Beyond Digital Incunabula: Modeling the Next Generation of Digital Libraries*

Gregory Crane, David Bamman, Lisa Cerrato, Alison Jones, David Mimno, Adrian Packel, David Sculley, and Gabriel Weaver

Perseus Project, Tufts University

Abstract. This paper describes several incunabular assumptions that impose upon early digital libraries the limitations drawn from print, and argues for a design strategy aimed at providing customization and personalization services that go beyond the limiting models of print distribution, based on services and experiments developed for the Greco-Roman collections in the Perseus Digital Library. Three features fundamentally characterize a successful digital library design: finer granularity of collection objects, automated processes, and decentralized community contributions.

1 Introduction

Potentially massive digital libraries such as the Google Library Project [17], the Open Content Alliance [20] and the EU i2010 Initiative [21] emphasize high volume digitization based primarily on automatically generated output of page images. While digitizing page images should be the first step in any digitization project and must comprise the core of any million book library, we need as well a sound infrastructure to augment our ability to search, browse and analyze the collection. Too great an emphasis on quantity can reinforce usage models that perpetuate limits from print distribution. This paper argues for a more aggressive, but still conservative, design strategy aimed at providing customization and personalization services that go beyond limiting models based on print distribution. While customization involves the user making explicit choices about the interface or system they are using, personalization involves the system adapting itself automatically to the user [35]. We base our argument on existing services and experiments developed for the Greco-Roman collections in the Perseus Digital Library [8, 6, 7]. The underlying methods are broader in application, have contributed to work completed for the National Science Digital Library [19], and lay the foundation for a range of services in Fedora and other digital repositories [27].

This paper has the following components. First, it describes several incunabular assumptions that impose upon early digital libraries the limitations drawn from print [34]. Second, it describes three features that fundamentally characterize emergent digital libraries. Third, it provides examples of customization and personalization built upon these three features.

2 Incunabular Models of Early Digital Libraries

New media begin by solving well known problems but also by imitating the forms that precede them. Consider three habits of thought drawn from paper libraries that constrain the design of

* The U.S. National Endowment for the Humanities (NEH) and the National Science Foundation provided primary support for this work under the Digital Library Initiative, Phase 2. Support from the Institute for Museum and Library Services also contributed substantially to this project.

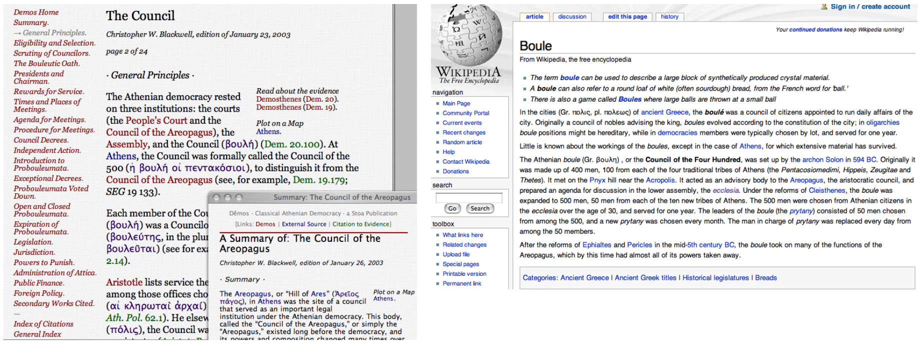
digital libraries. First, academic publications are based on coarse chunks of information, usually PDF or HTML files with heavily structured data (e.g., section headers, notes, bibliography). Prestructured documents perpetuate the primacy of the author, leaving readers to do what they can with the structure and content that the author has chosen to include. Second, the emphasis on metadata carries forward the card catalogue of the print library. While metadata is important, metadata repositories that do not also include content are of limited use and have imposed an elegant modularity that constrains, as much as it has enhanced, intellectual life [26]. Third, print libraries are static. We may replace the books on the shelves with new editions, but those old books do not generate new versions of themselves. Print libraries cannot learn about their holdings and generate new content on their own. Information retrieval systems, which reindex libraries as they grow, constitute only a partial step in this direction, for they do not cycle over and over generating new knowledge by learning from their collections and from their users.

Hand-crafted collections, their contents marked up according to XML DTD/schemas with RDF metadata, incorporate major advantages over, but still narrowly replicate, their print antecedents. Some projects have, however, begun to move beyond these limitations. Figure 1 (left) is drawn from the most current and best documented survey of Athenian democracy now available: an electronic publication called Demos, published on-line as part of the Stoa publishing consortium [18].

First, Demos combines traditional and emerging structural designs: it can be downloaded as PDF chapters resembling conventional publications, but it is also available as logical chunks, each containing a semantically significant unit of text, providing greater precision than chapter heading and greater accuracy than page numbers in a book.

Second, Blackwell based a densely hypertextual work on an academic digital library that contained most of the primary evidence about Athenian democracy. The digital library shapes the form of Demos. Every major statement contains links to the primary evidence: where print reference works avoid visual clutter and save space, style sheets can turn these links on and off in a digital publication. More substantively, documented writing diminishes authorial claims of authority, offering readers an opportunity to compare conclusions with their evidence and enabling discussion. Unlike their print antecedents, citations in a digital library can point not only to pre-existing documents but also to services such as morphological analysis of Greek words or the plotting of geographic locations on high-resolution maps. The author combines links to the digital library and contextualizing materials within Demos (one of which is pictured in the figure). These internal links include discussions of the primary sources and the issues that they raise. Demos does not, however, directly address the model of the static library: each chunk of Demos lists its last modification dates, and each date testifies to the fact that Demos, for all its strengths, is not changing.

Figure 1 (right) shows the Wikipedia article on the Athenian council (boule) as it appears on March 2, 2006 [14]. Taken together, these demonstrate the possibilities of both existing and emergent digital libraries. While the Demos discussion of the boule is rigorous, it is also dated (January 23, 2003) and grows slowly out of date with each passing day. Second the Demos article reflects the synthesis of a single author interacting with a small editorial community, and such a publication method could omit important perspectives from an authoritative discussion. The Wikipedia article, by contrast, is subject to constant modification. Wikipedia can thus capture broader perspectives and remain current if opinion shifts [5]. The Wikipedia article, however, contains no source citations: modeled on contemporary encyclopedias and reference works and their relatively superficial bibliographic apparatus, Wikipedia articles contain high statement/evidence ratios. They thus claim credibility, and the lack of systematic pointers to evidence is problematic. If every statement were linked to its source, gross misrepresentations would be much more readily identified and corrected.


