Machine Assistance in Collection Building: New Tools, Research, Issues, and Reflections

Steve Mitchell

Digital tool making offers many challenges, involving much trial and error. Developing machine learning and assistance in automated and semi-automated Internet resource discovery, metadata generation, and rich-text identification provides opportunities for great discovery, innovation, and the potential for transformation of the library community. The areas of computer science involved, as applied to the library applications addressed, are among that discipline's leading edges. Making applied research practical and applicable, through placement within library/collection-management systems and services, involves equal parts computer scientist, research librarian, and legacy-systems archaeologist. Still, the early harvest is there for us now, with a large harvest pending. Data Fountains and iVia, the projects discussed, demonstrate this. Clearly, then, the present would be a good time for the library community to more proactively and significantly engage with this technology and research, to better plan for its impacts, to more proactively take up the challenges involved in its exploration, and to better and more comprehensively guide effort in this new territory. The alternative to doing this is that others will develop this territory for us, do it not as well, and sell it back to us at a premium. Awareness of this technology and its current capabilities, promises, limitations, and probable major impacts needs to be generalized throughout the library management, metadata, and systems communities. This article charts recent work, promising avenues for new research and development, and issues the library community needs to understand.

This article is intended to discuss Data Fountains (http://datafountains.ucr.edu) project work and thinking (and its foundation in the iVia system, http://ivia.ucr.edu) regarding tools and services, for use in collection creation and augmentation. Both systems emphasize automated and semi-automated Internet resource discovery, metadata generation, and rich-text harvest. These areas of work and research occur within the larger realms of machine assistance and machine learning. They are of critical value to libraries as they currently or potentially concern: significant resource savings; amplification and re-tasking of expert effort to better match librarian expertise with tasks that truly require it (through the automation of routine tasks); and better scaling of collections by providing them the technological wherewithal to grow, as appropriate, and better match the explosion of significant available knowledge and information that the Internet has accelerated.

This article is organized into three major sections:

- Part I details machine assistance work to date in the Data Fountains and iVia systems project.
- Part II describes current and upcoming promising research directions in machine assistance.
- Part III delves into planning and organizational issues that may arise for the library community as a result of these technologies.

Steve Mitchell (smitch@ucr.edu) is the Science Librarian for iVia/NSDL Data Fountains/Data Fountains Projects, Science Library, University of California, Riverside.

Part I: Recent work in Data Fountains and iVia

Part I covers work to date on Data Fountains and iVia. Section 1, "A new service and open source software," describes concrete project work with Data Fountains, a new open service and suite of open-source software tools for the educational and library communities, in developing practical machine learning to provide machine assistance in collection building. Data Fountains is an expansion of work based upon the iVia systems foundation.[1] It is an effort that has been ongoing and evolving since 1994.[2] Section 2, "Role and niche definition for machine assistance in collection building," covers recent developments in our ongoing effort to better research and define roles and niches for machine assistance of the types offered by Data Fountains. The spectrum—ranging from collection building with an emphasis on expertise that receives small assists from machine tools to an emphasis on machine tools that are configured and thereafter assisted through small refinements by expertise—is examined. Results from an initial exploratory survey in these areas are summarized.

A new service and open-source software—Data Fountains

Description

Data Fountains is an Internet resource discovery, metadata-generation, and selected, full-text harvesting service as well as the open source (Lesser General Public License

(LGPL) and General Public License (GPL) licensed) software that makes the services possible. It is a set of tools for use by organizations and institutions serving the greater learning community that create and maintain Internet portals, subject directories, digital libraries, virtual libraries, or library catalogs with portal-like capabilities (IPDVLCs) containing significant collections of Internet resources. It is an evolved variant of the iVia system, with which it shares many components. The Data Fountains / iVia code base represents more than 250,000 lines of primarily C++ code.

On the systems level, Data Fountains operates as an array of independent systems containing crawler, text classifier, text extraction, portal, and database software components customized to the needs of participating projects. Each cooperator and subject community works with, fine tunes, and benefits from its own set of crawler(s), classifier(s), and database manager(s), i.e., its own specific Data Fountain. Note that in this article, Data Fountains' portal / metadata repository / database management, content management, import-export, or content search / browse capabilities, which are substantial, will not be discussed.[3] Instead, the article will focus on its machine assistance and machine-learning components.

The Data Fountains system and service has been developed through a research partnership among computer scientists and academic librarians that is beginning to provide technological solutions to some of the major overall problems associated with the scalability and efficient running of IPDVLCs. Much project effort is based on applying machine-learning techniques to partially automate and provide help in a number of laborious and costly IPDVLC activities. Included here, more specifically, are the following needs / scaling challenges: reducing to some degree the high costs of manually created metadata; better coverage of the ever-increasing number of important Internet resources (relatedly, the relatively small size of most library Internet collections, where searches yielding very few or no results are common); reducing or making more efficient expert-involved tasks requiring little expertise; and reducing redundant efforts among IPDVLCs (both in content and systems building).

By providing inexpensive, universally needed raw materials (i.e., metadata and rich full text representing important resources), the Data Fountains service is intended to offer major support and resource savings to cooperating IPDVLC participants that otherwise have strong ongoing commitments to their established institutional identity or "brand," interface or look, system, and, more generally, "established way of doing things." Data Fountains viability and sustainability is keyed to providing universally needed service and very generic information products that do not require IPDVLCs to change—this often being seen as prohibitively expensive in time and resources. Data Fountains is intended to lower barriers for substantive cooperation in collection building and resource savings on the part of large numbers of IPDVLCs by developing, sharing, and distributing the benefits of machine learning in its areas of application.

The Data Fountains service will be useful to a large spectrum of academic and library-based finding tools including metadata repositories and catalogs with Internet portal-like capabilities.[4] Increasingly, library-catalog software is developing more flexibility, including, hopefully, the means by which full MARC (MAchine-Readable Cataloging) records coexist with more streamlined (and less expensive) records, e.g., Dublin Core (DC) and other types, and, moreover, metadata records that include or can be closely associated with selected rich full-text, among many other catalog need areas.[5] Data Fountains offers multiple levels of products and services geared to fit the needs of IPDVLCs of differing sizes, subject needs, and desired data "completeness" or depth (this being the amount and type of metadata and full-text needed to properly represent each resource).

Uses, products, and services

Overall, Data Fountains automatically or semi-automatically supplies varying levels of what represents the basic "ore" required by IPDVLCs for Internet resource and article collection building: access to significant, previously undiscovered resources as well as the metadata and selected full-text that describe or represent them. This ore is available in both raw (relatively unprocessed) and more refined products depending on the needs of the participating IPDVLC including, perhaps most importantly, the degree to which expertise is available to provide further refinement and how and for whom the material is intended to be used. Data Fountains multiple product and usage models supports the building of a wide array of IPDVLC collections.

A number of usage or service models are supported by Data Fountains, including:

Collection development support for single hybrid record type collections

The first usage model, based on full automation, involves the utilization of Data Fountains metadata and rich, full-text "as is," without review, to populate a collection. These records can be used by themselves or mixed with other types of records. They can also be used as part of a hybrid collection to undergird another, more primary, or fully expert-created, collection.[6] While more accurate, expert-created collections are not only comparatively more labor intensive and expensive to create and maintain, but often smaller, with narrower and more limited coverage. This has been the INFOMINE (http: / / infomine.ucr.edu) model that features two distinct collections, with the automatically generated collection supporting, as a second tier of

data, the expert-built content in the primary collection. Users can search one or both.

Internet resource discovery service

A second model uses Data Fountains primarily as an Internet resource discovery service where links and titles and other minimal metadata are supplied but where the user's intent is to identify new resources and build metadata records emphasizing a considerable amount of metadata not generated by Data Fountains (e.g., different subject schema). This is done by utilizing the Targeted Link Crawler, Expert Guided Crawler, or Focused Crawler. Because little to no metadata / rich-text generation / extraction occurs, this is the least complex of the usage models.

Crème de la Data Fountains

A third approach, a variation of the second, utilizes only those Data Fountains records that have been automatically determined, through a user-set threshold, to represent the most highly significant resources (e.g., the top 20 percent). These can be flagged for expert review or automatically harvested without review. The Data Fountains metadata retained for expert review, post-processing, and improvement can be minimal or full.

Metadata records intended for expert refinement

A fourth approach, which is semi-automated, involves using Data Fountains as both a discovery service and as a metadata record-building service where employment of records from the Data Fountains data stream is selective but the machine-created record is routinely retained as a foundation record to be refined or augmented by the expert.

Metadata records plus full-text

A fifth approach is to use the rich full-text selectively identified and harvested from the Internet resource, either in addition to the metadata generated or by itself, to populate a collection and greatly boost retrieval. That is, some collections may want to utilize metadata differing from that produced by Data Fountains but have Data Fountains perform the service of augmenting their metadata with rich full-text. All or parts of the object and full-text can be harvested.

Standards, metadata, and full-text

Data Fountains' record format is Dublin Core (DC) and features standard research library subject schemas including slightly modified Library of Congress Subject Headings (LCSH) and Library of Congress Classification (LCC). As part of upcoming work, development of additional classifiers to apply other subject / classification schemas / vocabularies will occur, notably DDC and those that can be automatically invoked from the terminology found in the collection objects. Cooperators may choose to help develop new formats, subject schemas, and metadata to meet custom needs in collecting and classification. Other important metadata generated include: Title, Creators, Description (an annotation-like construct), Keyphrases, Capitalized Terms, and Resource Language, among a total of thirty-plus fields. In addition to fielded metadata, Data Fountains delivers selected rich text harvested from the resource. This is important for enhancing IPDVLC retrieval capabilities and user-searching success. The rich text can be harvested verbatim and offered as-is for search or, if this is problematical, further processed into keyphrases.

Data post-processing, transfer, and product relevance assurance

Participants determine and download resources of relevance automatically in batch mode via subject-profiled, custom Internet crawls and editable results sets created by and for each IPDVLC to reflect its particular interests. These profiled crawls and metadata generation routines are stored and can be re-executed at selected intervals. Results are transferred using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) or SDF (Standard Delimited Format) in DC, MARC, and eXtensible hypertext markup language (XHTML) formats. In addition to batch transfers, participants can manually and interactively identify individual records or groupings of records that suit their needs for harvest. Selective, interactive, sorting / browsing of results, followed often by evaluation and editing of metadata and full-text fields (as individual records or globally in patterns), is enabled prior to export. These capabilities allow precisely targeted, custom record identification, modification, and downloading. This in turn enables the most general, as well as the most subject-specialized, IPDVLCs certainty in identifying and receiving only records that meet their need criteria.

Open-source software

The software making the above possible is available to all for free through the LGPL / GPL open-source licenses and model. The open-source model should work well for tool development as fundamental as that described. Open source of this type generally means that users freely use and perhaps participate in further development of the functionality of the software and, at intervals, contribute their innovations back to the code base for all to use. LGPL / GPL supports a wide diversity of forms of com-

mercial service development. Open source has worked well for large applications such as many forms of the Linux operating system (a number of variants of this are supported), Apache server software, and MySQL database management software (all of which are used by the Data Fountains system). Using this model has the intent of cooperatively benefiting the community as a whole. It is the author's belief that tools of the Data Fountains type will have wide enough usage within and are crucial enough to the library community to support the development of an open-source community around them. Data Fountains software is of use to thousands of institutions that build IPDVLC collections.

Open source also means that the development and evolution of a core tool or system for a community can potentially occur faster and more flexibly, with the proper community support, than many types of proprietary effort. This is needed given the continuing and increasingly greater revolutions in computing power and software potential. The community needs to be able to evolve faster in response to changing conditions, and free, community-based, open-source software development is one strategy for achieving this.

Current systems design, development, and features

To date, most of the work has emphasized research and development leading to innovations in preferential focused crawling, subject classification using Logistic Regression, kNearest Neighbor (kNN) and other classifiers, and rich full-text identification and extraction. A major emphasis in systems development has been identifying points of intervention in crawling, classification, and extraction, whereby initial, periodic or ongoing interactive expert input can be employed to improve machine processes and results. That is, the work has emphasized usage not only of fully automated machine processes but semi-automated machine processes intended to interactively augment, amplify, and improve the efforts of experts. Experts assist machine processes, and machine processes assist expert judgment/labor. The programming has also been done with an eye toward modularity among different systems components.

Internet resource discovery/ identification—expert guided and focused crawling

A number of crawling systems have been used; currently, for Data Fountains, three are used that represent two approaches to crawling: expert guided and focused. Expert-guided crawling is accomplished by a Targeted Link Crawler (TLC) and an Expert Guided Crawler (EGC). TLC is concerned with crawling a user-specified link or list of links. EGC differs from TLC in that the single "Start URL" link given is only the beginning point from which the crawler will either drill down (find onsite links at multiple depths in a site) or drill out (find external links not on the Start URL site). The result is that, compared with TLC, many more links than just those given the EGC crawler initially are crawled. With all crawlers, a metadata record with accompanying rich full-text is generated for each resource crawled.

A preferential focused crawler, called the Nalanda iVia Focused Crawler (NiFC) after the name of the ancient seat of learning in India, continues to be developed. Focused crawling makes possible focused identification of significant Internet resources by identifying specific, interlinked, and semantically similar communities of sites of shared subject interest. Generally, NiFC traverses subject expert-targeted regions of the Internet to find resources that are strongly interlinked and thereby represent coherent subject-interest communities and sites of shared interest and mutual use (i.e., are often concerned with and contain content similar to one another). Communities sharing interests often identify and cite one another through linkages on their Internet resources. Through this mechanism, these communities and their sites/resources can be identified, mapped, and harvested. Preferential focused crawling makes focused crawling more efficient by employing algorithms that can respond to clues in Web resource page layout and structure (e.g., using document object models, visual cues, and text windows adjacent to anchor text, among others) that indicate the more "promising" links to crawl. The result is more efficient focused crawling (figure 1).[7]

The focused crawling process starts with exemplary sites/pages/URLs being supplied by participating IPDVLC experts. These highly on-topic exemplars are used to form a seed set of model pages used for training/guiding the crawler. As the crawling progresses, an interlinkage graph is developed of which resources link to one another (i.e., cite and co-cite). Highly interlinked resources are evaluated, differentiated, and rated as to the degree to which they are linked to/from as well as for their capacities as authoritative resources (e.g., a primary resource such as an important technical report that receives many in-links to it from other resources) or hubs (e.g., secondary sources such as expert virtual library collections that provide out-links to other, authoritative resources). As hubs, expert-created, high-quality IPDVLC collections of links (e.g., INFOMINE) play an important role as milestones and navigation aids in the guidance of many types of crawling. Another automated process works to rate resources, as a second indirect measure of resource quality, by comparing for similarity of content (e.g., similarities among key-word

vocabularies) between the potential new resources and model resources. The most linked to / from authorities and hubs, with terminology most similar to the exemplars, are thus identified and become prime candidates for adding to the collection and for indicating other resources to add. The overall architecture of Data Fountains involves multiple concurrent crawls and an array of multiple crawlers and associated classifiers on multiple machines (i.e., there are one or more Data Fountains for each major subject area or major cooperator).

