Collaboration Enabling Internet Resource Collection-Building Software and Technologies

Steve Mitchell

Abstract

Over the last decade the Library of the University of California, Riverside and its collaborators have developed a number of systems, service designs, and projects that utilize innovative technologies to foster better Internet finding tools in libraries and more cooperative and efficient effort in Internet link and metadata collection building. The open-source software and projects discussed represent appropriate technologies and sustainable strategies that we believe will help Internet portals, digital libraries, virtual libraries, library catalogs-with-portal-like-capabilities (IPDVLCs), and related collection-building efforts in academia to better scale and more accurately anticipate and meet the needs of scholarly and educational users.

Our work and its intent is best introduced by providing an overview of the projects, services, and software that we have been working on for the last several years: iVia, INFOMINE, and Data Fountains. iVia will be described in depth from the standpoints of its overall system, content and uses supported, end-user features, content development and management features for institutional collaborators, features for individual expert content builders, and incentives for collaborative collection building.

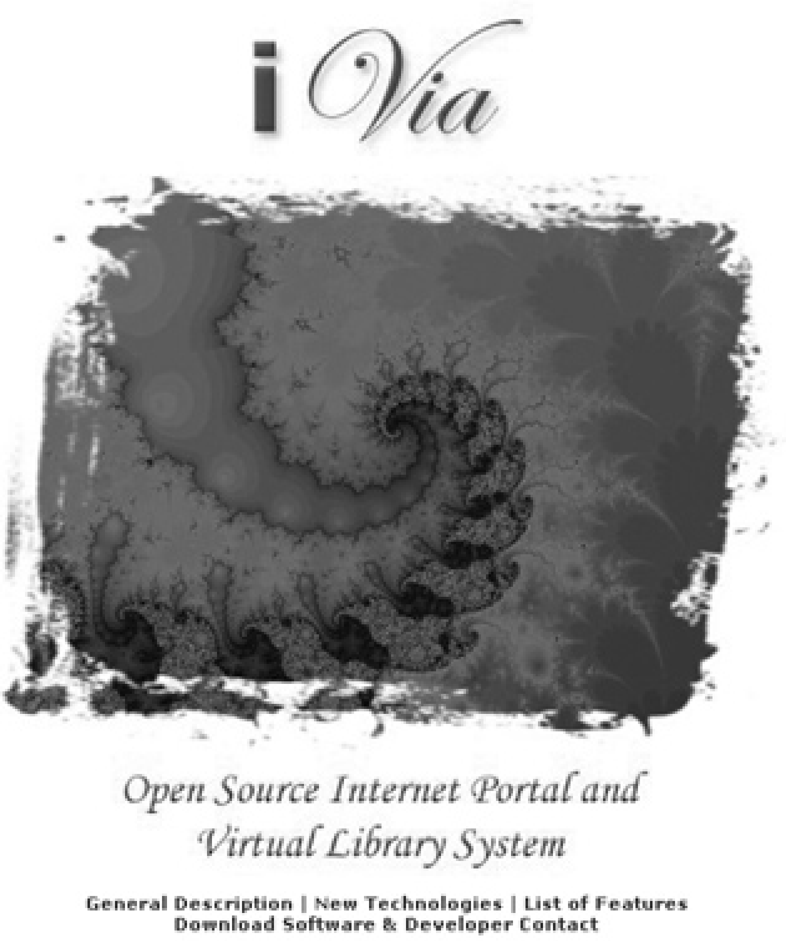
iVia

iVia (http://infomine.ucr.edu/iVia/) is a portal or virtual library collection-building software platform (Mitchell et. al., 2003). It was designed to support multiple institutions and projects in collaborative collection-building efforts. The system (or components) is used by INFOMINE and

