.w3.org/TR/xml-events/ (current 03/27/2004).

Mitchell, T. (1997). *Machine Learning*, New York:McGraw Hill.

Nonaka, I. (1994). "A Dynamic Theory of Organizational Knowledge Creation." *Organization Science* (5)1, pp. 14-37.

Pirolli, P., S. K. Card, et al. (2000). "The Effect of Information Scent on Searching Information: Visualizations of Large Tree Structures". Proceedings of the Working Conference on Advanced Visual Interfaces, New York: ACM Press.

Pirolli, P., S. K. Card, et al. (2001). "Visual Information Foraging in a Focus + Context Visualization." ACM Conference on Human Factors in Computing Systems, CHI Letters 3, pp. 506-513.

Plaisant, C., D. Carr, et al. (1995). "Image-browser Taxonomy and Guidelines for Designers." *IEEE Software* (12)2, pp. 21-32.

Sévigny, M. and M. Bottin (2000). "BiblioML Project" http://www.culture.fr/BiblioML/en/index.html (current 03/01/2003).

Tennant, R. (2000). "Open Archives: A Key Convergence." *Library Journal* (125)3, pp. 122-123.

Wibowo, W. and H. E. Williams (2002). "Strategies for Minimising Errors in Hierarchical Web Categorisation". Conference on Information and Knowledge Management (CIKM), McLean, VA.

Yang, C. C., H. Chen, et al. (2003). "Visualization of Large Category Map for Internet Browsing." *Decision Support Systems* (35)1, pp. 89-102.

APPENDIX I. LISTING OF A SAMPLE BIBLIOML RECORD WITH SINGLE ENCODING (CCS1998 CODE H.3.7)

In the following listing, note how the ACM CCS98 code *H.3.7* is encoded in

```
<Note Type="SubjectAccess">xml/acm.xml#H.3.7</Note> </Notes>
```

(the “Type” attribute of the “Note” field near the bottom of the listing). We utilized the “Note” field as a miscellaneous storehouse to encode options used by the TRV system. The Note field is an XLink pointer to the H.3.7 node in the acm.xml XTM file which in turn can be parsed to obtain more information about the parent node H.3 and its parent node, H.

```
<?xml version="1.0" encoding="UTF-8"?>
<BiblioRecord Language="eng">
<Meta>
<CreationDate value="2002-06-24"/>
<TransactionDate value="2002-06-24"/>
<Status Value="n"/>
<RecordType Value="a"/>
<BibliographicLevel Value="a"/>
<Origin>
<Country>U.S.</Country>
```

```
<Agency>UA</Agency>
<TransactionDate value="2002-06-24"/>
</Origin>
</Meta>
<Description>
<LanguageInfo>
<Language Role="Document" Code=""/>
</LanguageInfo>
<TitleAndResponsibility>
<Work>
<TitleGroup>
<Title>CoRR: A Computing Research Repository</Title>
</TitleGroup>
<Responsibility Type="Primary">Halpern, Joseph Y.</Responsibility>
</Work>
</TitleAndResponsibility>
<PublicationGroup>
<Publication>
<Date>2000-05-03</Date>
</Publication>
</PublicationGroup>
<ElectronicLocation>http://arXiv.org/abs/cs/0005003</ElectronicLocation>
<PhysicalDescription>  Discusses how CoRR was set up and some policy issues
involved with setting up
such a repository.
</PhysicalDescription>
</Description>
<Notes>
<Note Type="SubjectAccess">xml/acm.xml#H.3.7</Note>
</Notes>
</BiblioRecord>
```


APPENDIX II. LISTING OF AN EXCERPT OF A BIBLIOML RECORD DEMONSTRATING MULTIPLE ENCODING.
(TWO CCS1998 AND ONE CC2001 CLASSIFICATION)

We used the "Note" field to load bibliographic citation with multiple classifications. This supports the TRV2 functionality of run-time classification view switching. The following example shows the XLINK references to the various XTM classification files.


```
<Notes>
<Note Type="SubjectAccess">xml/acm.xml#I.2.8</Note>
<Note Type="secondary">xml/acm.xml#I.6.5</Note>
```