Areas of expert interaction in focused crawling

Expert interactive and semi-automated approaches to improve crawling are employed in and constitute special design areas of Data Fountains since many participating projects and communities have access to considerable subject expertise. There is much promise in amplifying the role of this expertise in the crawling process. Experts can create and refine crawls by:

- determining the most appropriate seeds (exemplary resources) to use (whether found in their own collections or generated from other sources);
- choosing degree of "on-topic-ness" desired (a precision versus recall setting);
- determining the total number of resources to be crawled;
- editing initial crawl results (e.g., de-selecting or blacklisting resources found) with an eye toward generally refining and developing a super seed set of very large numbers of increasingly on-target seeds that are then crawled anew. (This process of refinement and enlargement can be reiterated as desired in achieving increasing accuracy in and numbers of exemplars and therefore accuracy in the final crawl.)
- In addition, expert truing of crawler Web graph weightings (i.e., manually "lifting" the values of selected hubs and authorities) either during or after a crawling run is being explored to improve crawling accuracy. This lifting process can be aided through tools to visualize the crawl so that the expert can quickly identify, among the masses of results, the most promising areas of a Web graph for the crawler to emphasize.
- Expert-created blacklists of URLs for types of sites or pages that are not valuable can be stored to save future crawling and expert time. There is such a blacklist for each participating Data Fountains community group and individual.

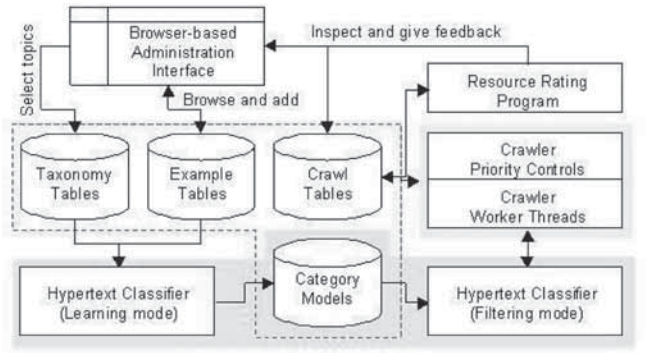

Figure 1. Focused and preferential crawling (courtesy of S. Chakrabarti)

Metadata generation—automated and semi-automated subject classification

Data Fountains and iVia embody innovations in automated metadata generation, including identifying and applying controlled subject terms (using academic library-standard subject schema), keyphrases, and annotation-like constructs (figure 2). Automated classifier programs apply these and other metadata and are part of a suite of programs known as the record builder. Controlled subject terminology applied currently includes LCSH, LCC, DDC, and Medical Subject Headings (MeSH). In assigning these, the system generally first looks for HTML and DC metatags and then extracts these data. With some fields, when these data are not present (which is common), original metadata are then generated automatically.

In the case of LCSH, LCC, and DDC, if not present in metatags, or if users choose to override metatag extraction (in cases where metatags are not accurate, such as when they are spammy or when top-page boilerplate metadata is carried onto all pages regardless of subject relevance), then classification processes are invoked. These derive a set of keywords and key phrases from the resource that serve as a surrogate in representing and summarizing its content. Then, using a model that encapsulates the relationships between these natural-language terms and the set of controlled-subject terms, the closest corresponding set of controlled terms is assigned. The model is learned from training data sets that consist of large sets of records (more than thirty million in corpora loaned for research purposes by the Cornell University Library, Library of Congress, California Digital Library [CDL], and OCLC)

from library catalogs and virtual libraries. With LCC, the aim has been to assign one or more LCCs to a resource based on the set of LCSHs associated with that resource. SVM, kNN, and Logistic Regression classifiers have been used. Generally, performance has been acceptable in cases where there were two hundred examples of the usage of a particular LCSH (in a record with a URL). Unfortunately, as large as the training data sets have been, there simply haven't been enough records for classification purposes with URLs and associated text. This problem will more than likely be resolved shortly as catalogs increasingly incorporate Web resources.

Metadata generation—Automated extraction of known, named entities

Named-entity (e.g., data elements that can be expected to be in a resource and that are placed by authors / publishers within a known textual / markup pattern) extraction is primarily practiced through the simple means of identifying and extracting data elements indicated by HTML / DC metatags, when present on a page. Data for more than thirty Dublin Core common (and not so common) fields are extracted. With some fields, extraction can be guided, as needed, in the interests of original metadata creation through pattern recognition and profiling, or through classification (e.g., title, subjects, description).

Rich-text identification and harvest

Refinement of our "aboutness" measure for identifying the most relevant pages or sections in a resource or document (i.e., those intended by the author to be rich in descriptive information about the topics within and the type of resource) from which to extract text is a continuing pursuit. Involved in this quest has been better determination of author-created structures and conventions in document or resource layout (e.g., locating introductions, summaries, etc., and determining / proportioning the amount of text to be extracted from each).

More accurate rich-text identification in turn yields more accurate identification, extraction, and application of key phrases and, from these, more accurate controlled subject term and other metadata application. This is at the foundation of many metadata generation processes. Crucially, rich full-text is also important from an end-user information-retrieval perspective because the natural-language terminology contained partially corrects for the limitations inherent in many controlled metadata and subject vocabulary / schema approaches (e.g., new or specialized subject terminology is often slow to appear or weakly represented in the often generalist library-standard subject schemas). Refinement of the "aboutness" measure in identifying terms indicating that rich text follows is an important and ongoing task that involves formulating fairly intricate text-extraction rules in reflecting conventions in rich-text placement in resources and documents of differing types (e.g., Web sites, articles, database interfaces), formats (e.g., HTML, PDF, postscript), and languages.

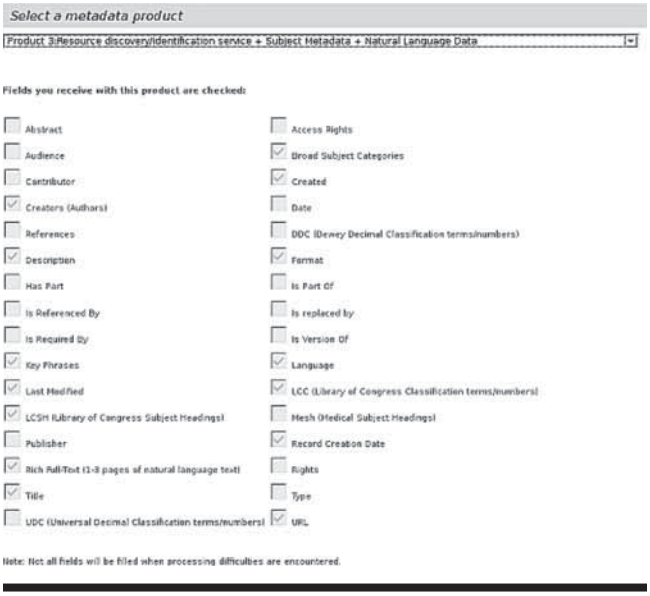

Figure 2. Metadata choices in Data Fountains

A modular architecture that supports a federated array of subject-specific focused crawlers and classifiers

The architecture that Data Fountains is based upon is shown in figures 3 and 4. Data Fountains operates on the systems level as an array of separate sets of bundled crawlers (both guided and focused), classifiers, and extractors; this bundled array of crawlers approach provides greater flexibility and efficiency, as compared with using a more monolithic, single-crawler, multiple-subject approach. A bundle can occupy a whole machine or several can exist independently, as virtual Data Fountains, on a single machine. Instead of one broad, multiple-subject, multiple-audience Data Fountain that follows a broad shotgun approach to Internet resource discovery and classification, there are several vertical, subject- and audience-focused Data Fountains. A Data Fountain is intended to exist for each distinct, major subject area and the subject-specific IPDVLC collections (e.g., visual arts, business, horticulture) associated with them.

Data Fountains systems architecture emphasizes modularity. It has been enabled and assumed that separate components of the system (e.g., the crawlers, classifiers, database management systems) could be developed further for other uses independent of the Data Fountains system. In addition, as technologies that the system is dependent upon advance, users will be able to more easily swap out and replace older modules. These capabilities contribute to system sustainability.

Service design and sustainability

Data Fountains was conceived to be a cooperative, nonprofit, low-overhead, cost-recovery-based service intended to sustain itself after start-up. Access will be provided to IPDVLC cooperators who demonstrate interest and support for the work and service. By so doing, cooperators share in supporting the continuing evolution and improvement of Data Fountains. As an additional sustainability consideration, the software has been released as open source so that it can develop and evolve in many directions (to directly fit unique needs) as well as benefit through distributed effort.

"Small is beautiful": Roles for and advantages of appropriate small- to medium-scaled tools

Approaches like those Data Fountains has taken may be among the few ways that Internet finding tools can continue to be relevant to the learning / library community and offer the accuracy and significant content needed by that community. The technical challenges faced by the large engines in their quest to cover an infinitude of audiences and Internet resources do not need to be grappled with by the community of research libraries and are not faced by focused crawlers and classifiers of the type Data Fountains relies upon. The latter are better able to develop targeted, more accurate approaches to their subjects because they enable machine assistance for, as well as amplification of, authoritative subject expertise (e.g., librarians) as a core interactive component in the process of finding and describing new resources. The processes involved target more narrowly defined, distinct, and finite subject universes and intellectual communities. This, in turn, allows them to scale appropriately for their tasks and to apply more complex and varied types of metadata for faculty, researchers, graduate students, and librarians, who generally require more precision (and authority) in their finding tools but still need to move beyond collections (even allied) that are essentially catalogs moved forward a notch. The

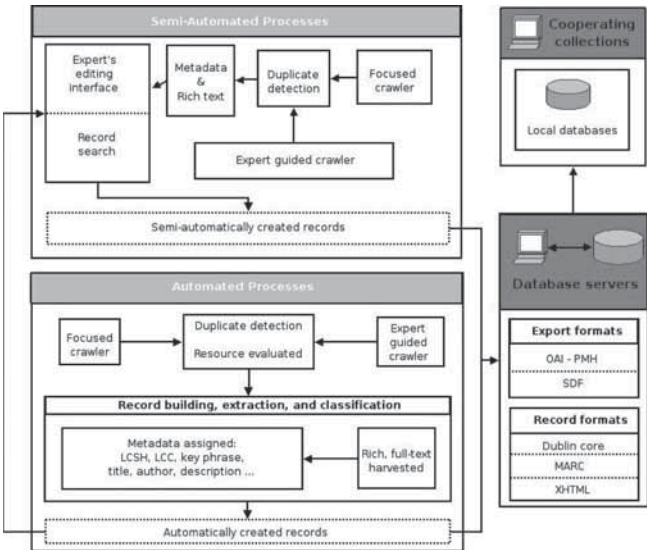

Figure 3. Interaction of fully and semi-automated and manual collection building processes in Data Fountains

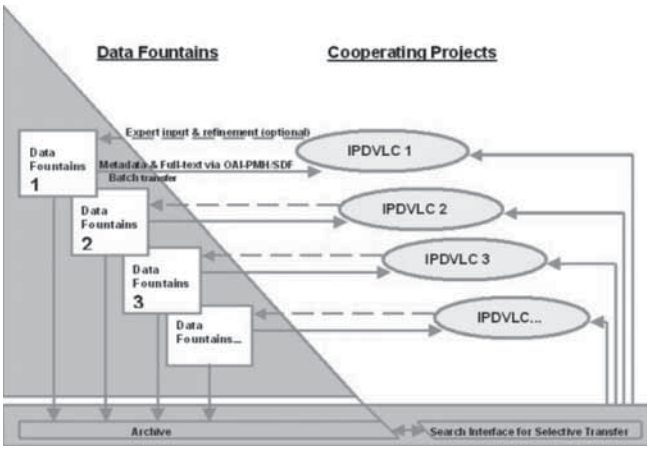

Figure 4. Overall Data Fountains architecture

smaller scale of this work also potentially enables innovations in effective linkage and similarity (i.e., semantic) analysis. Some experts note that the future of Internet searching as a whole may lie in searching federated finding tools based in these techniques.[8] Such a federation could be an academic's or librarian's Web of high-quality finding tools. Data Fountains may offer part of the foundation needed to support such a Web.

From a related perspective, these tools represent an appropriate approach for library and library-community-scaled resource identification and description tasks that emphasizes perhaps *the* great advantage the library com-

munity can bring to bear in creating useful finding and metadata generation tools, which no others have. That is, the community's unparalleled subject and description expertise in finding and ordering significant resources into coherent rich collections might be amplifiable shortly, through machine assistance. If such an effort was sensibly coordinated and focused, and minor modifications in approach and established standards made to enable best use of these new tools, then the best Internet finding tools / collections could be made possible yielding high-quality and significant coverage. These collections would benefit by having the capability to catalyze, out of the mass of the Web, the resources that constitute much of its intelligent fraction and make this coherently visible and available to learners and researchers. Moreover, this could be done in such a way that digital and print record and object collections could seamlessly interact as one, rendering what would be the best information-finding tools / collections without regard to type of resource. This effort in fact has been unfurling for a long time, though, to date, in small and somewhat sporadic, uncoordinated ways. For example, INFOMINE and similar collections have provided credible links to and for the academic community for well over a decade.

Role and niche definition for machine assistance in collection-building exploratory survey

An exploratory survey conducted in fall 2005 illuminated new perspectives, desired products and services, and research opportunities as perceived by a sampling of digital library and library leaders in regard to a number of areas involving machine assistance in collection building. Generally, areas explored concerned, among others: new roles projected for machine learning / machine assistance in libraries for metadata generation, resource discovery, and rich full-text identification and extraction; new finding-tool niches and opportunities existing in the service spectrum between Google and OPAC; acceptance of streamlined, more minimal, and cost-saving approaches to metadata creation or augmentation; the role of cost-recovery-based service and cooperative, participatory business models in digital libraries.