Steve Mitchell, iVia and Data Fountains Projects Coordinator, Science Library, University of California at Riverside, Riverside, CA 92521
LIBRARY TRENDS, Vol. 53, No. 4, Spring 2005 ("The Commercialized Web: Challenges for Libraries and Democracy," edited by Bettina Fabos), pp. 604–619

the National Science Digital Library (NSDL) of the National Science Foundation, among others. The software, written primarily in C++, is licensed as open source and is available to all. iVia features a very large number of custom-configurable user interfaces and information retrieval options to support the institutional identity management (that is, branding) and user finding needs of diverse, collaborating organizations. Institutional collaborators will also be able to avail themselves of multiple metadata creation options, including support for multiple "production lines" and levels of editorial control. Resource- and labor-saving machine assistance is featured and used to semi- and/or fully-automate a number of tasks in both Internet resource identification and metadata generation. The

Figure 1.


iVia
Open Source Internet Portal and
Virtual Library System
General Description | New Technologies | List of Features
Download Software & Developer Contact

former is made possible through new work in focused crawling and the latter through innovations in automated classification (which include the assignment of Library of Congress Subject Headings [LCSH] and Library of Congress Classifications [LCC]). iVia support has come from the Library of the University of California at Riverside, the U.S. Institute of Museum and Library Services (IMLS), NSDL, and the Fund for the Improvement of Post-Secondary Education of the U.S Department of Education (FIPSE).

INFOMINE

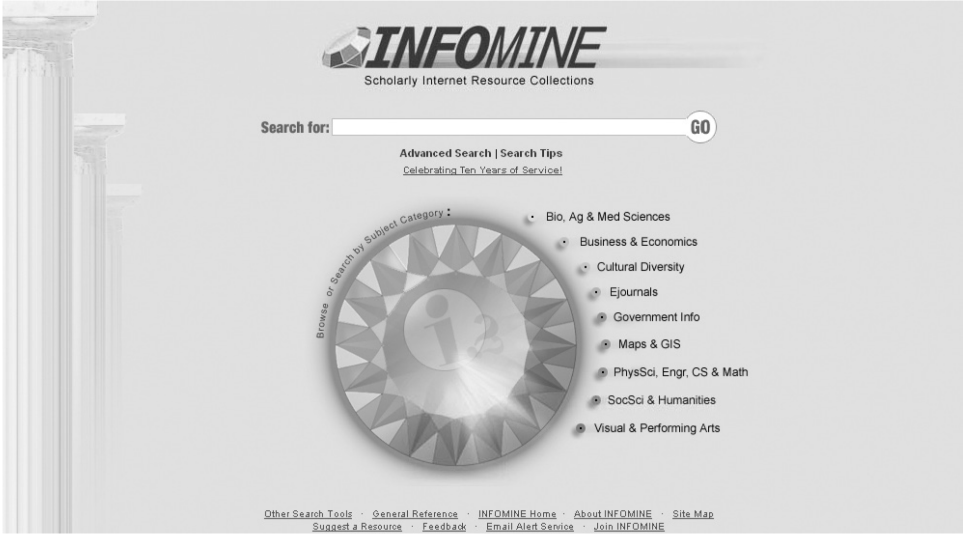
The INFOMINE (http://infomine.ucr.edu) virtual library service was conceived from inception as a multi-institutional, collaborative effort and has served the academic community since 1994. It has the mission of identifying, describing, and therefore making visible and useful to the academic community the significant scholarly and educational resources on the Internet. More than 230,000 resources populate the collection. These represent all major academic research disciplines and are the product of the collaborative efforts of librarians, faculty, and graduate students at the University of California (Riverside, Los Angeles, Santa Cruz, and Irvine campuses), Wake Forest University, and California State University (Fresno and Sacramento campuses).

INFOMINE draws upon a hybrid collection design that consists of metadata created by (1) subject experts (at INFOMINE and at collaborating institutions); (2) machine processes or machine processes with expert refinement; and (3) external collaborating institutions that share data streams of records, which are imported through OAI-PMH or other means, translated as needed, and then added to the INFOMINE collection (for example, MARC records of the University of California Shared Cataloging Project and Dublin Core records from some collections within the NSDL). INFOMINE represents a rich collection of records with rich metadata. For example, the number of subject and keyword terms applied in expert-created records that describe resource themes are much more numerous than in standard library catalogs. INFOMINE is used for both end-user searching and collection development on the part of other Internet portals, digital libraries, virtual libraries, and library catalogs-with-portal-like-capabilities (IPDVLCs). It uses iVia software as its system platform. INFOMINE support has come from the Library of the University of California at Riverside and the collaborating libraries mentioned above, as well as from IMLS, NSDL, and FIPSE.