Demos Home
Summary.
→ General Principles.
Eligibility and Selection.
Scrutiny of Councilors.
The Bouleutic Oath.
Presidents and Chairman.
Rewards for Service.
Times and Places of Meetings.
Agenda for Meetings.
Procedure for Meetings.
Council Decrees.
Independent Action.
Introduction to Probouleumata.
Exceptional Decrees.
Probouleumata Voted Down.
Open and Closed Probouleumata.
Expiration of Probouleumata.
Legislation.
Jurisdiction.
Powers to Punish.
Administration of Attica.
Public Finance.
Foreign Policy.
Secondary Works Cited.
—
Index of Citations
General Index

The Council

Christopher W. Blackwell, edition of January 23, 2003

page 2 of 24

· General Principles ·

Read about the evidence
Demosthenes (Dem. 20).
Demosthenes (Dem. 19).

Plot on a Map
Athens.

The Athenian democracy rested on three institutions: the courts (the People's Court and the Council of the Areopagus), the Assembly, and the Council (βουλή) (Dem. 20.100). At Athens, the Council was formally called the Council of the 500 (ἡ βουλὴ οἱ πεντακόσιοι), to distinguish it from the Council of the Areopagus (see, for example, Dem. 19.179; *SEG* 19 133).

Each member of the Cou
(βουλή) was a Councilo
(βουλεύτης, in the plur
βουλεῦται) (see for exa
2.14).

Aristotle lists service the
among those offices cho
(αἱ κληρωταὶ ἀρχαί)
Ath. Pol. 62.1). He elsew
(πόλις), the Council wa

Summary: The Council of the Areopagus

Dêmos · Classical Athenian Democracy · a Stoa Publication
[Links: Demos | External Source | Citation to Evidence]

A Summary of: The Council of the Areopagus

Christopher W. Blackwell, edition of January 26, 2003

· Summary ·

Plot on a Map
Athens.

The Areopagus, or "Hill of Ares" (Ἄρεῖος πάγος), in Athens was the site of a council that served as an important legal institution under the Athenian democracy. This body, called the "Council of the Areopagus," or simply the "Areopagus," existed long before the democracy, and its powers and composition changed many times over

Sign in / create account

article | discussion | edit this page | history

Your ***continued donations*** *keep Wikipedia running!*

WIKIPEDIA
The Free Encyclopedia

navigation
- Main Page
- Community Portal
- Current events
- Recent changes
- Random article
- Help
- Contact Wikipedia
- Donations

search
Go | Search

toolbox
- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link

Boule

From Wikipedia, the free encyclopedia

- *The term* ***boule*** *can be used to describe a large block of synthetically produced crystal material.*
- *A* ***boule*** *can also refer to a round loaf of white (often sourdough) bread, from the French word for 'ball.'*
- *There is also a game called* ***Boules*** *where large balls are thrown at a small ball*

In the cities (Gr. πολις, pl. πολεως) of ancient Greece, the ***boulé*** was a council of citizens appointed to run daily affairs of the city. Originally a council of nobles advising the king, *boules* evolved according to the constitution of the city; in oligarchies *boule* positions might be hereditary, while in democracies members were typically chosen by lot, and served for one year.

Little is known about the workings of the *boules*, except in the case of Athens, for which extensive material has survived.

The Athenian *boule* (Gr. βουλη) , or the **Council of the Four Hundred**, was set up by the archon Solon in 594 BC. Originally it was made up of 400 men, 100 from each of the four traditional tribes of Athens (the *Pentacosiomedimi*, *Hippeis*, *Zeugitae* and *Thetes*). It met on the Pnyx hill near the Acropolis. It acted as an advisory body to the Areopagus, the aristocratic council, and prepared an agenda for discussion in the lower assembly, the *ecclesia*. Under the reforms of Cleisthenes, the *boule* was expanded to 500 men, 50 men from each of the ten new tribes of Athens. The 500 men were chosen from Athenian citizens in the *ecclesia* over the age of 30, and served for one year. The leaders of the *boule* (the *prytany*) consisted of 50 men chosen from among the 500, and a new *prytany* was chosen every month. The man in charge of *prytany* was replaced every day from among the 50 members.

After the reforms of Ephialtes and Pericles in the mid-5th century BC, the *boule* took on many of the functions of the Areopagus, which by this time had almost all of its powers taken away.

Categories: Ancient Greece | Ancient Greek titles | Historical legislatures | Breads

Fig. 1. (Left) A page from Demos: Classical Athenian Democracy, ed. Christopher Blackwell. (Right) Wikipedia article on the Athenian Boule.

3 Three Features that Characterize Post-Incunabular Digital Libraries

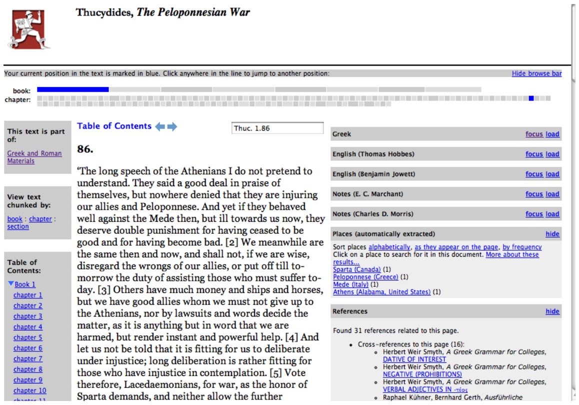
If we combine the scholarly rigor of Demos with the self-organizing qualities of Wikipedia, we can begin to see emerging a new model not only for digital libraries but also for the disciplined intellectual discourse which digital libraries should support. At least, three strategic functions distinguish emerging digital libraries from their predecessors.