More specifically, the purposes of the survey were to:

1. Elicit leading library attitudes in relation to the types of services, software development, and research that generally will constitute Data Fountains;
2. Test the waters in regard to attitudes toward implementing machine-learning / machine-assistance-based services for semi-automated collection building within the general context of libraries;
3. Probe for new avenues or niches for these services and tools in distinction to both traditional library services / tools and large Web search engines;
4. Concretely define Data Foundations' initial set of automatically and semi-automatically generated metadata / resource-discovery products, formats, and services;
5. Examine attitudes toward the value and roles of rich, full-text in library-related finding tools;
6. Examine attitudes toward hybrid databases containing heterogeneous records (e.g., multiple formats, types, and amounts of metadata);
7. Gather ideas on cooperatively organizing such services; and
8. Generally define new ideas in all interest areas for development of products and services.

The survey, comprised of fifty-nine questions, was sent to thirty-five managers of leading digital libraries / libraries / information projects.[9] There was roughly a 40 percent return from those targeted (fourteen out of thirty-five). Responding institutions and individuals were guaranteed anonymity of response.

Survey result summary

There was considerable agreement on most answers. As such, this initial definitional survey has proven helpful in design and product definition. Though the survey sample set / number of respondents was limited and while results need to be seen as tentative, the views expressed are from well-regarded experts in the fields of digital library and library technology, development, and services. In addition to helping define current Data Fountains services, the survey results also indicated the need for further exploration in the areas of services, tools, overall niche definition, and publicity. While conclusions remain tentative, barring future, larger surveys, some of the more relevant results are as follows:

- There appear to be significant niches for an automated / semi-automated collection-building / augmentation service given inadequacies in serving research-library users found in Google (and presumably other large commercial search engines) and commercial-library OPAC / catalog systems. Survey results indicate a need for services of the types characterized by Data Fountains.
- Generally, academic libraries get a slightly above middle-value (neutral) grade in terms of meeting shifting researcher and student information needs over the last decade. This indicates that, above and beyond specific library and commercial-finding tools,

there are information needs not being met by libraries in regard to information discovery and retrieval that new services may be able to help provide.
- There is support, above and beyond creating machine-assistance-based collection-building services, for developing and distributing the free, open-source software tools supporting these services. Tools that make possible machine assistance in resource description and collection development are seen as potentially providing very useful services.
- Automated metadata creation and automated resource discovery/identification, specifically, are perceived as potentially important services of significant value to libraries/digital libraries.
- There is support for the notion of automated identification and extraction of rich, full-text data (e.g., abstracts, introductions) as an important service and augmentation to metadata in improving user retrieval.
- The notion of hybrid databases/collections (such as INFOMINE) containing heterogeneous metadata records (referring to differing amounts, types, and origins of metadata) representing heterogeneous information objects/resources, of different types and levels of core importance, was supported in most regards.
- Many notions that were foreign to library and even leading-edge digital library managers/leaders (the respondents) two to three years ago appear to be acknowledged research and service issues now. Included among these are: machine assistance in collection building; crawling, extraction, and classification tools; more streamlined types of metadata; open-source software for libraries; limitations of Google for academic or research uses; limitations of commercial-library OPAC/catalog systems; and the value of rich full-text as a complement to metadata for improved retrieval.
- There is strong support, given the resource savings and collection growth made possible, for the notion of machine-created metadata: both that which is created fully automatically and, with even more support, that which is automatically created and then expert reviewed and refined.
- Amounts, types, and formats of desired metadata (very streamlined DC metadata was supported for most uses and contexts) and means of data transfer (OAI-PMH was preferred) were specified by respondents.

Summary of Part I

Data Fountains is a unique service and system for inexpensively supporting aspects of collection building among IPDVLCs. Developing and utilizing advances in focused crawling and classification, this service automatically and semi-automatically identifies useful Internet resources (both open-Web as well as closed-collection resources including articles and reports, etc.) and generates metadata (and selected rich text) to accompany them. Data Fountains is a cooperative service, a free open-source software system, and a research-and-development project exploring machine assistance as well as machine-expert interfaces and synergies in collection building. Several useful service niches and roles for the work have been identified and have been or are being developed.

Part II: New directions in research

This section discusses important new directions in research for machine assistance in collection building as they relate to upcoming and expanding research, development, and prototyping within Data Fountains and iVia. Among focus areas are promising means of: automated classification for applying library standard controlled subject vocabularies/schema, including hybrid and ensemble classification; smarter and more accurate named-entity extraction (e.g., capturing object/article metadata "facts" such as publisher and publishing date); improvements in rich-text identification and harvesting; article/report collection level co-citation and subject gisting functionality; and generally improved expert-guided and focused Web crawling.

New research in machine assistance for collection building

The iVia and Data Fountains projects have recently received a fourth National Leadership Grant from the United States Institute of Museum and Library Services that supports three years of research and development in machine assistance in collection building. In addition, the National Science Digital Library is continuing funding. The areas that will be worked in are discussed below. These have been determined through experience gained over the last eight years of work in machine-assistance-oriented systems development and dialogue with computer scientists and collection coordinators. These areas of technology work and application, though complex and challenging, are very important. That is, assuming it is important that the learning/library community not be dis-intermediated by such technologies but instead becomes more fully empowered by them. This can only occur through developing a much larger role in actively defining, guiding, and putting the technologies to best possible use.

Looking into the future, it is clear that libraries cannot simply continue to wait for or rely on good companies like

Google, OCLC, or OPAC creators to deliver them, much like a cargo cult, as they have in the past. To the degree that this is done, there is the risk of becoming vendor vectors blinded by the limitations of these companies and their product lines. These products are often incorrectly assumed to be the known technical and organizational universe of what is possible or doable.

The revolutions coming in computing power together with the low cost of this power—which will be almost ubiquitously distributed among users of library collections and services—promise much more change than libraries have seen in the last decade. Among the changes underway are those in machine learning and machine assistance in libraries.

As the changes take place, organization size may not guarantee much as, over the last decade, librarians and researches have witnessed large academic and other research libraries, with some exception, demonstrate a profound organizational entropy in almost direct proportion to the magnitude of what are essentially paradigm shifts in scholarly communications, information provision, and research. To some degree, these simply reflect larger blockages within the universities and institutions in which libraries are embedded. As these changes play out, it should be noted that history in information or library-related public or scholarly information provision / access probably will not end with Google or OCLC—wonderful and fairly open companies—just as history in automobile manufacture has not ended with GM, computer manufacture with IBM, or Web finding with Alta Vista.

With this as background and in the vein of open planning (as well as open services and open software) and given the size of the work areas addressed and their challenges, much of the projects' technical planning and direction are being presented in this paper. These areas of computer and information-science research and development, which will affect libraries in many ways, are evolving rapidly into practical application.

The current major research areas are:

Named-entity identification and extraction, and unified models of information extraction and data mining

Named-entity identification and extraction is concerned with finding and harvesting generally concise factual data—often common bibliographic metadata—present in the targeted resource such as publisher, title, and publishing date. This type of metadata usually is associated with particular collections containing information objects that are often homogeneous (e.g., scientific article collections) and in which author-intended placement of metadata (or data) elements follows an established pattern and location in the object (e.g., an abstract is typically present and indicated in a pattern following presentation of title / author). While making extraction easy is one of the functions of metatagged metadata in Internet resources, generally few authors or collection coordinators in academia, or elsewhere, use metatags or applicable naming schema in any significant or uniform way (often, in fact, it is used very sparingly or not at all). Extractors therefore must be able not only to identify and harvest metatag metadata, but must discern and then extract specific metadata elements interspersed in bodies of text, as made identifiable by detecting the patterns of occurrence unique to the type of element as it occurs in the object or collection.

Among the many advances planned for Data Foundations is the usage of conditional random fields.[10] Important as well are user interfaces or dash boards that allow configuration of extractors whereby, as patterns of placement for desired data for extraction change in differing collections and types of objects, the tool can be configured appropriately to match the context and task. Also under development consideration are more hybrid, unified approaches to and models for data extraction and mining (as applies to text classification), using each to inform and improve the other.[11] That is, a family of models is being developed for improving data mining of information in largely unstructured text by using methods that "have such tight integration that the boundaries between them disappear, and they can be accurately described as a unified framework for extraction and mining."[12] Much of this work is concerned with generating metadata for article / report-level collections.

Document-scale learning and classification

A strong emphasis in the new work will be on document-scale machine learning, classification, and named-entity extraction in regard to collections of research papers, reports, theses, and monographs.

Internet-object boundary detection is another important concern. Detecting and properly defining compound documents (e.g., Web hyper-books on multiple pages or sites) is a goal, as is identifying compound-document points of author-intended entry and intended-user paths (i.e., author-intended main connective threads in distributed or compound documents).[13] Relatedly, improved internal-document structure identification for better document-level classification and extraction is critical. Involved are standard-document internal-structure identification (e.g., abstract, introduction, summary text, captions for tables / figures) including units of rich text and micro-information units of text organized via subtopic.[14] Methods of document-level word-and-phrase graphing as per

TextRank and other means of identifying small-world and micro-information units are currently being pursued.[15]

A strong emphasis as well will be on examining and implementing new means of co-referencing among documents in collections and new means of identifying latent topics in a well-defined collection. By way of explanation, another term for co-referencing is co-citation. An example of such co-referencing is referencing work, described in papers, that has been funded through the same agency and program or that shares principal investigators in addition to standard bibliographic citation. This will improve on work done in CiteSeer.IST (ResearchIndex) and similar projects through integrating and advancing the promising approaches of Rexa open-source collection-management software.[16] The focus of this effort will be on integrating article-level named-entity extraction as well as co-citation and bibliometric-refined subject identification within collections of papers / reports.

Individual text-classification algorithm and training method improvement

New research on individual text-classification algorithms will be examined and applied. The emphasis here will be on prototyping and measuring how applicable recent promising scholarly work might be to library-related metadata-generation challenges. The major focus continues to be in the area of applying controlled, library standard subject vocabularies (e.g., LCSH, LCC, and DDC). Many of the improvements relate to advances in individual text-classification algorithms, classifier training and fine-tuning, training-corpora cleanup and normalization techniques, and creating the ability for the individual classifiers to hybridize with other classifiers. Of special interest are classifiers that perform well with very large numbers of classes, both small and large amounts of text, and that yield probabilistic estimates in class assignment (e.g., of a particular LCSH). The latter allows both provision of multiple class assignments for resources that have multiple subjects as well as greater accuracy and knowledge of the confidence level of the assignments (thresholds of confidence level in accuracy can be set in applying, or not, a particular classification).

More specifically, this work will examine, test, and—depending on test results—refine recently improved variants of the most promising of several classification algorithms.[17] Among those are:

- Support Vector Machines (SVMs)[18]
- Logistic Regression (LR)[19]
- Naïve Bayes (NB)[20]
- Hidden Markov Models (HMMs)[21]
- kNN / kNN Model[22]

A number of metrics to measure performance of these and other text classifiers in regard to controlled subject assignment, in both fully-automated and user-interactive (semi-automated) modes, will also be employed.[23]

Hybrid classifiers

An important effort will be to test and develop new hybrid classifiers that incorporate the best capabilities of two or more in one classifier. Much of the current research has involved developing and improving new hybrids that combine the best of discriminative (e.g., LR, SVMs, decision trees) and generative (e.g., NB and Expectation Maximization) techniques in classification. For example, NB is fast but lacking in accuracy, while SVMs are accurate but can be slow to train. Hybrid models can produce better accuracy / coverage than either their purely generative or purely discriminative counterparts.[24] Various combinations, among others, of LR, HMM, and SVM are among the most promising.[25]

Ensemble classification or classifier fusion

This constitutes one of the main current directions in classification research and has been applied to a wide range of real-world challenges. Classification ensembles are reputed to be more accurate than any individual classifier making them up.[26] An important focus is on experimenting with new approaches to automated and semi-automated ensemble classification that involves creating frameworks that support metaclassifiers or classifier-recommender systems to apply multiple classifiers, as appropriate, to the classification task.[27] Developing classifier ensembles, including the metaclassifiers to guide them, is a major element in making possible the self-service aspect of an open, automated metadata-generation service, given that the metaclassifier is intended to determine the nature of the collection and classification task and assign the appropriate classifier(s) to the job.[28] It is probable that expert interaction at suitable points in this process will improve performance.

Distributed classification

Classifier ensembles are often used for distributed data mining in order to discover knowledge from inherently distributed and heterogeneous information sources and to scale-up learning to very large databases (often the context for library-related tasks). However, standard methods

of combining multiple classifiers, such as stacking, can have high performance costs. New classifier combination strategies and methods of distributed interaction will be examined to better handle very large classification needs.[29] Distributed classification, by nature, would be focused on improving large-scale self-service classification.

Semi-automated, expert-interactive classification

Means of enabling semi-automated, expert-interactive classification will be presented.[30] There is much scope for building interactive classifiers that engage the tool user or collection coordinator in an active dialogue (e.g., multiple iterations of machine / expert actions and feedback loops) that leads to incorporation of expert knowledge about specific classification tasks, metadata, and collections into the classifier, thus improving performance. That is, an active learning model can be extended significantly for these processes to include both feature-selection and document-labeling conversations, exploiting rapidly increasing computing power to give the user immediate feedback on choices to improve the classification process.[31]

Several different models featuring domain expert-interactive classification and extraction will be evaluated. These vary from being extremely interactive, emphasizing frequent machine assists, to less interactive, where experts profile, launch, and only occasionally refine a primarily machine process. The initial focus will be on the latter models. Note that iVia and Data Fountains have included a metadata generator with semi-automated record builder for years. OLLIE and HIClass are examples of systems that are more intensively expert-interactive.[32] Classification tasks and collection types will be characterized as to which lend themselves to frequent expert interactions, occasional interactions, or more fully-automated modes (i.e., little interaction or initial profiling / definition only).

Classifier training and evaluation techniques

As important as direct work on the classifiers is work emphasizing assessment, cleaning, and testing of classifier-training data and classifier-evaluation techniques. Involved are training data / corpora-normalization techniques, document-clustering techniques, and classifier bias / variance-reduction techniques. Also involved on the classifier side are tuning issues in regard to the data at hand, including improved feature-selection techniques and determining and using confidence estimates in applying / not applying classifications. Different approaches to these will be examined, tested, and refined with a range of training corpora.