Data Fountains

Data Fountains (http://infomine.ucr.edu/Data_Fountains/) is an open-source software system and a service for automated or semi-automated Internet resource discovery and metadata generation. Based in the

Figure 2.


INFOMINE
Scholarly Internet Resource Collections
Search for:
GO
Advanced Search | Search Tips
Celebrating Ten Years of Service!
Browse or Search by Subject Category:
i2
Bio, Ag & Med Sciences
Business & Economics
Cultural Diversity
Ejournals
Government Info
Maps & GIS
PhysSci, Engr, CS & Math
SocSci & Humanities
Visual & Performing Arts
Other Search Tools · General Reference · INFOMINE Home · About INFOMINE · Site Map
Suggest a Resource · Feedback · Email Alert Service · Join INFOMINE

iVia system, it expands beyond iVia considerably by creating an array of independent, though federated, collection-building systems for collaborating projects with the goal of generating the basic "ore" (links to important Internet resources and associated metadata records and rich full text) for these projects. It also improves upon core crawling and classification techniques. Each collaborating project and/or subject community works with and fine tunes its own Data Fountain, that is, its own set of focused crawler(s) and classifier(s). The records and full text derived are exported to and utilized within the collaborator's own native interface, backend system, and databases. In iVia these crawlers and classifiers are shared, as is the backend. Expert-machine interaction, which relies upon the subject domain expertise and the wisdom and conventions in collection building of participating librarians, is emphasized more in Data Fountains than currently in iVia and should result in more accurate content. That is, semi-automated approaches are more fully designed into and featured in the system and are critical to improving its performance. Given that Data Fountains is currently under development, much of the following instead addresses iVia, its close relative. Data Fountains work is supported by IMLS and the Library of the University of California at Riverside. Please contact us if you are interested in implementing Data Fountains in your project.

Figure 3.

Data Fountains
A National, Cooperative Information Utility for Shared Internet Resource Discovery, Metadata Application and Rich, Full-text Harvest of Value to Internet Portals, Virtual Libraries and Library Catalogs with Portal-like Capabilities

Collaborative Service and Participatory Technology Development

We designed the technology behind the INFOMINE, iVia, and Data Fountains projects to enable and facilitate cooperative service building and effort. That is, we wanted the technology providing the foundation for these systems to be collaborative and participatory and to gain significant increases in accuracy and resource savings through this. While the system strongly supports fully automated and fully manual processes for collection building, the technology also supports semi-automated processes emphasizing interactive subject domain expertise. We see our work as building machine-assisted IPDVLC community-ware. We are developing and bringing to the library community new, machine learning–based technologies that are

- *Enabling:* These technologies provide systems that scale better in the Internet environment and save expert labor and other resources. They enable collaborative efforts of many types at the same time that they are supportive of multiple modes of collection building and user access. These technologies also enable us to reduce redundant effort by better distributing collection and metadata development efforts among similar projects.

- *Participatory:* Collaborating institutions, as co-designers, participate in developing and customizing the software to fit their needs (for example, in interface, data views/landscapes, record creation, and retrieval). Collaborators work in codesigning systems that emphasize identifying, enhancing, and/or developing synergies among collaborating projects. This is done as well by identifying promising expert/machine processes/interactions that will augment and improve the performance of both. Experts actively participate in improving machine processes and vice versa.
- *Supportive of Librarian Community Expertise, Values, and Effort:* These technologies help amplify and facilitate the transfer of academic librarian subject expertise, organizing expertise, public domain orientation, objectivity, service orientation, and other scholarly and educational community values and capabilities into efficient and effective Internet-based information. Tools such as iVia allow us to build very useful collections that are based on and express our considerable wealth of knowledge in subject domains, fully featured interfaces, sophisticated (that is, precise) user access, and rich, well-organized metadata. While Google-level accuracy and approaches suffice for many information-finding needs, they do not generally serve the in-depth finding needs of academics. Google may partially "disintermediate" the role of the expert librarian in some areas, but, in the long term, this will not extend to areas where superior information quality, sophisticated access, and accurate provenance verification are critical to major research and fact-finding efforts. It is incumbent upon the library community to work with this technology, to adapt it to its needs, and to come to own it just as physical collections usually own the facilities in which they are located. This is what our projects are about: bringing public domain community-ware and machine-learning technology in resource discovery and metadata generation, among other areas, into the library.