1. Finer granularity: while many documents (like this paper) cite other publications as chunks, users often do not want to plough through an entire information object but would rather use an overview or particular sub-objects (proposition, bibliographic references, etc.) [37]. As digital libraries evolve, these structures go beyond those implicit in traditional publication models and begin to include explicit propositional statements.
2. Autonomous learning: Improved granularity implies this second fundamental characteristic of emergent digital libraries. Digital libraries should be constantly learning and becoming more intelligent as they scan their contents – a phenomenon already visible in rudimentary form with existing search engines. In a model digital library, the articles should be constantly scanning for new secondary sources, new editions of the primary materials and, to the extent possible, shifts in language that suggest perspectives that differ from the content of the current document. Thus documents and their sub-components should be in constant, autonomous communication with the libraries of which they are a part, scrutinizing new materials submitted and rereading the rest of the collection as automated processes evolve [39].
3. Decentralized, real-time community contributions: Wikipedia presents one of the most important practical experiments in the history of publication. For all the criticism that it may deserve, the English language Wikipedia has generated more than 1,000,000 entries in five years and supports several million queries a day. If we go beyond the higher level prose and examine individual verifiable propositions, the accuracy is remarkable. A recent study of relational statements in Wikipedia demonstrated that 97.22% of basic propositional statements and 99.47% of disambiguating links prove to be correct [38]. Thus, even if we reject the expository prose of Wikipedia for bias, the propositions within Wikipedia demonstrate that decentralized communities will accumulate immense amounts of highly accurate propositional data [28].

The following sections describe concrete, if rudimentary, steps Perseus has taken toward all three of the above principles, and addresses the issues that digital libraries in general must confront in order to transcend the limitations of print distribution.

3.1 Granularity

Cultural heritage documents often have complex contents that do not lend themselves to simple hierarchical representation [3]. While modern publications pre-define form and structure for the sake of simplifying system design, cultural heritage documents often have multiple, overlapping hierarchies (e.g., Thucydides' *History of the Peloponnesian War*, for instance, can be organized by book/chapter or by alternations of third person narrative and speeches). Digital libraries need to be able to address these various parts of a document that users want to work with. Figure 2 illustrates a dynamically generated report on what the Perseus DL knows about a particular chapter in Thucydides (Thuc. 1.86).



Thucydides, *The Peloponnesian War*

Your current position in the text is marked in blue. Click anywhere in the line to jump to another position:

This text is part of:
Greek and Roman Materials

View text chunked by:
book : chapter : section

Table of Contents:
Book 1
chapter 1
chapter 2
chapter 3
chapter 4
chapter 5
chapter 6
chapter 7
chapter 8
chapter 9
chapter 10

Thuc. 1.86

86.

'The long speech of the Athenians I do not pretend to understand. They said a good deal in praise of themselves, but nowhere denied that they are injuring our allies and Peloponnese. And yet if they behaved well against the Mede then, but ill towards us now, they deserve double punishment for having ceased to be good and for having become bad. [2] We meanwhile are the same then and now, and shall not, if we are wise, disregard the wrongs of our allies, or put off till to-morrow the duty of assisting those who must suffer to-day. [3] Others have much money and ships and horses, but we have good allies whom we must not give up to the Athenians, nor by lawsuits and words decide the matter, as it is anything but in word that we are harmed, but render instant and powerful help. [4] And let us not be told that it is fitting for us to deliberate under injustice; long deliberation is rather fitting for those who have injustice in contemplation. [5] Vote therefore, Lacedaemonians, for war, as the honor of Sparta demands, and neither allow the further

Greek

English (Thomas Hobbes)

English (Benjamin Jowett)

Notes (E. C. Marchant)

Notes (Charles D. Morris)

Places (automatically extracted)

Sort places alphabetically, as they appear on the page, by frequency
Click on a place to search for it in this document. More about these results...
Sparta (Canada) (1)
Peloponnese (Greece) (1)
Mede (Italy) (1)
Athens (Alabama, United States) (1)

References

Found 31 references related to this page.

- Cross-references to this page (16):
  - Herbert Weir Smyth, *A Greek Grammar for Colleges*, DATIVE OF INTEREST
  - Herbert Weir Smyth, *A Greek Grammar for Colleges*, NEGATIVE (PROHIBITIONS)
  - Herbert Weir Smyth, *A Greek Grammar for Colleges*, VERBAL ADJECTIVES IN -τέος
  - Raphael Kühner, Bernhard Gerth, *Ausführliche*

Fig. 2. Information about Thucydides, *History of the Peloponnesian War*, book 1, chapter 86, with user focus upon one of several translations. The right hand of the screen illustrates other dynamically collected resources extracted from digital objects and organized into an ontology of document types.

First, information display depends upon an authority list of meaningful citations: "Thuc. 1.86" provides a common designation with which references to Thucydides are aligned. Using this citation, we can identify which digital objects mention this particular chunk of text. Second, the individual passages that cite this passage have also been classified. Using this classification, we can distinguish Greek source texts from English translations and annotations specifically about Thuc. 1.86, and both from texts that only mention this passage in passing. Third, all documents in the collection have been broken down into structural units. Thus, we can extract multiple Greek versions or foreign language translations of the precise chunk designated by Thuc. 1.86. This allows us to identify not only that Thuc. 1.86 shows up in a particular Greek grammar, for instance, but that it also appears specifically in the discussion on "dative of interest."

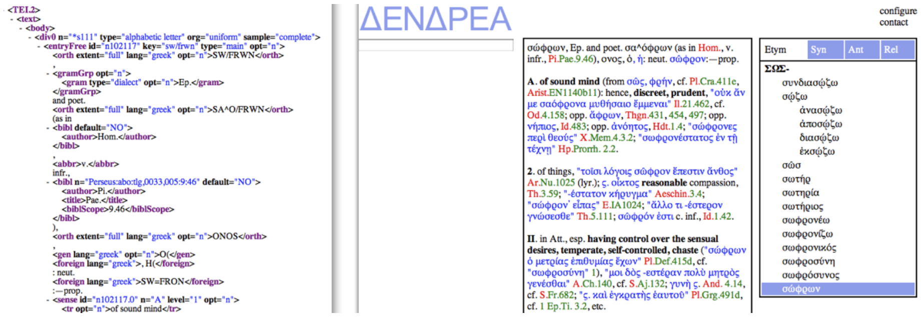
Such fine grained chunking can transform the value of information: the main Greek-English lexicon entry on the Greek preposition "pros" mentions Thuc. 1.86, but few readers would scan all fifty-two senses for the particular sense that cites Thuc. 1.86. Because we have captured the tree-structure of the dictionary, we know that this citation occurs at the third level down (as sense C.I.6), and we can extract this precise chunk from the much larger article. The right hand column aggregates and organizes these citations into a single report that could, in turn, be filtered and personalized for particular users.