Diverse training and test data from assorted collection "types" will include standardized corpora as well as data from participating library or educational community projects. That is, the techniques will be assessed with regard to how they perform with: (1) open Web resources, (2) collections of research papers, reports, theses, or monographs (working with Rexa), (3) typical campus Web-site pages, and (4) mixes of the above.[33] Each collection-type focus will require differing approaches, algorithms, training, and fine-tuning techniques and will be evaluated through a number of measures.[34]

Improved rich-text identification and extraction for improved classification and user search/ browse

Rich text is text that has the role of conveying through traditional or new document structures or conventions (e.g., introductions, tables of contents, FAQs, and captions for figures) the author-intended subject(s) and intent of the information object. Being able to accurately identify and extract this material greatly aids in classifier performance by improving significant keyphrase identification as well as in user retrieval by enabling full-text retrieval. The availability of natural-language text for searching is one means of helping to resolve problems encountered in searching controlled, library standard subject vocabularies (which in turn counteract problems searchers have when only natural-language retrieval is available). Both approaches are inherently complementary.

Improvements in rich-text identification and harvest through improved means of document-structure learning (e.g., identifying text windows around links or captions for figures and tables) will be sought. The lightweight semantic (e.g., use of terms that indicate "aboutness" such as "about", FAQs, introduction, and abstract; rating the frequency and uniformity of application of these terms in a given collection; and proportioning source of harvest) and markup clues will be refined as well. Identifying aboutness text, which can be seen as micro-information units of text organized via topic and subtopic, is being pursued through work with Rexa and others.[35]

Improved focused crawling

Focused crawling is an appropriately scaled method of crawling for many library collections (see Part I). It is used to discover new Internet resources by defined topic terminology and topic Web-link neighborhood. Topic similarity

and semantic analysis are key measures of significance that are combined with linkage co-citation measures to indicate significance or relevance of a new resource. Topic similarity among resources will be increasingly modeled through a topic-linkage matrix (i.e., semantic similarity map).[367] New means of evaluating, fine-tuning, and improving basic crawling will be examined.[37] Rules reflecting the specific semantics of each major subject area are to be developed by participants for crawls / classification.

Combined mining and extraction that support improved focused crawling in regard to best link pursuit and expert interaction

The development of hybrid, unified approaches to extraction and mining can be applied to focused crawling. The processes of data mining, rich-text identification and extraction, and the newest forms of focused crawling are starting to overlap and depend upon one another in important ways (as discussed in the section on preferential focused crawling). Another focus for development efforts will therefore be work to more systematically refine best-link pursuit with an eye toward combining advances in mining, extraction, and rich-text identification in focused crawling. This work will be undertaken to improve the work on NiFC. Focused crawling will improve in many situations, as well, through use of user-interactive components and data-visualization interfaces (e.g., control boards that visualize an interactive graph to aid in expert "lifting" of the values of specific sites / subtopic neighborhoods to better reflect their significance to the expert). This in turn that will help users guide and tune the crawling, in semi-automated fashion, to better fit the goals and context of a particular crawl.

Modeling different approaches for a self-service, openly accessible metadata-generation service(s)

The Data Fountains and iVia efforts have some experience with modeling metadata-collection related services, having provided collaborative, scholarly virtual-library service successfully for more than a decade. The Data Fountains project has improved upon earlier work and represents an automated and semi-automated resource discovery, metadata generation, and rich-text identification and harvest service for cooperating collections. The intent is that Data Fountains be a self-service operation. In related effort, with co-operators at the National Science Digital Library (NSDL) and Library of Congress (LC), the Data Fountains project has been striving to develop self-service dash boards that collection managers can use to configure, profile, and satisfy their needs. By complementing initial profiling with ongoing interactive dialogue, guidance, and refinement, more precise task definition and tool utilization can be achieved. The goal is to have a service that can, through advanced interfaces, engage users in dialogue to help them better determine their options, the tasks involved in achieving them, the capabilities and limitations of the tools available, and therefore, the best choice of tools and practices given their specific service needs and the nature of their collections.

Summary of Part II

There are many fronts of research in machine learning as applied to text processing and new-resource discovery in regard to collection building of various types, relevant to libraries, which have opened over the last few years. The Data Fountains / iVia research described is looking into just a few of these. For libraries, the borders between computer science, information science, and library science are dissolving rapidly. It would be hard to devise or project forward a five-year plan for a large working library without some understanding of current and oncoming machine-learning and machine-assistance work in each of these disciplines, the many inter-connected organizational / community / technical issues, and without an understanding that goes beyond the domain of current or developing products and services from existing vendors.

Part III: Issues and reflections

Part III is intended to define and address some of the many challenges and issues that are arising or may arise as a result of work on machine-assistance tools in the areas of automated and semi-automated resource discovery, metadata generation, and rich, full-text identification and harvest. Included here are reflections on and questions about some of the probable implications and impacts of, as well as roadblocks to, machine-learning technologies applied to collection building. Addressed are probable impacts leading to changing roles for libraries, librarian expertise, library standard vocabularies / schema, and the organizations that are the stewards of library standards. These include:

- What might be the effect of these technologies on library operations, including changes in the areas and nature of expenditure of expertise required, shifts in amount of expertise required, and changes in divisions of labor (both human / human and human / machine)?

- What are the effects on libraries and end users when the coverage of finding-tool content can be greatly and inexpensively broadened and deepened?
- How do current or traditional approaches to library-based practices and standards help foster or hinder these technologies?
- How will best practices develop in regard to machine-assisted activities?
- How do these technologies amplify and enable or simply prematurely dislodge librarian expertise?
- Who will own these technologies and tools?
- How open to evolution are library metadata standards and the organizations entrusted with their stewardship?
- How will these technologies impact these standards?

Unfortunately, most of these questions will remain as questions unanswered. The few answers offered here must remain as tentative, contradictory, and flawed as those of most who dabble in the cottage industry of imagining library futures. Still, in the effort to help map some of the new information landscape that is becoming apparent, these reflections, developed over the course of the last few years, may be small contributions toward defining and understanding what is coming.

Licensing for automatic agents of libraries

It will become increasingly important for libraries to develop licenses with commercial-resource vendors / publishers that allow crawlers / classifiers and other automated programs, to be seen as agents of and for these libraries. It is important that automated agents be allowed to work with (e.g., create or enrich metadata and therefore increase end-user success in finding) both free and fee-based materials in much the same way that an expert bibliographer, cataloger, or public-services librarian would when selecting, creating original metadata for, and providing access to a new commercially vended book intended to become part of a library or other well-defined collection. Automated agents accessing and processing fee-based, Internet-delivered information objects do so with the goal of improving the finding tools of the institution paying the fee to provide access for users to these objects (i.e., "library users"). Thus, they are engaged in a bona fide, fair use of the material by and for the purchasing / subscribing institution. The metadata and descriptive information these tools develop help make the materials they process more visible in collection / finding-tool contexts, a goal which should be desirable by all parties (i.e., end user, subscribing library, and owning author / publisher).

New medium, new organization, and an over-proliferation of electronic toll booths and borders

Another challenge is that Internet access to library-collection contents and library catalog-described data, both free and fee-based, is becoming increasingly restricted as libraries, library service organizations, and publishers grope to create special aggregations, with exclusive access for their clienteles. Countering this in their adherence to open access, have been, among others, services developed by, for example, arXiv, the Institute of Museum and Library Services open archive, CDL eScholarship, OAIster, CiteSeer, and NSDL.[38]

Differences in the two approaches may increasingly become an issue. On the one hand there is the broad, long-term community ethic favoring open access to an Internet with few walls or borders, and authors enabled to publish directly via the Internet through open eprint collections or dual commercial / personal-site publishing / copyrighting of their work. On the other hand there is the fairly narrow definition of an Internet information niche in which electronic / virtual services and collection access remain mapped restrictively to the sponsoring physical libraries / collections / institutions / publishers. Libraries face a contradiction or tension between these two approaches. The latter mode is a natural effort to retain a tightly held clientele and access model that has characterized physical libraries, reflecting narrowly conceived and decades-old organizational / budget / certification / user models of physical-library services and publisher controls. Much of this practice is necessitated by commercial publishers (for whom libraries often have no alternative but to act as vectors), together with the lack of vision for and outdated stereotypes held of libraries by the larger organizations in which they find themselves. At the same time, much of the problem is also due to the inability of libraries to develop new cooperative organizational modes, models, and services that map better to the new medium, map better to new author and user benefits enabled by this medium, and that are better able to exploit fully and fluidly the new medium's capabilities. The types of compartmentalization of collections, access, and services needed for physical libraries and print, or necessitated by publisher restrictions, are increasingly an obstacle when projected onto Internet access and service capabilities. Thorough rethinking is needed, just as the educational and scholarly missions of the university as a whole must be thoroughly rethought in the light of Internet-associated technologies and capabilities.[39]

While the information highway must be paid for, over-compartmentalization based on dated organizational and service models is yielding an over-multiplication of

toll booths and border crossings among aggregations and collections. An example has been the emphasis at many University of California campus libraries on the single campus OPAC rather than the pooling of resources across UC libraries for the strengthening and refinement of CDL's Melvyl Union Catalog. It is likely that with systemwide, multicampus shared resources, Melvyl could improve in all respects vastly beyond the single campus OPAC. This is noted in the Final Report of the Bibliographic Services Task Force of the University of California Libraries.[40]

Overall, institutional parochialism can and has greatly lessened the value and fluidity of the Internet as a medium for information provision. The booths and borders of tightly held collections make material harder to find, less visible, and less useful than would be true of more open, expansive collections and archives. As Dempsey stated, libraries need to find "better ways to match supply and demand in the open network. . . . We need new services that operate at the network level, above the level of individual libraries."[41] For crawlers and classifiers, the booths and borders that are proliferating in libraries can act disjunctively as barriers, reducing their performance.

There are few answers to the challenges that over-proliferation of booths and borders represent. They are often practical solutions to immediate needs. Still, projects that are exploring new avenues in organization and open, sharable collections (and the standards they are based upon) should be further encouraged and supported community-wide. These include the open archives already mentioned and systems such as those iVia / Data Fountains work upon that to provide services for such collections in an open, inclusive, cooperative, participatory manner. While the answer will probably remain a mix of open (reflecting capabilities of media) and closed (reflecting organizational and vendor restraints) collections, it would be progress to move the balance point more toward the middle and away from so many booths and borders.

Note on the related issue of meta-search

Libraries often respond to some of these open / closed / multiple-collection aggregator and "brand" challenges and issues with meta-search services. Meta-search can serve to mask the fundamental, growing problem of increasing booths and borders. Meta-search, unlike the Internet-borne conceptions of open service, collections, access, systems, software, and standards, does not really ask us to change our fundamental assumptions, organizations, or data architectures to match the capabilities of the new information medium. It does not ask us to cooperate more fully and share at the level of collection and data; it also doesn't encourage uniform-standards adoption and development. While meta-search is a fine answer to certain needs, sometimes it is used as a technical means to attempt to avoid these more fundamental issues.

In addition, meta-search can be constraining for user search / access—i.e., it frequently disallows use of significant or unique search and metadata capabilities of each individual database to which it is applied. Meta-search in libraries is becoming increasingly central, though it has many current operational flaws. Among these flaws are:

- simplification or dumbing-down of search in order to access lowest-common-denominator fields;
- clumsy cross-walking among fields, or metadata terminologies that really are not equivalents;
- difficulty in collating results / eliminating duplicates; and
- difficulty of matching differing results ranking weightings / systems held by different bases.

Libraries emphasizing this approach may be increasingly themselves perceived as dumbed down by academics, grad students, or serious researchers, who must reach beyond Google, the OPAC, and meta-search search and display. Instead of, or in addition to, meta-search, it might be wise to pursue more fully the hybrid database approach of combining heterogeneous records for multiple collections (and multiple retrieval languages as needed) in one database.[42] As computing power increases geometrically and price decreases drastically every couple of years, the challenge that the hybrid-database approach poses in regard to searching and maintenance of very large hybrid databases may soon become less of a problem. This power also implies that meta-search become more useful.

Library standard controlled subject schema/vocabularies

As the promise of automated and semi-automated metadata generation and related tools becomes better known, it may be important for the community as a whole to urge our major subject vocabulary standards organizations, i.e., LC and OCLC, to open more fully their standards and input in standard making for wider participation on the part of new communities of researchers, developers, and end users. Both organizations maintain important library standard subject vocabularies / schema, LCSH / LCC, and DDC, and related large bibliographic databases and classifier-training data embodying these standards. In this work, both organizations need to more actively seek out and encourage a wider variety of open innovation and development, both within and outside of the library community. This means involving more researchers, end users, and other perspectives in the effort of contributing to the more rapid evolution of these standards in an attempt both to better meet end-user finding needs and

to facilitate application of the standards through machine assistance. While OCLC and LC have been generous in providing their data and standards for iVia research (others that have been generous with training data have been the Cornell University Library and CDL), most known work on these standards is funneled through their organizations, allies, and organizational filters. This is, of course, critical to a point for coordination; however, if overdone it may unnecessarily inhibit wider pollinations, new perspectives (e.g., a wider variety of linguists, computer scientists, and subject vocabulary / schema experts from other disciplines such as medicine and the sciences), decision making, and faster movement forward.

Informing the perspective here is that, while there are major costs involved in maintaining and coordinating these vocabularies / schemas, such costs are being borne directly or indirectly by the community in fees paid, monies applied (often public monies through the large participating public university / land-grant libraries, among others), or labor volunteered / provided. LC is a public agency and OCLC a corporate cooperative. In many ways then, libraries, through their metadata expert / cataloger community, should be seen as "owning," as both co-author and funding agent, more of a share in these vocabularies (and other standards in library metadata) than their stewarding organizations. A significant portion of the success of thousands of individual libraries is dependent on the successful evolution (replacement?) of these standards through the facilitation efforts and new roles adopted by these two organizations.

Ultimately, it must be recognized that in many ways, OCLC and LC metadata schema and vocabularies (as well as conventions, styles, and customs in practical application) represent the codified wisdom, in the form of very large knowledge bases, of decades of resource description practice on the part of information professionals in thousands of institutions. The library community is the co-author of these, and OCLC and LC are their stewards. When viewing the community as owner, and when taking into account that the community needs to evolve more rapidly with its users to survive, then periodic clarification and renewal of the origin, intent, and understanding of the stewarding organizations and the standards they coordinate might help encourage more rapid, far-sighted change. Libraries may or may not sink to the degree that this is realized. In this light, it should be noted that some communities, including path-breaking projects within NSDL, have made well-reasoned decisions *not* to use these library subject vocabulary standards (Carl Lagoze, pers. comm.). These are just recent examples, given that abstracting and indexing services / databases, for the journal literature, have in most cases long ago chosen to use their own specialist vocabularies, often supplementing these by enabling key-word or natural-language searching of abstracts or complete full-text.