Focus on iVia—An Open-Source Software Platform for Collaborative Internet Collection Building

Hardware

The following hardware supports the INFOMINE application of iVia:

- Public search interface server: end-user and content-builder (including expert-guided crawler) interfaces are supported
- Public search interface server backup

- Database server (both the metadata and full-text databases are here)
- Database server backup
- Crawler/classifier processes server (for example, vlcrawler, Nalanda iVia focused crawler)
- OAI import/export server
- Additional mass storage equipment: 2 terabytes of storage including a RAID array (1 terabyte of storage) accessible via Network File System (NFS) (networked storage)

A standard machine would be an AMD XP 3200+ CPU, 1.5 GiB of high-speed RAM, and an 80GB disk storage.

Software

iVia software is licensed as open source (GNU GPL and LGPL). Open-source software is free software intended to be of use to and be further developed and refined by its users. In iVia's case this would be users in the library and Internet Portal community. The open-source approach enables institutions to pool resources and inexpensively develop and refine software that meets their needs. In fact, in addition to the software we have developed, our system is based on many very successful and well-known open-source packages, including the Linux operating system (including Debian, RedHat, and Suse variants), MySQL and Berkeley DB databases management packages, and Apache Web server software.

iVia code is in C++, this being one of the most powerful, flexible, and standardized of programming languages. Some of our interface code is in Java. Currently the iVia program size is close to 10 Mb (>230k lines).

Standards

iVia is based in standards. Metadata standards include Dublin Core and MARC (we use Dublin Core but can translate from/to MARC). Subject schema standards include Library of Congress Subject Headings (LCSH) and Library of Congress Classifications (LCC), these long being standards in the U.S. academic library. Using these will eventually allow iVia, as finding tool software, seamless subject access (no translations involved) to both the Internet and print records of knowledge. For data transfer among collaborators, iVia uses the Open Archives Initiative (OAI-PMH) approach as well as standard delimited formats (SDF). OAI-PMH is used as well internally to transfer/harvest records from our crawling and classification databases and our user databases.

Fields Supported

Forty-seven fields are supported in our database. Of most direct value to users are URL, title, alternative title, creator (author), subject—LCSH, subject—LCC, keywords, description, selected full-text (1–3 pages of rich text), MyI (a field that helps institutions create custom data views), and local URL (often of value for collaborators in accessing fee-based material). Other fields of note and their functions are general subject categories (for example, biological, agricultural, and medical sciences); created at; created by; modified by; last modified by; access restrictions; restricted to; publisher; audience levels; resource types; language; coverage begin; and coverage end.

Content Managed

Format types represented through iVia include HTML resources and, shortly, PDF, Postscript, and others. Metadata as well as representative, rich full text is generated or harvested from the resource being described and makes up the content of our databases. This data represents free and fee-based resources and includes resource types as varied as digital libraries, other virtual libraries and portals, e-journals, e-books, e-print archives, databases, hypertext fiction, maps, and more. Content retrieval is robust and quick. Berkeley DB indexing capabilities are used to augment performance through MySQL.

iVia Uses

Major applications of iVia to date have included INFOMINE, one of the first Web-based services offered by a library. INFOMINE (an Internet resources virtual library–type finding tool) has been supported by iVia in serving academic researcher and student end users both nationally and at specific institutions (for example, the University of California at Riverside and Wake Forest University). Collection development for others has been another major function, with many other academic virtual libraries using iVia/INFOMINE as a resource discovery service for their own collection-building efforts. iVia/INFOMINE is also used by librarians in creating Web-based subject guides or pathfinders in various subjects (this is facilitated through using our “canned search” generator and MyI field), as well as by faculty creating Web resource modules on their course pages in support of curriculum units.