Automatically organized reports on chunks of text (or museum objects, archaeological sites or other entities) build upon well structured documents which were designed for reference and subsequent citation. In order to address the complexity of cultural heritage documents that do not lend themselves to such representation, digital libraries need to confront the following issues, none of which are glamorous but each of which demands resolution:

1. Consistent markup for complex documents: For all of the progress that has been made, we do not have large, interoperable, richly structured documents in TEI or any other markup. Capturing chapters, sections, notes and similar well-defined elements is not the problem. Rather, we need consistent ways of managing documents within documents. Some carefully marked up collections (such as DocSouth [13] and American Memory [15]) may contain indices with accurate transcriptions of citations but have not included markup that captures the structure of the index or expands often idiosyncratic abbreviations into machine actionable data.
2. Mechanisms, automated and semi-automated, to identify meaningful chunks of individual documents: This desideratum addresses the need to generate large quantities of markup scalably as digital libraries with millions of books emerge: tables, notes, address/sender/ receiver/dateline tags for embedded letters and documents, indices, etc. are difficult to extract even from well transcribed documents. Such tools need to support both markup (e.g., where they add valid elements to existing structures) and extraction (where TEI texts might provide the source for generating external ontologies). Pilot projects integrating information extraction into digital libraries have begun to emerge (GATE/Greenstone), but developed solutions will be complex for large, heterogeneous collections and much needs to be done [40].
3. Digital library systems that can represent complex documents: We need as a starting point systems that support multiple, overlapping hierarchies within the same document [9]. We also need to be able to manage partial structures (e.g., browse all quoted speech, excerpts of poetry, personal names). Digital libraries need to model complex documents early in their development rather than concentrating on structurally simpler data types.
4. Conventions for exchanging complex content: Even if we address the first three issues, we still need conventions whereby we can exchange and recombine chunks of data from multiple collections [23]. These conventions include citation schemes, standard credit lines for author, institutions and funders, and an infrastructure for redundant, robust access in the present and for preservation into the future. Figure 3 illustrates an initial version of such a service, with a Perseus dictionary entry delivered as XML fragment (left) and a third party (Dendrea) representation of that data (right), with services added that are not currently present in the home digital library of the dictionary.

3.2 Automated Processes

While granularity can give us the ability to bring together related sections from different pre-existing documents, automatic analysis gives digital libraries the opportunity to create entirely new documents rather than quietly wait for new acquisitions.



```
<TEI.2>
- <text>
  - <body>
    - <div0 n="*s111" type="alphabetic letter" org="uniform" sample="complete">
      - <entryFree id="n102117" key="sw/frwn" type="main" opt="n">
          <orth extent="full" lang="greek" opt="n">SW/FRWN</orth>
          ,
        - <gramGrp opt="n">
            <gram type="dialect" opt="n">Ep.</gram>
          </gramGrp>
          and poet.
          <orth extent="full" lang="greek" opt="n">SA^O/FRWN</orth>
          (as in
        - <bibl default="NO">
            <author>Hom.</author>
          </bibl>
          ,
          <abbr>v.</abbr>
          infr.,
        - <bibl n="Perseus:abo:tlg,0033,005:9:46" default="NO">
            <author>Pi.</author>
            <title>Pae.</title>
            <biblScope>9.46</biblScope>
          </bibl>
          ),
          <orth extent="full" lang="greek" opt="n">ONOS</orth>
          ,
          <gen lang="greek" opt="n">O(</gen>
          <foreign lang="greek">, H(</foreign>
          : neut.
          <foreign lang="greek">SW=FRON</foreign>
          :—prop.
        - <sense id="n102117.0" n="A" level="1" opt="n">
            <tr opt="n">of sound mind</tr>
```

σώφρων, Ep. and poet. σα^όφρων (as in Hom., v. infr., Pi.Pae.9.46), ονος, ὁ, ἡ: neut. σῶφρον:—prop.

A. **of sound mind** (from σῶς, φρήν, cf. Pl.Cra.411e, Arist.EN1140b11): hence, **discreet, prudent**, "οὐκ ἄν με σαόφρονα μυθήσαιο ἔμμεναι" Il.21.462, cf. Od.4.158; opp. ἄφρων, Thgn.431, 454, 497; opp. νήπιος, Id.483; opp. ἀνόητος, Hdt.1.4; "σώφρονες περὶ θεούς" X.Mem.4.3.2; "σωφρονέστατος ἐν τῇ τέχνῃ" Hp.Prorrh. 2.2.

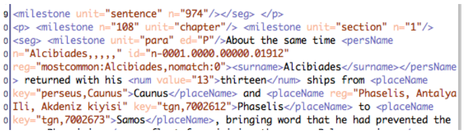
2. of things, "τοῖσι λόγοις σῶφρον ἔπεστιν ἄνθος" Ar.Nu.1025 (lyr.); ς. οἶκτος **reasonable** compassion, Th.3.59; "-έστατον κήρυγμα" Aeschin.3.4; "σώφρον᾽ εἶπας" E.IA1024; "ἄλλο τι -έστερον γνώσεσθε" Th.5.111; σῶφρόν ἐστι c. inf., Id.1.42.

II. in Att., esp. **having control over the sensual desires, temperate, self-controlled, chaste** ("σώφρων ὁ μετρίας ἐπιθυμίας ἔχων" Pl.Def.415d, cf. "σωφροσύνη" 1), "μοι δὸς -εστέραν πολὺ μητρὸς γενέσθαι" A.Ch.140, cf. S.Aj.132; γυνὴ ς. And. 4.14, cf. S.Fr.682; "ς. καὶ ἐγκρατὴς ἑαυτοῦ" Pl.Grg.491d, cf. 1 Ep.Ti. 3.2, etc.

ΣΩΣ-
συνδιασῴζω
σῴζω
ἀνασῴζω
ἀποσῴζω
διασῴζω
ἐκσῴζω
σῶσ
σωτήρ
σωτηρία
σωτήριος
σωφρονέω
σωφρονίζω
σωφρονικός
σωφροσύνη
σωφρόσυνος
σώφρων

Fig. 3. (Left) A well-formed fragment of XML representing an entry from the Liddell Scott Jones lexicon served over the Web. This service supports third party added-value services that exploit the rich underlying structure. Note that this document has a carefully captured structure, with each sense definition possessing a unique identification number. Thus, individual senses can be precisely extracted and reused in third-party hybrid documents. (Right) A third-party representation of the same dictionary article. Dendrea.org not only provides a different front end but also information extraction services that scan for etymological data, related words, and semantic relationships such as synonymy and antonymy.