<note type="descriptiveinformation">xml/cc2001.xml#pf3</note>

</Notes>

APPENDIX III. LISTING OF AN EXCERPT OF THE ACM CCS1998 XML TOPIC MAP

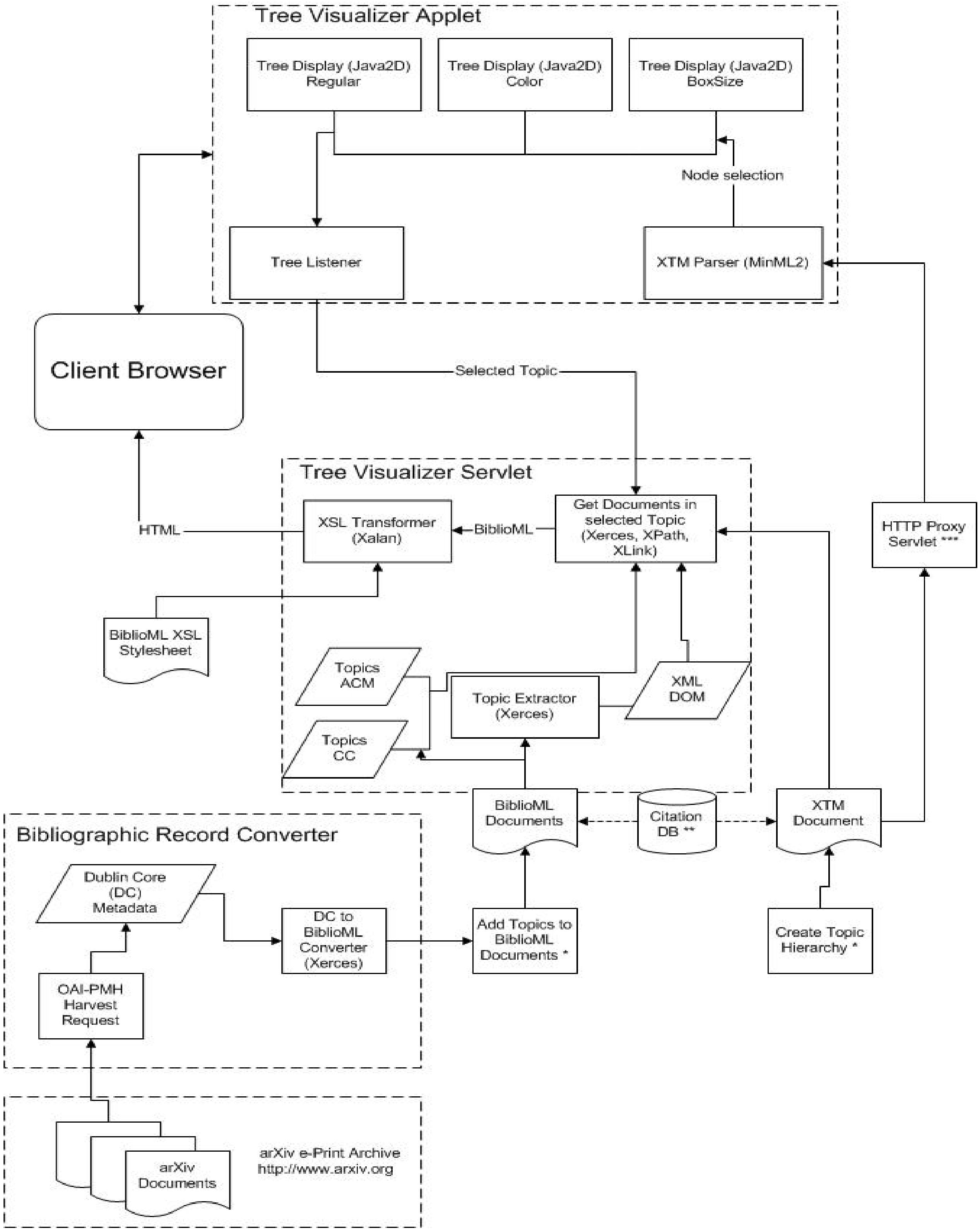
Our final listing shows an XML Topic Map (XTM) representation of the ACM CCS1998 tree. The two major nodes that we harvest from ArXiv in this project are "H" (Information Systems) and "I" (Computing Methodologies) as shown in this excerpt. This example also shows how sub-nodes, in this case `H.1` (Models and Principles) and `H.2` (Database Management) are encoded via XLink as children of H.

```
<?xml version="1.0" ?>
<!DOCTYPE topicMap PUBLIC "-//TopicMaps.Org//DTD XML Topic Map (XTM) 1.0//EN"
"xtm1.dtd">
<topicMap id="acm" xmlns="http://www.topicmaps.org/xtm/1.0/"
xmlns:xlink="http://www.w3.org/1999/xlink">
        <topic id="H">
            <baseName>
                 <baseNameString>Information Systems (H)
                 </baseNameString>
           </baseName>
        </topic>
        <topic id="I">
          <baseName>
              <baseNameString>Computing Methodologies (I)
              </baseNameString>
           </baseName>
        </topic>
        <topic id="H.1">
                <instanceOf>
                        <topicRef xlink:href="#H" />
                </instanceOf>
           <baseName>
               <baseNameString>Models and Principles (H.1H)
              </baseNameString>
          </baseName>
        </topic>
        <topic id="H.2">
                <instanceOf>
                        <topicRef xlink:href="#H"/>
                </instanceOf>
        <baseName>
          <baseNameString>Database Management (H.2H)
          </baseNameString>
        </baseName>
        </topic>
```