Among other core practical concerns here are that the library community's standards may not be seen as useful and as widely applicable as other information communities may desire. That is, if an important goal is to evolve and expand standards long associated with and emanating from the library community into becoming the standards of new, larger communities outside of libraries, then a more-guarded-than-not approach, which is slow to respond to early adaptors or innovators and slows sensible change, may not be the best path.

Here it should be said that there are significant ongoing efforts to overcome some of the challenges and better evolve LCSH / LCC. OCLC's Faceted Application of Subject Terminology (FAST) may represent a step in the right direction.[43] Having an entry-level vocabulary to translate end-user terminology to appropriate library subject standard vocabulary terms would be of great importance to most types of end user.[44] OCLC has also been working with the Resource Description Network (RDN) to streamline DDC application.[45] There just need to be more of these efforts moving at a more rapid clip. As MacEwan concluded in 1998, "if LCSH does not change it will sooner or later be abandoned. . . ."[46] The same might be said of library subject vocabulary / classification standards.

However, in the worst-case scenario, assuming the existing subject standards cannot evolve more rapidly to meet new user needs in information access, collection building, and metadata creation, now may even be an appropriate juncture for a large-scale rethinking and rebuilding, from the ground up.[47] The architecture, intent, end-user audience, form, and substance of these standards would need to be rebuilt and expanded. A capability for organizationally responding more quickly to what has amounted over the last few years to far-reaching paradigm shifts would be enabled. Now may be the time because, in addition to the questions of the openness / innovation / evolutionary adaptability of these standards, they exhibit significant, long-noted, functional flaws in terms of a non-librarian end user finding success. Among others often noted are:

- Misuse / lack of understanding on the part of end users (and, rarely, poor learning materials and guidance supplied by librarians) due to real or perceived complexity, often associated with the use of subheadings and arcane terms that are far from intuitive for users).[48]
- Typically sparse application that doesn't fully represent the number or depth of topics addressed by a work. Despite the time needed to create the MARC record manually, very few LCSHs are applied (often three or less in the University of California's Melvyl Union Catalog).
- The arcane and overly general nature of many terms that sometimes do not accord with terminology used by practitioners in the field.[49]

- The lack of currency of terms describing new or recent phenomenon (see discussion of entry vocabulary.[50]
- The lack of uniformity of subject granularity in their application across multiple cataloging institutions for the same / similar works.
- The significant amounts of expensive expert labor involved in their application.
- Their complexity often at least partially assumes some expert mediation (that may not be available, given that access is increasingly from outside the library) or long-term experience with the vocabulary.
- Overdone detail / complexity, some of it either not extremely useful to researchers and nonlibrarian end users or already instantly verifiable by users.
- Their arcane-ness and complexity, which limits capabilities for machine assistance in application and, thus thwarts a major, inexpensive means for future collection growth, increased coverage, and more useful collections.

Fortunately, and this is crucial, it turns out that much of the tonic needed for improvement may reside in the areas of inexpensively augmenting, as opposed to changing, the LCSH / LCC / DDC schema / vocabularies. For example, it is probable that most significant objects, when not digitized themselves, will be accompanied increasingly by digitized, representative cores of searchable natural-language rich text, as LC is doing with its table of contents digitization.[51] Automated and semi-automated tools for rich-text identification, extraction, and end-user searching are showing applicability now (see part I). Similarly, keyphrase identification and application can be accomplished automatically with a good degree of reliability; these processes play a role similar to rich text in providing useful retrieval terms and in augmenting subject searching with / without these controlled vocabularies. Finally, reasonably good overall subject gisting is occurring in the creation of annotation-like constructs. All of these—rich text, keyphrases, and annotation-like constructs alike—are of great potential value in addressing controlled subject vocabulary / schema inadequacies and in complementing LCSH / LCC / DDC in end-user finding.

It is also probable that use of machine means to augment overarching standard subject vocabularies with complementary and much more granular / detailed specialist vocabularies (both expert created and controlled as well as those that are automatically invoked) will shortly be practical and prove very useful. Streamlined LCSH / LCC / DDC could be made perhaps to function as linguistic "switching yards" with specialist vocabularies oriented to them and acting as extensions via the spine provided by the generalist vocabularies (similar to work being explored by Vizine-Goetz). All of this could be hinged on the synonomy and other term / concept relationships supplied by WordNet or other whole natural-language corpora.[52] In such a manner, reconceived LCSH / LCC / DDC can basically work as multi-vocabulary integration and translation tools in cases where the granularity of the subject becomes very fine-grained or specialized.[53] Such synonymy, linguistic linkages, and switching capabilities would make possible more meaningful and accurate interrelations and more fluid user movement among the vocabularies and concepts of multiple disciplines and multiple-controlled vocabularies / schema. This would also better enable the end user when employing terms actually used by practitioners / researchers / students in their disciplines.[54]

These and other efforts are crucial because, despite their problems, LCSH / LCC / DDC *are* comprehensive, overarching vocabularies and schema that, though complex (as are the subject vocabularies of BIOSIS and Pubmed / Medline, which successfully represent very large subject universes of their own), have done a generally useful job of representing and coherently organizing finding terminology for most known worldly (and unworldly) phenomena. This, on any basis, is no easy task.

These library standard vocabularies might best be seen as both essential connective tissue and as spines that could coherently thread many disciplines and interests, and many of the more specific vocabularies, together. Without such a spine, interdisciplinarians, researchers / students new to an area, and generalists—whose focus requires wide knowledge often across among many disciplines (and therefore subject vocabularies)—may find themselves handicapped. Each sub- and then sub-sub-specialization might develop its own mutually exclusive and contradictory terminology in a manner that natural-language substitutions such as keyphrase and rich-text availability can only partially fix. Many end users and librarians noted the downsides of natural-language-text-only searching two decades ago while using newspaper and other full-text databases offered by Dialog or BRS. Finally, one cannot ignore that LCSH / LCC / DDC have huge established bases of practitioners and metadata records employing them. Therefore, their value is large.

To summarzie, the solutions to the problems inherent in using library standard subject vocabulary / schema and other controlled metadata will involve the following:

- openness to extensive hybridization of approaches to rethinking subject vocabularies / schema and other metadata;
- awareness of, design for, guidance of, and incorporation of new machine-assisted technologies to boost collection coverage and reduce costs of application;
- embracing machine assistance, as appropriate, as a means of amplifying and extending expertise and application;
- applying existent technologies for generation of key-

phrases, description-like constructs, and rich text in order to augment controlled subject vocabularies;
- developing a better conception of end-user metadata expectations and needs against the backdrop and expectations generated by the Web, such as instant end-user access / verification; and
- making use of specialist vocabularies that might be dovetailed well with and coordinated through standard vocabularies.

Invoked subject vocabularies—hierarchical and otherwise

It is important to track recent research into automated and semi-automated means for creating (often referred to in the computer-science literature as "inducing" or extracting) hierarchical and other subject vocabularies / ontologies from natural-language corpora (see part II). The intent of this work is to have the natural-language terms used by practitioners directly populate and structure the subject-finding approach. Automated induction of subject vocabularies will be useful to augment and increase the capabilities, flexibility, and interactivity of standard subject vocabularies / schema.[55]

At the very least, and this is important, they could function to automatically suggest synonyms or new terminology for ongoing vocabularies / schema. And these approaches could be put to use in building entry-level vocabularies that front the vocabularies of the standards.[56]

They could also be used to aid in the semi-automated or automated repopulation / reworking of the standards, if large-scale, from-the-ground-up reworking is deemed necessary at some point. This would be done on a discipline-by-discipline, subject-by-subject basis.

Resource discovery, search engines, and your library's subject portal

Library collections, virtual libraries, portals, and Internet-enabled catalogs of openly accessible, significant Internet resources all function as "hubs" (see part I). Along with other types of expert-created hubs, they have played a role in providing most large, sophisticated, commercial search engines with a significant means for modeling and determining high-quality resources and, when accurate, a considerable portion of their accuracy. Though Google and others do not detail how their search algorithms work, most advanced crawlers highly weight (give authority to) sites that contain large numbers of links to research and other significant resources, especially when expert created. Similarly, resources from specific domains such as .edu, .org, and .gov, and institutions such as libraries, universities, and scholarly societies can be identified and more highly weighted. This is another case of the community's expertise / authority functioning as a knowledge base that, when offered as a public good (as library-created hubs often are), helps better enable directional tools for these commercial and noncommercial crawlers. There is nothing wrong with this as long as the community is aware of its contribution and as long as its efforts are recognized by these businesses. Expert library-based subject portals often reciprocate usage by using commercial engines for resource discovery, though this usually represents a minor way of collecting because other expert sources are preferred.

Enumeration of catalysts for, impacts of, and issues in machine assistance in the library community

Related to these research and technical developments, the library community needs to think through a great many interrelated and diverse issues and questions regarding (1) impacts of the machine assistance we have been discussing; (2) the possible massive automation of metadata generation and resource discovery in libraries, (3) who will "own" these technologies and ideas, and (4) changes in expectations / roles of metadata practitioners and standards and their stewards, in the following areas:

- When will machine learning / machine assistance yield reliable, inexpensive, and therefore massive application of metadata on an Internet scale, that meets librarian, and more importantly, end-user expectations in terms of usefulness? Machine assistance should begin to be factored into long-term planning.
- What will be the effects of this machine amplification in changing the importance / roles / content of subject standards? That is, how and to what degree will a new means and scale of application change these standards generally, and how they're perceived and used by end users and librarians and, therefore, be applied by the library community? How might these standards themselves change both in terms of changes in and approaches to vocabulary and schema? That is to say, how would massive, machine-assisted application in and of itself change the makeup of the vocabulary, schema, and the styles / conventions with which they are applied?
- How might the roles of the stewards of these standards change, given massive application as well as possible interest on the part of other communities? Can library standards penetrate and be effectively used by other information communities? What changes in the standards would be required to achieve this?

- What are the trade-offs between highly manual or craftsman / guild approaches and highly automated or more industrial approaches to applying metadata? Within which contexts, collections, resources, and budgets are these approaches to be best used, either singly or combined in various proportions, in building / expanding a collection? How does each approach best complement the other in library collections?
- To what degree will changing end-user information usage and access patterns change approaches in regard to collection design and access assumptions, the metadata standards the collections are based upon, and the stewarding organizations of the standards?
- To what degree may labor and resource savings, as well as the ability to provide for more comprehensive collections, as offered by this technology, dictate changes within the library community in regard to expectations for metadata quality and specificity? In which information-seeking contexts and collections and to what degree will the Google-type record or minimal, streamlined DC become, if not a necessity themselves, then a pole toward which library bibliographic metadata evolves?
- A question self-evident to most but not to all is: to what degree will the nature of the Internet itself continue to change our approach to supplying metadata? Again, researchers in academic departments no longer need walk across campus to the library by virtue of having many bibliographic details of an object present in a metadata record. Increasingly, they can go to the object on the Internet and instantly verify the detail for themselves. Should libraries deemphasize data elements / fields that are dependably and quickly end-user verifiable in favor of expending more expertise, time, and resources in gisting / describing the subject, intent, and perhaps even estimated quality or significance of the work?
- In which specific ways will labor be saved and machines be capable of assisting in resource discovery and metadata generation? That is, what level of automation / semi-automation is acceptable to the community and reliably deployable in production over horizons of one to five years? What level of quality / depth will users accept in metadata designed to occupy the continuum existing between the MARC record and the Google "record" (this being a large and significant service area; see part I)? How will this technology change old and enable new roles, tasks, and production routines for library subject experts and other staff? How will libraries ramp up and transition into this?
- Will the substantial potential economic advantages of automated or semi-automated generation of library standard metadata such as LCSH / LCC / DDC vocabularies / schema drive a rethinking toward greater uniformity / simplicity / streamlining of these standards and conventions in their application, explicitly with machine application in mind? For example, perhaps only a subset of a whole vocabulary will be used and those that are used will become less detailed and less rich for experts but also—for most end users—less complex and arcane, and more intuitive.[57]
- In some ways, the existence of DC is a recognition that this kind of rethinking and streamlining of library description standards, in the interest of representing and providing access to a much larger scale of communities and resources, is already well under way. What are the obstacles to greater usage of DC?
- What should the balance be in streamlining metadata for automated application, in relation to its current complexity / depth while augmenting with rich text? From another perspective, what is the balance when considering the oversimplification and loss of descriptive power when using machine methods as compared with that otherwise achievable through use of subject expertise? How will libraries determine best balances of expert and machine in regard to different tasks? How will this be quantified and determined through examination of user retrieval success / satisfaction—with this, in turn, factored against the backdrop of metadata creation costs, full-text data harvesting and retrieval, and the need for collections with much greater reach?
- As accurate means of metadata and rich-text generation for / from text objects improve, machine assistance will allow a shifting of expertise to provide better collection coverage and expression of subject-domain expertise (e.g., in abstracts). How will this new capability for breadth and depth be defined and used in library collections? For example, will new visual, multimedia, and data objects—which the Web has made possible on a mass basis and which libraries generally do not cover well—become a major goal in repurposing expertise since these do not easily lend themselves to machine processing (Karen Calhoun, pers. comm.)?
- Might streamlining and the usage of multiple depths / types of metadata application first require the acceptance within the community of the concept of the multitiered collection / database that supports multiple levels and types of heterogeneous resources representing differing levels of importance to users?[58] Or, can this need be met through more fully evolved meta-search approaches?
- Helping to structure this metadata heterogeneity might be the sliding-scale application of varying levels of metadata-generation labor expenditures and amounts / type of metadata, with the lower-

and middle-value resources receiving application of streamlined standard vocabularies/schema and rich text, automatically or semi-automatically, at low cost. High-value resources would continue to receive expert-applied, expensively created, complex, and high-quality metadata as well as rich text. Libraries already make such distinctions in quality/significance to some degree through purchasing (e.g., departmental collecting profiles/weightings by subject and object type and cost) and order-of-cataloging priority decisions, as well as by student/faculty input on specific items. More specifically, we would need to discuss and develop criteria in determining the core or peripheral value of a resource for its subjects and user communities and then, based on the judgments derived, appropriately apportion amount and type of metadata and expert labor or machine assistance, on a sliding scale. Again, while it should be noted that the library community has generally avoided rendering judgments on the possible use/relevance of a resource to a subject community, libraries nevertheless do routinely make general calls that effectively function this way to some degree. In making this judgement, it would be critical to involve resource users. Reviewer-researcher, library user, and librarian evaluations for purchases as well as finding tool/collection-usage statistics for the specific subject or author and item all could be woven into the means by which the core weighting of a resource could be assigned and be refined over time via usage. Developing this value is important from a library standpoint. It is a key that may help unlock solutions for some of the community's bigger challenges, including those revolving around the best marriage of machine assistance with librarian expertise. How do libraries go about making these sliding-scale evaluations with some uniformity, among different collection types and interests, with an eye toward tasking expert and machine?