Digital libraries need to include a range of automated processes that add value to their collections, including both classification (matching pre-determined patterns) and data mining (discovering new patterns) [1, 25]. While fields such as machine learning and data mining are major topics in their own right, integrating such technologies into digital libraries raises a range of problems [11, 10]. We need systems that can draw upon the contents of the digital library, continually applying new knowledge sources (e.g., gazetteers, machine readable dictionaries) as these become available, recalculating its results and assessing its performance [31, 41].



```
<milestone unit="sentence" n="974"/></seg> </p>
<p> <milestone n="108" unit="chapter"/> <milestone unit="section" n="1"/>
<seg> <milestone unit="para" ed="P"/>About the same time <persName
n="Alcibiades,,,,," id="n-0001.0000.00000.01912"
reg="mostcommon:Alcibiades,nomatch:0"><surname>Alcibiades</surname></persName
> returned with his <num value="13">thirteen</num> ships from <placeName
key="perseus,Caunus">Caunus</placeName> and <placeName reg="Phaselis, Antalya
Ili, Akdeniz kiyisi" key="tgn,7002612">Phaselis</placeName> to <placeName
key="tgn,7002673">Samos</placeName>, bringing word that he had prevented the
```

Fig. 4. Named entity analysis for classical texts. Notice the lack of culturally appropriate markup, with the TEI SURNAME tag used problematically to capture the primary name, as we begin adapting instruments for modern sources to Greco-Roman documents.

Not only do we need scalable methods to identify the semantically significant document chunks such as tables, embedded letters, and notes, we need more ways to analyze raw text and classify it as propositional knowledge, bibliographic citations, quotations, and named entities (e.g., is "London" a person or a place, and, which person or place?). People, places, dates and organizations are fundamental data which any mature digital library must track. Figure 4 displays the results of a named entity recognition system in place at Perseus since 2000. In an excerpt from Thucydides 8.108 ("About the same time Alcibiades returned with his thirteen ships from Caunus and Phaselis to Samos, bringing word ..."), "Alcibiades" has been automatically annotated as a person, and "Caunus," "Phaselis" and "Samos" have been annotated as places, two of them matched with the Getty Thesaurus of Geographic Names via TGN identification numbers.

Automated analysis also includes multi-lingual services such as cross-language information retrieval and machine translation [30, 12]. Identifying the fundamental meanings of a word is a notoriously slippery problem – human lexicographers do not agree among themselves as to what constitutes a separate word sense. One pragmatic approach involves examining translation equivalents: where translators use distinct words in the translation language, we have evidence of a substantive different meaning. Thus Table 1 displays one cluster of word meanings for the polysemous Greek word *arche*, derived from comparison of a Greek source text and five separate English translations. *Arche* can mean "empire," "government," "political office," and "beginning"; by grouping together the words that occur around it, we are able to use translations to identifyits intended sense. The six texts as a whole are aligned according to the standard citation scheme rather than at the word or sentence level (c. 42 Greek words per chunk), with sections themselves in four of the five English translations automatically aligned with the Greek original. The experiment thus explores what can be done with translations of canonical texts, with minimal extra tagging, that will populate large emerging digital libraries.

Table 1. Parallel text analysis: word clusters associated with uses of the Greek word *arche* in Thucydides (c. 150,000 words) and five English translations. Translation equivalents are underlined. The clusters capture the senses "empire," "government," "political office," and "beginning." The cluster headed "ancient" (marked in bold) captures a distinct word that happens to share the stem *arch-*.

empire	dominion	power	government
office	government	magistrates	people
dominion	power	rule	Hellenes
magistrates	Theseus	people	council
ancient	descendants	temples	Pythian
whom	beginning	pits	just
called	Zancle	Pangaeus	originally

The translation analysis points to four elements of text mining relevant to digital libraries. First, this function will improve as digital libraries grow larger, because we will have access to more and more translations of source texts into a range of languages. Second, the clustering of word groups is computationally intensive and the current algorithm is not suited to providing real time results. Third, while such exploratory data may not begin as part of the general digital library, the results of such analysis, once generated, may become a domain specific service available to those reading Greek (or similar languages for which this service is suitable). We may well find domain specific front ends, integrating data from several larger collections into a new hybrid information space designed for particular communities. Fourth, the output of the parallel text analysis is useful in its own right to human analysts, but this output also provides a foundation for cross-language information retrieval, machine translation and other multi-lingual services.

3.3 User Contributions

Every large collection contains errors and, while these are finite and may be corrected in a finite period of time, by the time original errors may be fully corrected, scholarship will have marched on, rendering bibliography, front matters, and annotations in need of revision. While some automated processes do approach perfection, even these still generate errors, and most automated processes have error rates far removed from 100%. Some processes (such as assigning a sense to a given instance of a word or analyzing the syntax of a sentence) will yield disagreement among experts.

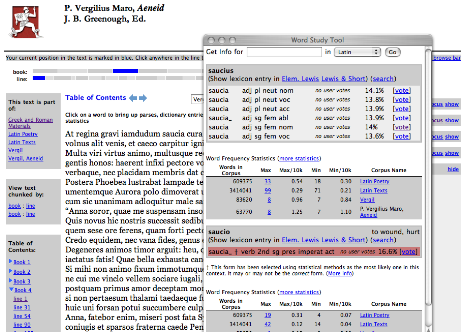
We need to consider mechanisms to collect information from our user communities [33, 4]. In some cases, the amateurs will probably perform better than the academics: professional historians may chafe at genealogists fixated on precise identifications of people, places and organizations or antiquarians fretting about the precise buttons a particular person might have worn, but such attention to detail can be of immense value when we are refining the results of our collections.

There are two categories of contribution. First, we need annotations and reference articles that incorporate at least some full text. Wikipedia has demonstrated both immense strengths and troubling weaknesses in this area. More conservative efforts such as PlanetMath [16], based on Noosphere, have arguably produced more consistent results, but they are much more focused efforts and have created around 5,000 encyclopedia entries rather than the 1,000,000 in Wikipedia [24]. The NEH has funded "Pleaides: An Online Workspace for Ancient Geography," which will explore community created scholarly content [22]. The Perseus DL will include Wiki pages for every data object and every addressable text chunk. Our optimistic hypothesis is that community driven commentaries, created by dedicated amateurs struggling to understand the texts, may prove more useful than commentaries produced by experts: we expect many errors at first but that the churning of user responses will weed out mistakes and that Wiki commentaries will evolve into accurate instruments. We are less concerned with the potential errors than with whether such Wiki commentaries will attract a critical mass of contributors.

Second, we need to collect user feedback on propositional data: e.g., whether "London" in passage X is London, Ontario, rather than London, England; whether "saucia" is a nominative singular adjective rather than ablative; whether a certain dative is an indirect object of the verb rather than a dative of possession with a nearby noun. Propositional data does not always have a single, clearly correct answer, but we can collect alternatives and store them in a structured format.