While INFOMINE has been the major application of iVia so far, with most aspects of iVia as described in this article being applied in INFOMINE, we have been working with the National Science Digital Library (NSDL) to develop an NSDL iVia. Among the major goals of this project are the integration of our Web crawlers and classification software into NSDL’s core system for purposes of open Internet resource discovery and related classification (that is, resource identification and metadata generation). Just as crucial here will be the use of this software to generate metadata for existent, “deep Web” collections (for example, article databases or e-print collections or other databases where access is through a search front-end) in many different document formats other than HTML.

iVia User Features

Through the INFOMINE application, iVia has demonstrated sophisticated and flexible user features geared toward varying levels of searching expertise. Most searchers will use defaults that are transparent to them as they use the basic search (http://infomine.ucr.edu/). Librarians, information specialists, and researchers may choose to use the many user configurable features found in Advanced Search and Browse (http://infomine.ucr.edu/cgi-bin/search). Advanced Search and Browse features are present in each individual collection (for example, http://infomine.ucr.edu/cgi-bin/search?category=bioag).

In more detail, iVia's search and browse features include the following: multiple subject and resource type collections or categories, including Biological, Agricultural and Medical Sciences; Business and Economics; Cultural and Ethnic Diversity; E-journals; Government Information; Maps and GIS; Physical Sciences, Engineering, Computer Science, and Math; Social Sciences and Humanities; and Visual and Performing Arts.

The availability of standardized, fielded metadata, as well as rich full-text, enables advanced searching capabilities including Boolean (for example, and, or, not) and Proximity operators (for example, near 1–20); exact searching using quotes or stem searching using asterisk; nested searching using parentheses; and various types of limit searching. One can limit to expert or expert plus robot-originated records (the latter being those that have been automatically identified and described), or combine general subject categories (for example, BioAgMed or E-journals), any combination of fields (for example, title, keywords, subjects, and/or description, and so on), resource type (for example, article databases, electronic journals, or e-print collections), and/or type of access to resource (such as free, fee, or a mix).

In iVia, search interfaces are presented on the bottom of each results page if search modification is desired. In the event of zero result searches, spelling is checked and possible spelling alternatives are suggested. Finally, in full display, most indexing terms are presented as links, which can be clicked on to narrow or broaden a user's search.

Browse indexes are available for both all subject categories and individual subject categories. Specific browse indexes are available for titles, creators (including authors), subjects—LCSH, subjects—LCC, keywords (these often include minor subjects and lay-person terminology), resource types (for example, standards, style manuals) and Whats New! (that is, recent expert additions to the collection).

Records are displayed in three formats: title only, regular (title, description, and origin of record as either expert or robot created), and long (accessed by clicking on "More Info" in the full display). The latter includes a great number of fields of interest to users or collection builders including URL, title, description, broad subject categories, creators, subject—LCSH, subject—LCC, keywords, access, audience level (academic, K–12, or lifelong learner), institutional owner (which collaborator contributed the record if expert in origin), URL checker information, and INFOMINE collection information (mostly for record keeping: who added, who modified, record number, record origin). Results pages can be displayed in groups of 30, 50, or 100. They can be ordered alphabetically by title or by relevance to the query as judged by how many query terms were hits, how many were hits in major or minor fields (for example, title being more highly weighted than keyword, which is more highly weighted than full text), and whether terms in a specified phrase were found in exact or approximate adjacency.

iVia Content Development and Management—Features, Tools, and Machine Assistance for Institutional Collaborators and Expert Content Builders

iVia emphasizes numerous innovations for improving and making more efficient collection development and management efforts for both individual or multiple collaborating projects. These translate into significant labor and resource savings in building collections. These innovations can be best understood from the standpoints of institutional collaborators and individual experts creating new content, as detailed below.