- Can some of the general end-user search deficiencies commonly acknowledged for LCSH/LCC/DDC be rectified to some extent by automatically/semi-automatically providing rich full-text accompaniment for each record/resource, either in the form of "selected" excerpts verbatim or as processed into significant keyphrases representing this text? How could the presence of this rich text not so much change as augment these standards? For example, rich full-text might be relied upon to contain detail that obviated the need to use certain LCSH subdivisions or other types of MARC metadata. Could inadequacies/inaccuracies in expert-applied and machine-applied metadata be partially countered, for end-user retrieval purposes, through the presence of rich full-text? Rich text, as well as keyphrases/terms and descriptions that serve the same purpose in this context, can now be reliably generated in many cases automatically. What would be the right mix of subject-vocabulary standard metadata and accompanying, selected natural-language text for best end-user success? How might rich-text extraction and searching improve upon searching of whole-object full-text? How much rich text is needed and how distilled should it be? Large, whole-object full-text searching can often be a searcher's quagmire, clouding results rankings and weightings.
- Could a new scale of application and interest on the part of new communities be better catalyzed through the incentive offered by opening up the LCSH/LCC/DDC subject vocabularies/schema on an open-standards/open-source, free-software model?
- If development of these technologies is constrained with regard to action/inaction on the part of the community and its stewards, will the standards be replaced—or become obsolete—for major existing or prospective sectors of users? If so, what does this mean for the library community?
- By and for whom is such standard subject vocabulary/schema application technology developed within the community? Classifiers are actually trained through great amounts of what, in many cases, is really community-created knowledge in order to apply community-developed schema/vocabularies. Smart crawlers and extractors similarly use (have "learned") collectively created information patterns, derived from open-knowledge bases of various sorts. Who should own these tools/models and how open/closed should the programming code/ideas be, considering they could not be built without using the collective wisdom embodied in these knowledge bases? These tools exploit decades of labor by thousands of institutions, whose assumption has generally been that the knowledge base and, by extension, the tools that are built on and benefit from it, are and should remain directly or indirectly, public goods.
- For whom is machine learning/assistance in collection building patented? The ideas, training corpora, algorithms, and data models discussed need to be observed and protected for the public domain to encourage their widespread and inexpensive availability, as well as their evolution. The U.S. Patent and Trademark Office is now more commonly supporting the patenting of whole, generic processes that have heretofore had one or both feet in the commons, as compared with solely granting patent rights in more discrete areas of original invention. It would be unfortunate to find one day that machine assistance in collection building had been patented. This is especially an issue, given that there is little machine learning of interest to libraries that does not mine, apply, and extend the stored wisdom and knowledge that the community has built for decades.

Summary of Part III

It is important to think through and anticipate a great number of issues and concerns—including those of open models and open development—regarding machine-assistance tools (e.g., classifiers, extractors, and related algorithms / models) that generate library standard metadata, and identify and extract useful natural-language data. It is important because these tools could become central activities in libraries over the next one to five years. Reflection here is especially appropriate, given the degree that these tools are trained on exemplars from library collections and come to distill and embody models of library metadata, standards, and expertise that represent the knowledge created over decades through the effort of a whole community. It is important to think through what machine-assistance technologies in collection building imply for the future role of the librarian's expertise. Specifically, libraries need to reconceptualize machine-assistance software not as fully automated "AI" but rather, as enabling expert driven, strongly interactive, "servo-mechanisms" that semi-automate some work to increase the reach, quality, and user-finding success within library collections. While it will probably start out with ten or fifteen minutes of expert time saved per record by such tools, this is a lot of time saved when aggregated across the entire community and will only increase. And the community needs to think through what this implies for the evolution of library-standard metadata, given that machine assistance will increasingly allow for massive and economic application, if a convergence of machine capabilities and machine-friendly metadata standards is architected.

This large-scale amplification of usage will quite likely involve changing the value / roles of these standards for the community, as well as for the larger communities that may come to use them at the cost of simplification, streamlining, and a greater reliance on end users to verify some of their own metadata details (often interacting directly with the digital resource). The tools also imply a restructuring of expertise and its application in metadata creation in libraries to reflect a division of labor, with semi-automated machine description processes spent on the mass of useful but mid- to lower-value materials; with and expert time being spent on high-value resources; and with both types of records residing in the same multitiered, heterogeneous collection.[58] Finally, needing examination will be the roles of the stewardship organizations in:

- shepherding the community's metadata standards during a period of great change;
- openly evolving the application of metadata standards within the context of machine assignment for the greatest possible good;
- rapidly evolving the application of metadata standards to retain guidance of and to keep pace with open and proprietary developments in these areas;
- distilling the metadata knowledge base and wisdom created by the community as this is transformed into the programmatic knowledge (rule bases and models) used by new tools.

This knowledge base is a priceless asset for the library community in sustaining service roles in an age of the large-scale advent of commercial-information access, delivery, and ownership.

Conclusion

This article discusses work over the last several years in machine-learning software and services relevant to collection building in libraries. A number of promising avenues for exploration and research are detailed. Deeper understanding of and more direct involvement in areas of machine learning are urged for libraries in order to reflect advances in the computer sciences and other disciplines as well as to meet changing end-user needs among information seekers.

Acknowledgements

The author would like to thank the U.S. Institute of Museum and Library Services; the Library of the University of California at Riverside; the National Science Foundation's National Science Digital Library; the Fund for the Improvement of Post-Secondary Education of the U.S Department of Education; the Librarians Association of the University of California; and the Computing and Communications Group of the University of California at Riverside for current or past funding support. The author would also like to thank the Library of Congress; Cornell University Library; OCLC; and the California Digital Library for providing training data and other assistance for the research. Thanks to Karen Calhoun (Cornell University Library) and two anonymous readers for some excellent comments and suggestions. Finally, the author would like to commend iVia lead programmer Johannes Ruscheinski, primary author of the Data Fountains and iVia code bases, for his excellent work over the years, as well as Gordon Paynter, Walt Howard, Jason Scheirer, Keith Humphries, Anthony Moralez, Paul Vander Griend, Artur Kedzierski, Margaret Mooney, John Saylor, Laura Bartolo, Carlos Rodriguez, Jan Herd, Carolyn Larson, Diane Hillmann, and Ruth Jackson for their invaluable contributions to the

projects. The views expressed here are solely those of the author and not intended to represent those of the Library of the University of California, Riverside, our funding agencies, or cooperators. ■

References and notes

1. S. Mitchell et al., "iVia: Open Source Virtual Library Software," *D-Lib Magazine* (January 2003). http://www.dlib.org/dlib/january03/mitchell/01mitchell.html (accessed Oct. 20, 2006); G. Paynter, "Developing Practical Automatic Metadata Assignment and Evaluation Tools for Internet Resources," in *Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries* (Denver: ACM Pr., 2005), 291–300 (Winner of the JCDL 2005 Vannevar Bush Best Paper Award), http://ivia.ucr.edu/projects/publications/Paynter-2005-JCDL-Metadata-Assignment.pdf, (accessed Oct. 20, 2006); S. Mitchell, "Collaboration Enabling Internet Resource Collection-Building Software and Technologies," *Library Trends* 53, no. 4 (May 2005): 604–19; J. Mason et al., "INFOMINE: Promising Directions in Virtual Library Development," *First Monday* (2000), http://www.firstmonday.dk/issues/issue5_6/mason/ (accessed Oct. 20, 2006).

2. S. Mitchell, "INFOMINE: The First Three Years of a Virtual Library for the Biological, Agricultural, and Medical Sciences," in *Proceedings of the Contributed Papers Session, Biological Sciences Division, Special Libraries Association Annual Conference* (Seattle: Special Libraries Assocation, 1997).

3. Mitchell, "Collaboration Enabling Internet Resource Collection-Building Software and Technologies."

4. J. Phipps et al., "Orchestrating Metadata Enhancement Services: Introducing Lenny," in *Proceedings of DC-2005: International Conference on Dublin Core and Metadata Applications* (Madrid, Spain: Universidad Carlos III de Madrid, 2005), http://arxiv.org/pdf/cs.DL/0501083, (accessed Oct. 20, 2006).

5. Mason et al., "INFOMINE: Promising Directions in Virtual Library Development."

6. Ibid.

7. S. Chakrabarti, *Mining the Web: Discovering Knowledge from Hypertext* (San Francisco: Morgan Kauffman, 2003); S. Chakrabarti et al., *Accelerated Focused Crawling through Online Relevance Feedback*, http://www2002.org/CDROM/refereed/336/ (accessed Oct. 20, 2006); S. Chakrabarti, *The Structure of Broad Topics on the Web*, http://www2002.org/CDROM/refereed/338/index.html (accessed Oct. 20, 2006); S. Chakrabarti, *Integrating the Document Object Model with Hyperlinks for Enhanced Topic Distillation and Information Extraction*, http://www10.org/cdrom/papers/489 (accessed Oct. 20, 2006).

8. Chakrabarti et al., *Accelerated Focused Crawling*; F. Menczer, "Mapping the Semantics of Web Text and Links" *IEE Internet Computing*, 9, no. 3 (May/June 2005): 27–36; F. Menczer, G. Pant, and P. Srinivasan, "Topical Web Crawlers: Evaluating Adaptive Algorithms" *Transactions on Internet Technology*, 4, no 4 (2004): 378–; F. Menczer, "Correlated Topologies in Citation Networks and the Web" *European Physical Journal B*, 38 no. 2 (March 2004): 211–21.

9. S. Mitchell, "Data Fountains Survey," 2005, http://datafountains.ucr.edu/datafountainssurvey.doc, (accessed Oct. 20, 2006).

10. A. Culotta and A. McCallum, "Confidence Estimation for Information Extraction," in *Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics* (Boston: Association for Computational Linguistics, 2004), http://www.cs.umass.edu/~mccallum/papers/crfcp-hlt04.pdf, (accessed Oct. 20, 2006); F. Peng and A. McCallum, "Accurate Information Extraction from Research Papers Using Conditional Random Fields," in *Proceedings of the Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics* (2004). http://ciir.cs.umass.edu/pubfiles/ir-329.pdf, (accessed Oct. 20, 2006); C. Sutton and A. McCallum, "An Introduction to Conditional Random Fields for Relational Learning," in *Introduction to Statistical Relational Learning*, Lise Getoor and Ben Taskar, eds. (Cambridge, Mass.: MIT Pr., 2006). http://www.cs.umass.edu/~mccallum/papers/crf-tutorial.pdf, (accessed Oct. 20, 2006).

11. A. McCallum and D. Jensen, "A Note on the Unification of Information Extraction and Data Mining Using Conditional-Probability, Relational Models," in *Proceedings of the IJCAI 2003 Workshop on Learning Statistical Models from Relational Data, Acapulco, Mexico: IJCAI*, http://www.cs.umass.edu/~mccallum/papers/iedatamining-ijcaiws03.pdf, (accessed Oct. 20, 2006); U. Nahm and R. Mooney, "A Mutually Beneficial Integration of Data Mining and Information Extraction," in *Proceedings of the American Association for Artificial Intelligence/Innovative Applications of Artificial Intelligence* (Austin, Texas: American Association for Artificial Intelligence, 2000). http://www.cs.utexas.edu/users/ml/papers/discotex-aaai-00.pdf, (accessed Oct. 20, 2006); R. Raina et al., "Classification with Hybrid Generative/Discriminative Models," in *Proceedings of Neural Information Processing Systems* (2003). http://www.cs.umass.edu/~mccallum/papers/hybrid-nips03.pdf, (accessed Oct. 20, 2006); G. Bouchard and B. Triggs, "The Trade-Off Between Generative and Discriminative Classifiers," *COMPSTAT 2004*. (Prague: Springer, 2004) http://lear.inrialpes.fr/pubs/2004/BT04/Bouchard-compstat04.pdf, (accessed Oct. 20, 2006).

12. McCallum and Jensen, "A Note on the Unification of Information Extraction."

13. N. Eiron and K. McCurley, "Untangling Compound Documents on the Web," in *Conference on Hypertext* (Nottingham, UK: ACM Conference on Hypertext and Hypermedia, 2003), http://citeseer.ist.psu.edu/eiron03untangling.html, (accessed Oct. 20, 2006). http://www.almaden.ibm.com/cs/people/mccurley/pdfs/pdf.pdf, (accessed Oct. 20, 2006); P. Dimitriev et al., "As We May Perceive: Inferring Logical Documents from Hypertext," presented at *HT 2005, 16th ACM Conference on Hypertext and Hypermedia* (Salzburg: ACM, 2005); K. Tajima, "Finding Context Paths for Web Pages," in *Proceedings of ACM Hypertext* (Darmstad, Germany: ACM, 1999), http://www.jaist.ac.jp/~tajima/

papers/ht99www.pdf, (accessed Oct. 20, 2006); K. Tajima et al., "Discovery and Retrieval of Logical Information Units in Web," in *Proceedings of the Workshop of Organizing Web Space* (in conjunction with *ACM Conference on Digital Libraries*) (Berkeley, Calif.: ACM, 1999), 13–23, http://www.jaist.ac.jp/~tajima/ papers/wows99www.pdf, (accessed Oct. 20, 2006); E. de Lara et al., "A Characterization of Compound Documents on the Web," TR99-351, University of Toronto (1999), http://www.cs.toronto.edu/~delara/papers/compdoc.pdf, (accessed Oct. 20, 2006), http://www.cs.toronto.edu/~delara/ papers/compdoc_html/, (accessed Oct. 20, 2006); L. Xiaoli et al., "Web Search Based on Micro Information Units," (Honolulu, Hawaii: Eleventh International World Wide Web Conference, 2002), http://www2002.org/CDROM/poster/78.pdf, (accessed Oct. 20, 2006); W. Lee et al., *Retrieval and Organizing Web Pages by Information Unit*, http://www10.org/cdrom/papers/466/, (accessed Oct. 20, 2006).