We created two initial mechanisms to collect propositional user input, allowing users to match a particular word sense and morphological analysis appropriate to a given word in a given passage. Figure 5 illustrates the results of three processes. First, a morphological analyser generates an exhaustive list of possible analyses for the form "saucia." Second, another module examines the immediate context and the relative frequency of the forms and possible dictionary entries (if the word is lexically ambiguous), then calculates probabilities for each alternative analsysis. Accuracy of the automated disambiguation stands at 76%. Third, users can cast votes of their own, whether to reinforce the automatic analysis or to suggest an alternative.

As of March 7, 2006, users have cast 7,597 votes to disambiguate Greek and Latin words with more than one possible morphological analysis. The overall accuracy for individual voting stands at 89%, but the improvement over the performance of the automated disambiguation is substantially higher, since users overwhelmingly vote on words for which the system has assigned the wrong analysis. (While the overall accuracy of automatic disambiguation is 76%, its accuracy on words that users vote on is only 34%.) And since 43% of word tokens have only one morphological sense (and do not need therefore to be disambiguated), user voting with 89% accuracy on ambiguous forms has the potential to deliver an overall system with 93.7% accuracy if every ambiguous word



Your current position in the text is marked in blue. Click anywhere in the line t

Click on a word to bring up parses, dictionary entrie
statistics

At regina gravi iamdudum saucia cura
volnus alit venis, et caeco carpitur igni
Multa viri virtus animo, multusque re
gentis honos: haerent infixi pectore vo
verbaque, nec placidam membris dat c
Postera Phoebea lustrabat lampade te
umentemque Aurora polo dimoverat u
cum sic unanimam adloquitur male sa
“Anna soror, quae me suspensam inso
Quis novus hic nostris successit sedibu
quem sese ore ferens, quam forti pect
Credo equidem, nec vana fides, genus
Degeneres animos timor arguit: heu, c
iactatus fatis! Quae bella exhausta can
Si mihi non animo fixum immotumque
ne cui me vinclo vellem sociare iugali,
postquam primus amor deceptam mo
si non pertaesum thalami taedaeque f
huic uni forsan potui succumbere culp
Anna, fatebor enim, miseri post fata S
coniugis et sparsos fraterna caede Pen

saucius
(Show lexicon entry in Elem. Lewis Lewis & Short) (search)

saucia	adj pl neut nom	*no user votes*	14.1%	[vote]
saucia	adj pl neut voc	*no user votes*	13.8%	[vote]
saucia	adj pl neut acc	*no user votes*	13.9%	[vote]
saucia_	adj sg fem abl	*no user votes*	13.9%	[vote]
saucia	adj sg fem nom	*no user votes*	14%	[vote]
saucia	adj sg fem voc	*no user votes*	13.6%	[vote]

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
609375	33	0.54	18	0.30	Latin Poetry
3414041	99	0.29	71	0.21	Latin Texts
83620	8	0.96	7	0.84	Vergil
63770	8	1.25	7	1.10	P. Vergilius Maro, Aeneid

saucio to wound, hurt
(Show lexicon entry in Elem. Lewis Lewis & Short) (search)

saucia_ †	verb 2nd sg pres imperat act	*no user votes*	16.6%	[vote]

† This form has been selected using statistical methods as the most likely one in this context. It may or may not be the *correct* form. (More info)

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
609375	19	0.31	4	0.07	Latin Poetry
3414041	42	0.12	14	0.04	Latin Texts

Fig. 5. System to register votes against machine generated choice of correct morphological analysis: heuristic driven morphological analysis has been a long-term service in the Perseus DL, but we have only recently applied machine learning to estimate the correct morphological analyses for a given word in a given passage. Users can vote for alternative analyses if they disagree with the automatic choice. This system has been adopted to evaluate unfiltered, anonymous contributions, providing a possible baseline for more demanding models.

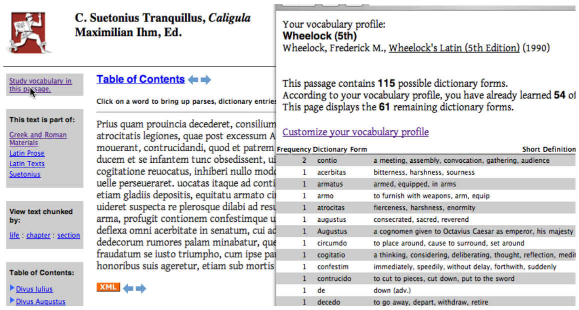
receives one vote. We could solicit expert users to fill in the remaining accuracy gap, but a better solution may simply be to focus on enlarging the contributor base: the more individual votes per word, the more likely that all, when taken together, will be correct.

4 User-Centered Digital Libraries: Customization and Personalization

The three incunabular constraints of early digital libraries all act as much against as for the user. The catalogue model tells users what exists and sometimes points them to on-line versions of the source object, but then its job is done and the users must do what they can with what they find. The digital codex model may incorporate searching and convert citations to links, but the author creates fixed content and structure around which users must work. The static library can learn neither on its own nor from its users. Early digital libraries thus, not surprisingly, replicate the hegemony of library, author, and publisher.

Digital libraries can, however, shift the balance further toward the user and toward the active life of the mind. More finely grained data objects, automated processes and decentralized user contributions all should interact, with the digital library progressively growing better structured and more self-aware. Initial data structures seed automated processes which classify and mine data. Users evaluate classification results and feed their contributions back into the system. Data mining suggests new patterns, which in turn complement or revise previous schemes, leading to the discovery and classification of new structures within the same set of digital objects.

Two fundamental strategies should be available to the user. First, users should be able to customize the environment [2]. Such customization needs to reflect not only default page layouts and simple preferences but also much more elaborate models of user knowledge. Figure 6 shows a customized report for a user who has studied Latin with a particular textbook (one of about thirty Greek and Latin textbooks whose vocabularies we have modeled on a chapter by chapter basis). Multiple readers with different backgrounds can thus see in a given passage which terms are likely



C. Suetonius Tranquillus, *Caligula*
Maximilian Ihm, Ed.

Study vocabulary in this passage.