Support for Institutional Collaborators

Institutional identity management or branding is important for iVia collaborators. Access to collaborative resources needs to reflect, within reason, the established ongoing Web presence and interface of the collaborating institution. To this end iVia provides multiple interfaces and methods of accessing data in collections it supports. The user interfaces and desired data views of collaborator project sites are supported. For example, the interface that the user is accessing from can be detected by iVia, which activates searching and other interface capabilities that meet existent profiles set up for this by the collaborating institution. Access is also enabled for selected external collections that rely on metasearching.

Custom Data Views and Access Supported iVia provides pre-constructed interface modules that can be quickly assembled and customized by collaborators in building interfaces to iVia data. These interface modules reflect the themes and presentation of the collaborating project while still taking full advantage of unique iVia retrieval and other user features. The suite of programs that facilitate this is known as "Theme-ing." Special fields, such as MyI (which allows institutions to create custom data views), support Theme-ing and custom interface access. For example, retrieval filters can be created by participating institutions to channel user searches through selected subsets of iVia data (for example, perhaps only the records for fee-based resources in the collection that have been subscribed to by the particular institution). This is done by identifying and tagging, in the MyI field, those records that the institution wants its users to view. Parallel fields are also supported for similar reasons. For example, some collaborators want short descriptions and others long. Hence, there are two, parallel, description fields. Users coming from the institution desiring short descriptions will see only these.

Metasearching Access iVia also enables access to its content through the interfaces of selected, completely external finding tools, which rely on general methods of metasearching. For example, the Ex Libris online public access library catalog system provides access to INFOMINE content, as does the California Digital Library Searchlight system. The nice thing about metasearching is that large numbers of diverse collections from multiple projects can be searched simultaneously. However, significant downsides exist because of the need to include generally very simplified, lowest common denominator searching of only the shared fields among the databases searched, which can be very few; this eliminates search access to unique, useful fields. Another problem is the limited ability to eliminate duplicate, overlapping results returned from the databases searched.

Multiple Modes of Content-Building Supported Even if collaborating institutions have been building Internet resource collections for some time and have established ways or styles of doing things, iVia takes this into account by providing multiple means for new collaborators to ramp up and begin creating content in ways with which they are comfortable. To this end iVia supports from one to three levels of editorial review as well as a pending record database that holds records in the process of being built and reviewed prior to their being approved and moved to the main working database. Some collaborators use just one level of review, that of the editor of the subject file (for example, the BioAgMed file in INFOMINE). Others have developed a well-defined division of labor whereby catalogers review the subject content of records created by public service librarians or metadata specialists prior to review by the editor of the subject file.

Similarly, in support of various divisions of labor and optimum utilization of staff with varying skill sets, each content builder can be assigned a different level of access to iVia content-building features. Managing editors of a subject file have full permission of many kinds, including batch deletes and batch changes, to the content of the whole database. Metadata specialists, on the other hand, may only be allowed to add content to the pending record database, with their records going through multiple levels of review before being added, by the subject file editor, to the working database.

Hybrid Collections of Heterogeneous Metadata—Support for Multiple Incoming Data Streams and Types of Records Just as one of the main benefits of collaboration in mutual content building is sharing the collection development load among participants, iVia also makes it possible to utilize the work of other collection-building projects that choose to not be an integral part of the project. To do this, iVia has a hybrid collection design that supports diverse, heterogeneous record types and record origins (Mason, Mitchell, Mooney, Reasoner, & Rodriguez, 2000).

As manifested in the INFOMINE application of iVia, the system builds content by ingesting and threading together a number of diverse data streams. The first of these is, of course, the records created within the iVia system by experts. Sources for these currently include content builders from the University of California at Riverside, UCLA, and individuals from other UCs; Wake Forest University; and California State University at Fresno and Sacramento. There are about 20,000 of these expert-built records internally created for and through INFOMINE's iVia system. INFOMINE's iVia also imports and, as needed, translates from collaborating external data streams. For example, MARC records for Internet resources cataloged by the UC Shared Cataloging Project (SCP) are imported, translated to Dublin Core, and utilized (about 25,000 records in INFOMINE are of this origin). Through collaborators at UC Santa Cruz, Lexis Nexis serial titles are imported (accounting for close to 6,000 records). INFOMINE's iVia also uses OAI-PMH to import records from selected NSDL-associated collections (about 10,000). In INFOMINE, there is a total of close to 60,000 expert-created records either of internal origin from closely allied institutions or that have been created externally by sharing institutions and imported. All of these expert-driven data streams form a first tier of records in the architecture of iVia.