14. Tajima et al., "Discovery and Retrieval of Logical Information Units in Web"; Xiaoli et al., "Web Search Based on Micro Information Units"; Lee et al., *Retrieval and Organizing Web Pages.*

15. R. Mihalcea, "Graph-Based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, companion volume (Barcelona, Spain: Association for Computational Linguistics, 2004), http://www.cs.unt.edu/~rada/papers/mihalcea.acl2004.pdf, (accessed Oct. 20, 2006); R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Texts," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Barcelona, Spain: Empirical Methods in Natural Language Processing, 2004), http://www.cs.unt.edu/~rada/papers/mihalcea.emnlp04.pdf, (accessed Oct. 20, 2006); R. Mihalcea, P. Tarau, and E. Figa, "PageRank on Semantic Networks, with Application to Word Sense Disambiguation," in *Proceedings of the 20th International Conference on Computational Linguistics* (Geneva, Switzerland: COLING 2004). http://www.cs.unt.edu/~rada/papers/ mihalcea.coling04.pdf, (accessed Oct. 20, 2006); Y. Matsuo et al., "KeyWorld: Extracting Keywords in a Document as a Small World," in *Proceedings of Discovery Science* (Berlin, New York: Springer, 2001), 271–81 (Lecture Notes in Computer Science, v. 2226), http://www.miv.t.u-tokyo.ac.jp/papers/matsuoDS01.pdf, (accessed Oct. 20, 2006); Y. Matsuo and M. Ishizuka, "Keyword Extraction from a Single Document Using Word Co-Occurrence Statistical Information," *International Journal on Artificial Intelligence Tools* 13, no.1 (2004): 157–69, http://www.miv.t.u-tokyo.ac.jp/papers/matsuoIJAIT04.pdf, (accessed Oct. 20, 2006); Xiaoli et al., "Web Search Based on Micro Information Units"; Lee et al., *Retrieval and Organizing Web Pages*; G. Forman and Ira Cohen, "Learning from Little: Comparison of Classifiers Given Little Training," Tech Report: HPL-2004-19R1 20040719 (Palo Alto, Calif.: Hewlett-Packard Research Labs., 2004), http://www.hpl.hp.com/techreports/2004/HPL-2004-19R1.pdf, (accessed Oct. 20, 2006).

16. G. Mann et al., "Bibliometric Impact Measures Leveraging Topic Analysis," (in press), in *Proceedings of the Joint Conference on Digital Libraries* (2006). http://www.cs.umass.edu/~mccallum/papers/impact-jcdl06s.pdf, (accessed Oct. 20, 2006).

17. R. Bouckaert and E. Frank, "Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms," in *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining.* (Berlin, New York: Springer-Verlag, 2004), 3–12 (Lecture Notes in Computer Science, v. 3056), http://www.cs.waikato.ac.nz/~ml/publications/2004/bouckaert-frank.pdf, (accessed Oct. 20, 2006); R. Bouckaert, "Estimating Replicability of Classifier Learning Experiments," in *Proceedings of the International Conference on Machine Learning* (2004), http://www.cs.waikato.ac.nz/~ml/publications/2004/bouckaert-estimating.pdf, (accessed Oct. 20, 2006); R. Caruana and A. Niculescu-Mizil, "Data Mining in Metric Space: An Empirical Analysis of Supervised Learning Performance Criteria," in *KDD-2004: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York: ACM Press, 2004), http://perfs.rocai04.revised.rev1.ps, (accessed Oct. 20, 2006).

18. J. Zhang et al., "Modified Logistic Regression: An Approximation to SVM and Its Application in Large-Scale Text Categorization," in *Proceedings: Twentieth International Conference on Machine Learning* (Menlo Park Calif.: AAAI Press, 2003), 888–97, http://www.informedia.cs.cmu.edu/documents/icml03zhang.pdf, (accessed Oct. 20, 2006); Y-C. Chang, "Boosting SVM Classifiers with Logistic Regression," Technical Report. (Taipei: Institute of Statistical Science, Academia Sinica, 2003), http://www.stat.sinica.edu.tw/library/c_tec_rep/2003-03.pdf, (accessed Oct. 20, 2006); T. Zhang and F. Oles, "Text Categorization Based on Regularized Linear Classification Methods," *Information Retrieval* 4, no. 1 (2001): 5–31, http://www.research.ibm.com/people/t/tzhang/pubs.html, (accessed Oct. 20, 2006); T. Joachims, "SVMlight," (including SVMmulticlass, SVMstruct, SVMHMM) (software, 2005), http://svmlight.joachims.org/, (accessed Oct. 20, 2006); C. Chang and C-J. Lin, "LIBSVM," (software, 2005), http://www.csie.ntu.edu.tw/~cjlin/libsvm/, (accessed Oct. 20, 2006); C-W Hsu and C-J Lin, "BSVM," (software, 2003), http://www.csie.ntu.edu.tw/~cjlin/bsvm/index.html, (accessed Oct. 20, 2006); T. Finley and T. Joachims, "Supervised Clustering with Support Vector Machines," in *Proceedings of the International Conference on Machine Learning* (New York: ACM Press, 2005), http://www.cs.cornell.edu/People/tj/publications/finley_joachims_05a.pdf, (accessed Oct. 20, 2006); I. Tsochantaridis et al., "Support Vector Machine Learning for Interdependent and Structured Output Spaces," in *Proceedings of the International Conference on Machine Learning* (New York: ACM Press, 2004), http://www.cs.cornell.edu/People/tj/publications/tsochantaridis_etal_04a.pdf, (accessed Oct. 20, 2006); S. Godbole and S. Sarawagi, "Discriminative Methods for Multi-Labeled Classification," in *Proceedings of the Pacific-Asia Conferences on Knowledge Discovery and Data Mining* (2004), http://www.it.iitb.ac.in/~shantanu/work/pakdd04.pdf, (accessed Oct. 20, 2006); L. Cai and T. Hofmann, "Hierarchical Document Categorization with Support Vector Machines," in *Proceedings of the ACM 13th Conference on Information and Knowl-*

edge Management (2004), http://www.cs.brown.edu/people/th/publications.html, (accessed Oct. 20, 2006); T. Hofmann et al., "Learning with Taxonomies: Classifying Documents and Words," in *Proceedings of the Workshop on Syntax, Semantics, and Statistics, Neural Information Processing* (2003), http://www.cs.brown.edu/people/ th/publications.html, (accessed Oct. 20, 2006); A. Tveit, "Empirical Comparison of Accuracy and Performance for the MIPSVM Classifier with Existing Classifiers," Technical Report, Division of Intelligent Systems, Department of Computer and Information Science, Norwegian University of Science and Technology. (Trondheim, Norway, 2003), http://www.idi.ntnu.no/~amundt/publications/2003/MIPSVMClassificationComparison.pdf, (accessed Oct. 20, 2006); C-W Hsu and C-J Lin, "A Comparison of Methods for Multi-Class Support Vector Machines," *IEEE Transactions on Neural Networks* 13, no. 2 (2002): 415–25, http://www.csie.ntu.edu.tw/~cjlin/papers/multisvm.pdf, (accessed Oct. 20, 2006).

19. P. Komarek, "Logistic Regression for Data Mining and High-Dimensional Classification" (Ph.D. thesis, Carnegie Mellon University, 2004), 138; P. Komarek and A. Moore, "Fast Robust Logistic Regression for Large Sparse Data Sets with Binary Outputs," *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics. January 3–6, 2003, Hyatt Hotel, Key West, Florida*, ed. By Christopher M. Bishop and Brendan J. Frey. http://research.Microsoft.com/conferences/AIStats2003/proceedings/174.pdf (accessed Nov. 23, 2006); A. Popescul et al., "Towards Structural Logistic Regression: Combining Relational and Statistical Learning," in *MRDM 2002: Workshop on Multi-Relational Data Mining*, http://www-ai.ijs.si/sasodzeroski/MRDM2002/proceedings/popesul.pdf (accessed Nov. 23, 2006); J. Zhang and Y. Yang, "Probabilistic Score Estimation with Piecewise Logistic Regression," in *Proceedings: Twenty-first International Conference on Machine Learning* (Menlo Park, Calif.: AAAI Press, 2004), http://www-2.cs.cmu.edu/~jianzhan/papers/icml04zhang.pdf, (accessed Oct. 20, 2006); Zhang et al., "Modified Logistic Regression"; Zhang and Oles, "Text Categorization"**;** Multi-class LR is discussed in Zhang et al., 2003, and Chang, 2003 (reference 18).

20. Some recent work on NB can be seen in J. Rennie, "Tackling the Poor Assumptions of Naive Bayes Text Classifiers," in T. Fawcett and N. Mishra, eds., *Proceedings of the 20th International Conference on Machine Learning* (Washington, D.C.: AAAI Pr., 2003), 616–23, http://haystack.lcs.mit.edu/papers/rennie.icml03.pdf, (accessed Oct. 20, 2006); K. Schneider, "Techniques for Improving the Performance of Naive Bayes for Text Classification," in *Computational Linguistics and Intelligent text processing: Sixth International Conference, CICLing2005, Mexico City, Mexico, February 13–19, 2005: Proceedings* (New York: Springer, 2005). (*Lecture Notes in Computer Science*, 3406). 682–93, http://www.phil.uni-passau.de/linguistik/schneider/pub/cicling2005.html, (accessed Oct. 20, 2006); E. Frank et al., "Locally Weighted Naive Bayes," in *Proceedings of the 19th Conference in Uncertainty in Artificial Intelligence* (Acapulco: Morgan Kaufmann, 2003), 249–56, http://www.cs.waikato.ac.nz/~eibe/pubs/UAI_200.ps.gz, (accessed Oct. 20, 2006); G. Webb et al., "Not so Naive Bayes: Aggregating One-Dependence Estimators," *Machine Learning* 58, no. 1 (Jan. 2005): 5–24, http://www.csse.monash.edu.au/~webb/Files/WebbBoughtonWang05.pdf, (accessed Oct. 20, 2006); E. Keogh and M. Pazzani, "Learning the Structure of Augmented Bayesian Classifiers," *International Journal on Artificial Intelligence Tools* 11, no. 4 (2002): 587–601, http://www.ics.uci.edu/~pazzani/Publications/tools.pdf (accessed Oct. 20, 2006).

21. McCallum and Jensen, "A Note on the Unification of Information Extraction and Data Mining"; Joachims, "SVM-light"; Y. Altun et al., "Hidden Markov Support Vector Machines," in *Proceedings of the 20th International Conference on Machine Learning* (Menlo Park, Calif.: AAAI Press, 2003), http://www.cs.brown.edu/people/th/publications.html (accessed Oct. 20, 2006); A. Ganapathiraju et al., "Hybrid SVM/HMM Architectures for Speech Recognition," in *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference* (Cambridge, Mass.: MIT Press, 2001), http://www.nist.gov/speech/publications/tw00/pdf/cp210.pdf (accessed Oct. 20, 2006); D. Freitag and A. McCallum, "Information Extraction with HMM Structures Learned by Stochastic Optimization," in *Proceedings of the 18th Conference on Artificial Intelligence* (Austin, TX.: AAAI Press, 2000) http://www.cs.umass.edu/~mccallum/papers/iehill-aaai2000s.ps (accessed Oct. 20, 2006); S. Basu et al., "A Probabilistic Framework for Semi-Supervised Clustering," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Seattle, Wash.: 2004), 59–68, http://www.cs.utexas.edu/users/ml/papers/semi-kdd-04.pdf, (accessed Oct. 20, 2006).

22. T. Liu et al., "Efficient Exact kNN and Nonparametric Classification in High Dimensions," in *Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference* (Cambridge, Mass.: MIT Press, 2001). http://www.autonlab.org/autonweb/showPaper.jsp?ID=Liu-knn, (accessed Oct. 20, 2006); G. Guo et al., "KNN Model-Based Approach in Classification," in *Lecture Notes in Computer Science*, vol. 2888 (Heidelberg: Springer Berlin, 2003), 986–96, http://www.icons.rodan.pl/publications/%5BGuo2003%5D.pdf (accessed Oct. 20, 2006)

23. Bouckaert and Frank, "Evaluating the Replicability of Significance Tests"; Bouckaert, "Estimating Replicability of Classifier Learning Experiments"; Caruana and Niculescu-Mizil, "Data Mining in Metric Space"; R. Caruana and T. Joachims, "PERF (Data Mining Evaluation Software)," in *Proceedings of the Conference on Knowledge Discovery and Data Mining* (2004). http://kodiak.cs.cornell.edu/kddcup/software.html (accessed Oct. 20, 2006)**;** Paynter, "Developing Practical Automatic Metadata."

24. Raina et al., "Classification with Hybrid Generative/Discriminative Models"; Bouchard and Triggs, "The Trade-Off Between Generative and Discriminative Classifiers."

25. Ibid; Zhang et al., "Modified Logistic Regression"; Chang, "Boosting SVM Classifiers with Logistic Regression"; Joachims,

"SVMlight"; L. Shih et al., "Not Too Hot, Not Too Cold: The Bundled SVM Is Just Right!" in *Proceedings of the ICML-2002 Workshop on Text Learning* (2002). http://people.csail.mit.edu/u/j/jrennie/public_html/papers/icml02-bundled.pdf (accessed Oct. 20, 2006); F. Fukumoto and Y. Suzuki, "Manipulating Large Corpora for Text Classification," in *Proceedings of the Conference on Empirical Methods in Natural-Language Processing* (Philadelphia: Association for Computational Linguistics, 2002), 196–203, http://acl.ldc.upenn.edu/W/W02/W02-1026.pdf (accessed Oct. 20, 2006); Altun et al., "Hidden Markov Support Vector Machines; Ganapathiraju et al., "Hybrid SVM/HMM Architectures"; Liu et al., "Efficient Exact k-NN"; A. Ng and M. Jordan, "On Discriminative versus Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes," in *Advances in Neural Information Processing Systems 14: Proceedings of the 2001 Conference* (Cambridge, Mass.: MIT Press, 2002), http://www.robotics.stanford.edu/~ang/papers/nips01-discriminativegenerative.ps (accessed Oct. 20, 2006); K. Nigam et al., "Text Classification from Labeled and Unlabeled Documents Using EM," *Machine Learning* 39, nos. 2/3 (2000): 103–34, http://www.kamalnigam.com/papers/emcat-mlj99.pdf (accessed Oct. 20, 2006).

26. G. Valentini and F. Masulli, "Ensembles of Learning Machines," in *Neural Nets WIRN Vietri-02, Series Lecture Notes in Computer Sciences*, M. Marinaro and R. Tagliaferri, eds. (Heidelberg: Springer-Verlag, 2002), http://www.disi.unige.it/person/MasulliF/papers/masulli-wirn02.pdf (accessed Oct. 20, 2006).