This text is part of:
Greek and Roman Materials
Latin Prose
Latin Texts
Suetonius

View text chunked by:
life : chapter : section

Table of Contents:
Divus Iulius
Divus Augustus

Table of Contents

Click on a word to bring up parses, dictionary entries

Prius quam prouincia decederet, consilium
atrocitatis legiones, quae post excessum A
mouerant, contrucidandi, quod et patrem
ducem et se infantem tunc obsedissent, ui
cogitatione reuocatus, inhiberi nullo modo
uelle perseueraret. uocatas itaque ad conti
etiam gladiis depositis, equitatu armato ci
uideret suspecta re plerosque dilabi ad resu
arma, profugit contionem confestimque u
deflexa omni acerbitate in senatum, cui ad
dedecorum rumores palam minabatur, que
fraudatum se iusto triumpho, cum ipse pa
honoribus suis ageretur, etiam sub mortis

XML

Your vocabulary profile:
Wheelock (5th)
Wheelock, Frederick M., Wheelock's Latin (5th Edition) (1990)

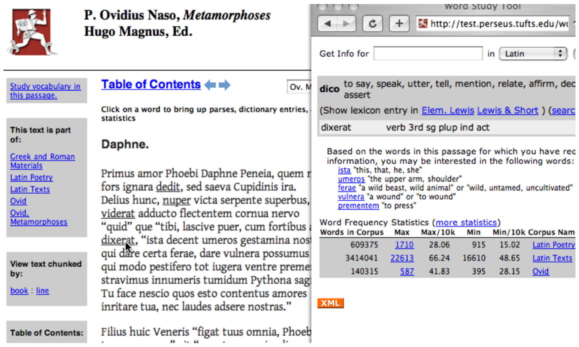
This passage contains **115** possible dictionary forms.
According to your vocabulary profile, you have already learned **54** of
This page displays the **61** remaining dictionary forms.

Customize your vocabulary profile

Frequency	Dictionary Form	Short Definition
2	contio	a meeting, assembly, convocation, gathering, audience
1	acerbitas	bitterness, harshness, sourness
1	armatus	armed, equipped, in arms
1	armo	to furnish with weapons, arm, equip
1	atrocitas	fierceness, harshness, enormity
1	augustus	consecrated, sacred, reverend
1	Augustus	a cognomen given to Octavius Caesar as emperor, his majesty
1	circumdo	to place around, cause to surround, set around
1	cogitatio	a thinking, considering, deliberating, thought, reflection, medit
1	confestim	immediately, speedily, without delay, forthwith, suddenly
1	contrucido	to cut to pieces, cut down, put to the sword
1	de	down (adv.)
1	decedo	to go away, depart, withdraw, retire

Fig. 6. Customized knowledge profile: the digital library knows the textbook with which the user has worked and analyzes probable known and unseen vocabulary in a given passage. Because the user has specified a profile and the system has responded, this is an example of customization.

to be novel and which they have encountered before. The fundamental principle can be applied to scientific terminology – which is, of course, easier to recognize than natural language.



P. Ovidius Naso, *Metamorphoses*
Hugo Magnus, Ed.

Study vocabulary in this passage.

This text is part of:
Greek and Roman Materials
Latin Poetry
Latin Texts
Ovid
Ovid, Metamorphoses

View text chunked by:
book : line

Table of Contents:

Table of Contents

Ov. M

Click on a word to bring up parses, dictionary entries, statistics

Daphne.

Primus amor Phoebi Daphne Peneia, quem n
fors ignara dedit, sed saeva Cupidinis ira.
Delius hunc, nuper victa serpente superbus,
viderat adducto flectentem cornua nervo
"quid" que "tibi, lascive puer, cum fortibus a
dixerat, "ista decent umeros gestamina nost
qui dare certa ferae, dare vulnera possumus
qui modo pestifero tot iugera ventre preme
stravimus innumeris tumidum Pythona sag
Tu face nescio quos esto contentus amores
inritare tua, nec laudes adsere nostras."

Filius huic Veneris "figat tuus omnia, Phoeb

Word Study Tool
http://test.perseus.tufts.edu/wc

Get Info for [ ] in Latin

dico to say, speak, utter, tell, mention, relate, affirm, dec assert
(Show lexicon entry in Elem. Lewis Lewis & Short) (searc
dixerat verb 3rd sg plup ind act

Based on the words in this passage for which you have req information, you may be interested in the following words:
ista "this, that, he, she"
umeros "the upper arm, shoulder"
ferae "a wild beast, wild animal" or "wild, untamed, uncultivated"
vulnera "a wound" or "to wound"
prementem "to press"

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Nam
609375	1710	28.06	915	15.02	Latin Poetry
3414041	22613	66.24	16610	48.65	Latin Texts
140315	587	41.83	395	28.15	Ovid

XML

Fig. 7. Personalized knowledge profile: the digital library does not know the background of the user but has analyzed four initial questions, compared these with past question patterns and suggested which of the remaining three hundred words the current user most likely to query. User behavior clusters into distinct classes and this approach has been able to predict 67% of subsequent queries. Because the system has taken the initiative rather than the user, this is an example of personalization.

Second, personalization augments the user-initiated decisions of customization: digital libraries should be able to analyze user behavior and background and offer new automatically generated configurations of information [36, 29, 32]. Figure 7 illustrates a recommender system that compares records of questions from earlier readers who had read a particular text. By mining past behaviors, the system can quickly learn to predict most of the subsequent questions that new users will pose.

5 Conclusion

Google with its massive library project, more recently Microsoft, Yahoo and others in the Open Content Alliance, and potentially the EU in its i2010 initiative are poised to assemble massive, but coarse, libraries of digitized books that have the potential to reinforce usage models based on print distribution. This paper provides initial examples of a post-incunabular design strategy utilized at the Perseus Project for its Greco-Roman collection but, we hope, scalable to other domains, focused on the principles of customization and personalization built upon fine grained digital objects, automated processes and decentralized user contributions.

References

[1] H. S. Baird, V. Govindaraju, and D. P. Lopresti. Document analysis systems for digital libraries: Challenges and opportunities. In *Document Analysis Systems*, pages 1–16, 2004.

[2] N. Beagrie. Plenty of room at the bottom? Personal digital libraries and collections. *D-Lib Magazine*, 11(6), 2005. http://dlib.anu.edu.au/dlib/june05/beagrie/06beagrie.html.

[3] J. Bradley. Documents and data: Modelling materials for humanities research in XML and relational databases. *Literary and Linguistic Computing*, 20(1), 2005.

[4] M.A.B. Burkard. Collaboration on medieval charters–Wikipedia in the humanities. In *Proceedings of the XVI International Conference of the Association for History and Computing*, pages 91–94, 2005.