The second-tier collection supported by iVia consists of records that have been created automatically by crawler/classifier robots. There are also records that are of robot origin but that have been refined, augmented, and vetted by experts. This is an example of semi-automation with experts receiving machine assistance in resource discovery and metadata development. Currently, there are three crawler/classifiers (to be described below) that have created over 170,000 records. As in Google, these records, while far from MARC perfect, remain very useful and have been created relatively inexpensively. In the architecture of iVia they form a large second-tier collection that is used to support the first-tier collection of expert-built records. Complemented by the 60,000 expert-created records, INFOMINE's total collection size is around 230,000 records and growing rapidly.

Importantly, the content of iVia records ranges from just metadata to metadata augmented by selected, rich full text that has been robotically harvested from the resource itself. Judicious use of full text is of great help to user retrieval by drastically increasing the amount of material that can be searched and therefore the granularity or detail in searching that can be supported. Full text also helps correct for controlled subject vocabularies that are often too removed from common parlance and/or too general or specialized to adequately serve a wide variety of user audiences.

The collection designs discussed above have been very successful. They have been able to reflect and provide intelligent organization and access to content from many different sources and of many different types. In a world of multitudes of important collections and approaches to metadata, the iVia hybrid collection approach has been very useful for end-user access.

Support for Expert Content Builders

Just as iVia provides means for facilitating and aggregating the mutual efforts of multiple institutions, it also provides a great amount of time saving, machine assistance, and other means of expediting the work of expert collection builders. Machine assistance is provided in new resource discovery (that is, collection development), metadata generation (that is, indexing), and in a great number of smaller collection-building tasks.

Machine Assistance through Automated and Semi-Automated Resource Discovery Automated and semi-automated resource discovery (that is, collection development) is a major boost in collection building and saving the time of experts in finding relevant new resources. iVia uses several Web crawlers to scour the Web (or selected parts of it) to identify scholarly and educational resources of interest (Chakrabarti, 2003). The crawling technology can run fully automatically, but it has been built to include important roles for experts in guidance, refinement, and truing. For example, experts work with the crawlers to monitor and adjust resource acceptance weighting thresholds or the criteria by which a crawler will identify a resource as relevant. Screening for duplicates or resources already in the database is a perennial challenge. This is done through automated means as well as through experts monitoring lists of potential duplicates found through either exact or fuzzy matches of title and URL information. For irrelevant sites that keep re-occurring in crawls, iVia content-builder community blacklists are maintained that prohibit future crawler visits.

For custom, finite crawls, we have built crawlers that are fully expert guided in the sense that well-defined crawling targets are provided by experts and crawling occurs in a very directed manner. iVia's "Expert Guided Crawler with Drill Down/Drill Out" takes expert-provided individual or multiple URLs and crawls them. Experts specify the number of levels down into a site that should be crawled (most sites being organized hierarchically) as well as the distance of other sites linked to from the expert-provided site that should be pursued (for example, options are one to two jumps from the original URL). This semi-automated crawler gives the expert the ability to "mine" for new resources/links in a very precise way. A single page or site can be crawled, or a community of closely linked sites can be crawled. Likewise, we are building a focused crawler that will take a topic that is very well defined by experts and concentrate on just that topic. This is a semi-automated focused crawler that will be dependent on feedback and truing from participating experts for best results.

Just as experts interact with and improve crawler processes and accuracy, the interaction can be reversed with crawlers suggesting the most promising of sites as needing expert attention from content builders. That is, the most highly weighted sites that are automatically included in the crawler collection are flagged for expert review and refinement. Similarly, iVia database and record usage statistics are kept so that the most used or visited records of crawler origin can be flagged for expert attention, whereby the automatically created metadata present can be improved. Such a record is then moved from the second-tier, robot-created collection to the first-tier, expert-created collection. These are both important collection development tools and provide useful assists for experts.