27. Ibid.; R. Caruana et al., "Ensemble Selection from Libraries of Models" in *Proceedings: Twenty-first International Conference on Machine Learning* (Menlo Park, Calif.: AAAI Press, 2004). http://www.cs.cornell.edu/~alexn/shotgun.icml04.revised.rev2.pdf (accessed Oct. 20, 2006); G. Tsoumakas, "Effective Voting of Heterogeneous Classifiers," in *Machine Learning ECML 2004: 15th European Conference on Machine Learning, Pisa, Italy, September 20–24, 2004: Proceedings*. (Berlin, New York: Springer, 2004), http://users.auth.gr/~greg/Publications/tsoumakas-ecml2004.pdf (accessed Oct. 20, 2006); J. Fürnkranz, "On the Use of Fast Sub-Sampling Estimates for Algorithm Recommendation," Technical Report TR-2002-36 (Wien: Österreichisches Forschungsinstitut für Artificial Intelligence, 2002), http://www.ofai.at/cgi-bin/get-tr?paper=oefai-tr-2002-36.pdf (accessed Oct. 20, 2006); A. Seewald, 2002. "Meta-Learning for Stacked Classification," (extended version) in *Proceedings of the 2nd International Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support, and Meta-Learning* (University of Helsinki, Department of Computer Science, Report B-2002-3, 2002), http://www.ofai.at/cgi-bin/get-tr?download=1&paper=oefai-tr-2002-05.pdf (accessed Oct. 20, 2006); A. Seewald and J. Fürnkranz, "An Evaluation of Grading Classifiers," in *Advances in Intelligent Data Analysis: Proceedings of the 4th International Symposium* (Lisbon, Portugal: Springer-Verlag, 2001), http://www.ofai.at/cgi-bin/get-tr?paper=oefai-tr-2001-01.pdf (accessed Oct. 20, 2006); P. Bennett et al., "The Combination of Text Classifiers Using Reliability Indicators," Technical Report. *Microsoft and Information Retrieval* 8, no. 1 (2005): 67–100, http://research.microsoft.com/~horvitz/tclass_combine.pdf (accessed Oct. 20, 2006); Y. Kim et al., "Optimal Ensemble Construction via Meta-Evolutionary Ensembles," *Expert Systems With Applications* 30, no. 4 (in press 2006), http://www.informatics.indiana.edu/fil/Papers/mee-eswa.pdf (accessed Oct. 20, 2006).

28. S. Godbole, "Document Classification as an Internet Service: Choosing the Best Classifier" (masters thesis, IIT Bombay, 2001). http://www.it.iitb.ac.in/~shantanu/work/mtpsg.pdf (accessed Oct. 20, 2006).

29. K. Liu and H. Kargupta, "Distributed Data Mining Bibliography: Release 1.7," (Baltimore: University of Maryland, Computer Science Department, 2006), http://www.csee.umbc.edu/~hillol/DDMBIB/ (accessed Oct. 20, 2006); A. Prodromidis and P. Chan, "Meta-Learning in Distributed Data Mining Systems: Issues and Approaches," in *Advances of Distributed Data Mining*, Hillol Kargupta and Philip Chan, eds. (Menlo Park, Calif.: AAAI/MIT Press, 2000). http://www1.cs.columbia.edu/~andreas/publications/DDMBOOK.ps.gz (accessed Oct. 20, 2006); G. Tsoumakas and I. Vlahavas, "Distributed Data Mining of Large Classifier Ensembles," in *Methods and Applications of Artificial Intelligence: Second Hellenic Conference on AI, SETN 2002, Thessaloniki, Greece, April 11–12, 2002: Proceedings*, (Berlin, New York: Springer, 2002), 249–56, http://users.auth.gr/~greg/Publications/ddmlce.pdf (accessed Oct. 20, 2006); R. Khoussainov et al., "Grid-Enabled Weka: A Toolkit for Machine Learning on the Grid," *ERCIM News* no. 59, (Oct. 2004), http://www.ercim.org/publication/Ercim_News/enw59/khussainov.html (accessed Oct. 20, 2006).

30. S. Godbole et al., "Document Classification through Interactive Supervision of Document and Term Labels," in *Knowledge Discovery in Databases: PKDD 2004: 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, Pisa, Italy, September 20–24, 2004: Proceedings* (Berlin; New York: Springer, 2004), http://www.it.iitb.ac.in/~shantanu/work/pkdd04.pdf (accessed Oct. 20, 2006).; H. Yu et al., "PEBL: Positive Example Based Learning for Web Page Classification Using SVM," in *KDD-2002: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (New York: ACM Pr., 2002), 239–48, http://eagle.cs.uiuc.edu/pubs/2002/pebl-kdd02.pdf (accessed Oct. 20, 2006); T. Kristjannson et al., "Interactive Information Extraction with Constrained Conditional Random Fields," in *Proceedings: Nineteenth National Conference on Artificial Intelligence (AAI-04)* (Menlo Park, Calif.: AAAI Press; Cambridge, Mass.: MIT Press, 2004), http://www.cs.umass.edu/~mccallum/papers/addrie-aaai04.pdf (accessed Oct. 20, 2006); V. Tablan et al., "OLLIE: On-Line Learning for Information Extraction," in *Proceedings of the HLT-NAACL Workshop on Software Engineering and Architecture of Language Technology Systems: Edmonton, Canada: 2003.* (New York: ACM, 2003), http://gate.ac.uk/sale/hlt03/ollie-sealts.pdf (accessed Oct. 20, 2006).

31. Godbole et al., "Document Classification."

32. Ibid.; Tablan et al., "OLLIE: On-Line Learning for Information Extraction."

33. G. Mann et al., "Bibliometric Impact Measures," (in press).

34. Bouckaert and Frank, "Evaluating the Replicability of Significance Tests"; Caruana and Niculescu-Mizil, "Data Mining in Metric Space"; Caruana and Joachims, "PERF (Data Mining Evaluation Software)."

35. Mann et al., "Bibliometric Impact Measures"; Matsuo et al., "KeyWorld"; Matsuo and Ishizuka, "Keyword Extraction from a Single Document"; Lee et al., *Retrieval and Organizing Web Pages*; Tajima et al., "Discovery and Retrieval of Logical Information." (See also the sections on Hybrid, Unified Models, and Document Scale Learning and Classification, above.)

36. Menczer, "Mapping the Semantics of Web Text and Links."

37. P. Srinivasan et al., "A General Evaluation Framework for Topical Crawlers," *Information Retrieval* 8, no. 3 (2005): 417–47, http://www.informatics.indiana.edu/fil/Papers/crawl_framework.pdf (accessed Oct. 20, 2006); A. Maguitman et al., "Algorithmic Computation and Approximation of Semantic Similarity," (in press, 2006). To appear in *World Wide Web Journal.* http://www.informatics.indiana.edu/fil/Papers/semsim_extended.pdf (accessed Oct. 20, 2006).

38. ArXiv. Cornell University Library, http://arxiv.org/ (accessed Oct. 20, 2006); CiteSeer.IST (formerly ResearchIndex), http://citeseer.ist.psu.edu/ (accessed Oct. 20, 2006); eScholarship Repository, California Digital Library, http://repositories.cdlib.org/escholarship/, (accessed Oct. 20, 2006); National Science Foundation, National Science Digital Library, http://nsdl.org/ (accessed Oct. 20, 2006); OAIster. Digital library production service (University of Michigan), http://oaister.umdl.umich.edu/o/oaister/ (accessed Oct. 20, 2006); U.S. Institute of Museum and Library Services. Digital collections and content, http://imlsdcc.grainger.uiuc.edu/ (accessed Oct. 20, 2006).

39. K. Calhoun, "The Changing Nature of the Catalog and Its Integration into Other Discovery Tools," (report to the Library of Congress, Mar. 17, 2006), http://www.loc.gov/catdir/calhoun-report-final.pdf (accessed Oct. 20, 2006); Mitchell, "Collaboration Enabling Internet Resource Collection-Building Software and Technologies"; W. Wulf, "Higher Education Alert: The Railroad is Coming," in *EDUCAUSE, Publications from the Forum for the Future of Higher Education (2002),* http://www.educause.edu/ir/library/pdf/FFPIU022.pdf (accessed Oct. 20, 2006).

40. University of California Libraries, "Rethinking How We Provide Bibliographic Services at the University of California," final report of the Bibliographic Services Task Force of the University of California Libraries, 2005, http://librariesuniversityofcalifornia.edu/sopag/BSTF/Final.pdf (accessed Oct. 20, 2006).

41. L. Dempsey, "Libraries and the Long Tail: Some Thoughts About Libraries in a Network Age," *D-Lib Magazine* 12, no. 4 (2006), http://www.dlib.org/dlib/april06/dempsey/04dempsey.html (accessed Oct. 20, 2006).

42. Mason**,** J. et al., "INFOMINE: Promising Directions in Virtual Library Development," *First Monday* 5, no. 6 (June 5, 2000), http://www.firstmonday.dk/issues/issue5_6/mason/ (accessed Oct. 20, 2006).

43. E. O'Neill and L. M. Chan, "FAST: Faceted Application of Subject Terminology," in *Proceedings of the World Information Congress, IFLA General Conference and Council* (Berlin: IFLA, 2003). http://www.ifla.org/IV/ifla69/papers/010e-ONeill_Mai-Chan.pdf (accessed Oct. 20, 2006); See also: OCLC 2003–2006, "FAST: Faceted Application of Subject Terminology," http://www.oclc.org/research/projects/fast/default.htm) (accessed Oct. 20, 2006).

44. M. Bates, 2003, "Improving User Access to Library Catalog and Portal Information," *Task Force Recommendation 2.3, Final Report* (Washington, D.C.:Library of Congress, 2003), 30, http://www.loc.gov/catdir/bibcontrol/2.3BatesReport6-03.doc.pdf (accessed Oct. 20, 2006).

45. RDN (Resource Description Network), http://www.rdn.ac.uk/projects/eprints-uk/, (accessed Oct. 20, 2006); OCLC "ePrints-UK" (2005), http://www.oclc.org/research/projects/mswitch/epuk.htm, (accessed Oct. 20, 2006).

46. A. MacEwan, "Working with LCSH: The Cost of Cooperation and the Achievement of Access: A Perspective from the British Library," presented at the *IFLA General Conference, 1998,* http://www.ifla.org/IV/ifla64/033-99e.htm (accessed Oct. 20, 2006).

47. Ibid.; R. Larson, "The Decline of Subject Searching: Long-Term Trends and Patterns of Index Use in an Online Catalog," *Journal of the American Society for Information Science* 42, no. 3 (1991): 197–215.

48. K. Drabenstott et al., "End-User Understanding of Subject Headings in Library Catalogs," *Library Resources & Technical Services* 43, no. 3 (Jul. 1999): 140–60; Bates, "Improving User Access."

49. Bates, "Improving User Access," (see discussion of entry vocabulary).

50. Ibid.

51. BEAT (Bibliographic Enrichment Advisory Team, Library of Congress), "Digital Tables of Contents," (2005), http://www.loc.gov/catdir/beat/digitoc.html (accessed Oct. 20, 2006).

52. D. Vizine-Goetz, "Terminology Services, OCLC," (2004), http://www.oclc.org/research/projects/termservices/default.htm (accessed Oct. 20, 2006).

53. C. Fellbaum, *Wordnet: An Electronic Lexical Database* (Cambridge, Mass.: MIT Pr., 1998), http://wordnet.princeton.edu/ (accessed Oct. 20, 2006); A. Csomai, "Wordnet Bibliography," (2006). http://lit.csci.unt.edu/~wordnet/ (accessed Oct. 20, 2006).

54. Bates, "Improving User Access."

55. A. Maedche and R. Volz, "The Ontology Extraction and Maintenance Framework: Text-to-Onto," in *Proceedings of the ICDM 2001 Workshop* (San Jose, Calif.: IEEE Computer Society (2001), http://cui.unige.ch/~hilario/icdm-01/DM-KM-Final/Volz.pdf (accessed Oct. 20, 2006); V. Parekh, J. Gwo, and T. Finin, "Mining Domain Specific Texts and Glossaries to Evaluate and Enrich Domain Ontologies," in *Proceedings of the 2004 International Conference on Information and Knowledge Engineering: IKE '04* (Las Vegas: CSREA Press, 2004), http://ebiquity.umbc.edu/v2.1/paper/html/id/171/ (accessed Oct. 20, 2006);

D. Sleeman et al., "Enabling Services for Distributed Environments: Ontology Extraction and Knowledge Base Characterization," in *Proceedings of Workshop on Knowledge Transformation for the Semantic Web/Fifteenth European Conference on Artificial Intelligence* (Lyon, France: ECAI, 2002), http://www.csd.abdn.ac.uk/~sleeman/published-papers/p129-final-ontomine.pdf (accessed Oct. 20, 2006). ; B. Omelayenko, "Learning of Ontologies for the Web: The Analysis of Existent Approaches," in *Proceedings of the International Workshop on Web Dynamics* (London: WebDyn, 2001), http://dcs.bbk.ac.uk/webdyn/webDynPapers/omelayenko.pdf (accessed Oct. 20, 2006); R. Dhamankar et al., "Imap: Discovering Complex Semantic Matches Between Database Schemas," in *SIGMOD 2004: Proceedings of the ACM SIGMOD International Conference on Management of Data, June 13–18, 2004, Paris, France* (New York: Association for Computing Machinery, 2004), http://www.cs.washington.edu/homes/pedrod/papers/sigmod04.pdf (accessed Oct. 20, 2006); P. Cassin et al., "Ontology Extraction for Educational Knowledge Bases," *Lecture Notes in Computer Science,* vol. 2926 (Heidelberg: Springer-Verlag, 2004), 297–309; *Revised and Invited Papers from Agent-Mediated Knowledge Management: International Symposium* (Stanford, Calif., Mar. 24–26, 2003), ftp://mas.cs.umass.edu/pub/Cassin_Ontology-AMKM03.pdf (accessed Oct. 20, 2006); T. Wang et al**.,** "Extracting a Domain Ontology from Linguistic Resource Based on Relatedness Measurements," in *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence: Proceedings: September 19–22, Compiègne University of Technology, France* (Los Alamitos, Calif.: IEEE Computer Society, 2005), 345–51, http://csdl2.computer.org/persagen/DLAbsToc.jsp?resourcePath=/dl/proceedings/&toc=comp/proceedings/wi/2005/2415/00/2415toc.xml&DOI=10.1109/WI.2005.63 (accessed Oct. 20, 2006).

56. Bates, "Improving User Access to Library Catalog and Portal Information."

57. O'Neill and Chan, "FAST: Faceted Application of Subject Terminology."

58. Mason, et al., "INFOMINE: Promising Directions in Virtual Library Development."