[5] D. J. Cohen and R. Rosenzweig. Web of lies? Historical knowledge on the internet. *First Monday*, 10(12), Dec. 2005. http://www.firstmonday.org/issues/issue10_12/cohen/.

[6] G. Crane. Cultural heritage digital libraries: Needs and components. In *ECDL*, Rome, Italy, 16-18 Sept. 2002.

[7] G. Crane, R. F. Chavez, A. Mahoney, T. L. Milbank, J. A. Rydberg-Cox, D. A. Smith, and C. E. Wulfman. Drudgery and deep thought: Designing a digital library for the humanities. *Communications of the ACM*, 44(5):35–40, 2001.

[8] G. Crane, C. E. Wulfman, L. M. Cerrato, A. Mahoney, T. L. Milbank, D. Mimno, J. A. Rydberg-Cox, D. A. Smith, and C. York. Towards a cultural heritage digital library. In *JCDL*, pages 75–86, Houston, TX, June 2003.

[9] A. Dekhtyar, I. E. Iacob, J. W. Jaromczyk, K. Kiernan, N. Moore, and D. C. Porter. Support for XML markup of image-based electronic editions. *International Journal on Digital Libraries*, 6(1):55–69, Feb. 2006.

[10] J. S. Downie, J. Unsworth, B. Yu, D. Tcheng, G. Rockwell, and S. J. Ramsay. A revolutionary approach to humanities computing?: Tools development and the D2K data-mining framework. In *Annual Joint Conference of The Association for Computers and the Humanities & The Association for Literary and Linguistic Computing*, 2005.

[11] F. Esposito, D. Malerba, G. Semeraro, S. Ferilli, O. Altamura, T. M. A. Basile, M. Berardi, M. Ceci, and N. Di Mauro. Machine learning methods for automatically processing historical documents: From paper acquisition to XML transformation. In *DIAL*, volume 1, pages 328–335. IEEE Computer Society, 2004.

[12] F. C. Gey, N. Kando, and C. Peters. Cross-language information retrieval: the way ahead. *Information Processing and Management*, 41(3):415–431, 2005.

[13] http://docsouth.unc.edu/.

[14] http://en.wikipedia.org/wiki/Boule.

[15] http://memory.loc.gov/ammem/index.html.

[16] http://planetmath.org/.

[17] http://print.google.com/googleprint/library.html.

[18] http://seneca.stoa.org/projects/demos/article_council?page=1&greekEncoding=UnicodeC.

[19] http://www.nsdl.org.

[20] http://www.opencontentalliance.org/.

[21] http://www.theeuropeanlibrary.org/portal/index.htm.

[22] http://www.unc.edu/awmc/pleiades.html.

[23] Y. E. Ioannidis, D. Maier, S. Abiteboul, P. Buneman, S. B. Davidson, E. A. Fox, A. Y. Halevy, C. A. Knoblock, F. Rabitti, H. J. Schek, and G. Weikum. Digital library information-technology infrastructures. *International Journal on Digital Libraries*, 5(4):266–274, 2005.

[24] A. Krowne. Building a digital library the commons-based peer production way. *D-Lib Magazine*, 9(10), 2003. http://www.dlib.org/dlib/october03/krowne/10krowne.html.

[25] A. Krowne and M. Halbert. An initial evaluation of automated organization for digital library browsing. In *JCDL*, pages 246–255, New York, NY, USA, 2005. ACM Press.

[26] C. Lagoze, D. B. Krafft, S. Payette, and S. Jesuroga. What is a digital library anymore, anyway? Beyond search and access in the NSDL. *D-Lib*, 11(11), 2005. http://www.dlib.org/dlib/november05/lagoze/11lagoze.html.

[27] Carl Lagoze, Sandy Payette, Edwin Shin, and Chris Wilper. Fedora: an architecture for complex objects and their relationships. *Int. J. Digit. Libr.*, 6(2):124–138, 2006.

[28] M. Lesk. The qualitative advantages of quantities of information: Bigger is better. *J. Zhejiang Univ. Science*, 11(6A), 2005.

[29] E. J. Neuhold, C. Niederée, and A. Stewart. Personalization in digital libraries: An extended view. In *ICADL*, 2003.

[30] D. W. Oard. Language technologies for scalable digital libraries. In *International Conference on Digital Libraries*, 2004. Invited Paper.

[31] G. Pant, K. Tsioutsiouliklis, J. Johnson, and C. L. Giles. Panorama: Extending digital libraries with topical crawlers. In *JCDL*, pages 142–150, New York, NY, USA, 2004. ACM Press.

[32] M. E. Renda and U. Straccia. A personalized collaborative digital library environment: a model and an application. *Information Processing and Management*, 41(1):5–21, 2005.

[33] M. Richardson and P. Domingos. Building large knowledge bases by mass collaboration. In *K-CAP*, pages 129–137, New York, NY, USA, 2003. ACM Press.

[34] M. Riva and V. Zafrin. Extending the text: digital editions and the hypertextual paradigm. In *HYPERTEXT*, pages 205–207, New York, NY, USA, 2005. ACM Press.

[35] John Russell. Making it personal: information that adapts to the reader. In *SIGDOC '03: Proceedings of the 21st annual international conference on Documentation*, pages 160–166, New York, NY, USA, 2003. ACM Press.

[36] A. F. Smeaton and J. Callan. Personalisation and recommender systems in digital libraries. *International Journal on Digital Libraries*, 5:299–308, 2005.

[37] E. S Villamil, C. González Muñoz, and R. C. Carrasco. XMLibrary search: an XML search engine oriented to digital libraries. *Lecture Notes in Computer Science - Research and Advanced Technology for Digital Libraries*, 3652:81–91, 2005.

[38] Gabriel Weaver, Barbara Strickland, Alison Jones, and Gregory Crane. Quantifying the accuracy of relational statements in Wikipedia: A methodology. In *To appear in JCDL 06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, 2006.

[39] R. Witte. An integration architecture for user-centric document creation, retrieval, and analysis. In *Proceedings of the VLDB Workshop on Information Integration on the Web*, 2004.

[40] I. H. Witten, K. J. Don, M. Dewsnip, and V. Tablan. Text mining in a digital library. *International Journal on Digital Libraries*, 4(1):56–59, 2004.

[41] Z. Zhuang, R. Wagle, and C. L. Giles. What's there and what's not?: Focused crawling for missing documents in digital libraries. In *JCDL*, pages 301–310, 2005.