APPENDIX IV. XML TECHNOLOGIES USED IN THE TRV PROTOTYPE

XML Component	What does it do?	Type of Standard (Conceptual Product; Implementation Product) and Source	How it was used in the TRV Prototype
MinML2	A minimal XML parser for small documents	Implementation Product – UK commercial venture	To parse multiple small XML documents for XTM
Xerces	A reliable full-featured parser	Implementation Product - Apache	Used by DOM and SAX to extract topics and to transform XML document from DC to BiblioML
Xalan	XSLT processor for transforming XML documents into HTML	Implementation Product - Apache	Used to convert XML documents to HTML
XTM	XML definition and interchange of Topic Maps	Conceptual Product – an official ISO/IEC standard	Used to represent the ACM classification tree for the Java applet hyperbolic visualizer
DOM	Defines a programmatic interface to process smaller XML documents	Conceptual Product – W3C	Used to access and extract data from smaller XML documents
SAX	Defines a programmatic interface to process larger XML documents	A de facto Conceptual Product; individually developed by David Meggison and turned over to the open source community.	Used by the harvester program to process the large OAI XML document
JDOM	Defines DOM and SAX for JAVA programmers	A de facto Implementation Product issued under the Open Source Apache license; invented by Brett McLaughlin and Jason Hunter in the Spring of 2000	Used in the ArXiv harvester program
XSL-FO	A language for reformatting XML	Conceptual Product - W3C	Used by Xalan to format the HTML document
XSLT	A language for transforming XML	Conceptual Product – W3C	To convert Dublin Core to BiblioML
Xpath	An expression language used by XSLT to refer to parts of XML document	Conceptual Product – W3C	Used in XSLT code
Xlink	XML elements to create and describe links between resources.	Conceptual Product – W3C	Used in XTM Classification Tree to define child-parent relationships.

APPENDIX V. ARCHITECTURE OF THE TRV2 SYSTEM



ABOUT THE AUTHOR

Mark Ginsburg is Assistant Professor of MIS at the Eller College of Business, University of Arizona. His research interests are in the areas of collaborative computing, visualization of large-scale digital libraries, socio-technical issues of virtual communities, and practical problems pertaining to document management, knowledge management, and e-business strategies. His research articles appear in Communications of ACM, J. CSCW, IEEE Computer, J. Autonomous Agents and Multi-Agent Systems, and Electronic Markets Journal. From 1994-6, he was the lead developer on the flagship EDGAR on the Internet project; the first large-scale dissemination of government data (corporate disclosure statements) to the general public. He is a member of IEEE, ACM, and SIGGROUP.

Communications of the Association for Information Systems

ISSN: 1529-3181