Machine Assistance through Automated Record/Metadata Generation or Import Automated and semi-automated metadata generation provides expert content builders with a great advantage (Chakrabarti, 2003; Frank & Paynter, 2004). Collection size and depth is greatly improved through records created in these ways. Specifically, iVia's second-tier collection of records, those that have been created fully automatically, provides a great boost for the utility and value of the collection as a whole to users and greatly augments and complements expert content-building work. At the same time, the existence of automatically created records provides great assists for expert record-building activities when they are viewed as "foundation records" or records that have been partially built (from a librarian standpoint) and that can be improved upon through some expert effort. Working with these automatically created records as foundation records and improving them saves expert time compared with creating records from scratch. Foundation records can be seen as the basic "ore" that can be easily refined for more demanding or discerning uses where more rigorous (though more expensive) metadata may be the norm.

Expert content builders are also aided, as mentioned above, by iVia's ability to import and share records with other collections though OAI-PMH and standard delimited formats. This also contributes to boosting collection size, depth, and value for the end user.

Specific Machine Assistance to Experts in Record Building Numerous small machine assists are supplied by iVia to make expert record building more efficient. In the aggregate, these are crucial and save much expert time. For example, iVia supports

- Duplicate checking: prior to building an expert record, the iVia checker finds both exact and fuzzy matches within the URL and title fields for experts to review. Also identified and deleted, by checking exact lengthy character strings, are mirror sites.
- Record cloning: multiple records can be built representing closely related sites, authors, or organizations. Similarly, multiple records on the same or related subjects can be cloned and the subject and keyword indexing, among other metadata, saved and re-utilized.

- Batch editing: just as multiple records can be imported or exported in batches, their metadata can be edited and changed globally in batches. This saves much time in cases, for example, where a convention on naming a resource type has changed.
- URL Canonization: variants of URLs are canonized to proper form when this is needed.
- URL change notification: always a challenge is keeping up with changing URLs. To do this iVia has developed a "URL Checker and Pursuit" utility that flags problem URLs, notes the nature of the problem, notes potential locations indicated by forwarding messages, and (after three consecutive failures of a URL over a period of three weeks) flags the editor of the subject file with the record with the problem URL and suggests possible working URLs.
- Pull down menus of various controlled vocabularies: these would include resource types, keywords, and broad subject disciplines.
- User corrections/suggestions/new content: these are encouraged and funneled to content builders. This has been a major source for identifying possible new content and correcting errors.
- Online and point-of-need guidance: help is provided via manuals, style guides, and pop-up screens with pointers.
- Collection development assistance: this is supplied to other collections through iVia's email-based "New Resources Alert Service" and through the Whats New! index.

Under the Hood

The techniques, approaches, and algorithms that make machine assistance to experts in collection building and, more generally, iVia possible are described in more depth at the iVia site, http://infomine.ucr.edu/iVia.

A Collaboration-Inducing System

There are a number of catalysts that should stimulate increasing collaboration with iVia and its participants. The foremost is that, working together, a powerful, far-reaching, and high-quality finding tool and both internally developed and allied, externally developed collections, with proven value to researchers and students, will continue to grow and thrive. Working together, collaborators reduce redundant efforts by sharing and distributing collection development tasks and by unifying system building and support activities. Collaborators participate in a state-of-the-art system incorporating resource-saving machine assistance in numerous tasks.

Furthermore, the iVia system is in the public domain, free, and open to custom development. At the same time, iVia and the collections it provides access to can be utilized through custom interfaces and data views that meld well with the Web presence of the collaborating institution. Additionally, as one of the first library-based Web services, iVia/INFOMINE developers have a great deal of experience in meeting scholarly Internet user finding needs. Finally, the collections that populate iVia through INFOMINE are significant, well-organized, and useful. INFOMINE is among the largest librarian-built collections of its type.

Summary

iVia is a powerful and flexible, collaboration-enabling, open-source, Internet collection-building, and finding tool system. It is of use in building Internet collections of metadata and full-text data representing resources from the Web as exemplified through INFOMINE, one of the earliest and more significant of academic virtual libraries. The metadata generated includes library standard subject schema. iVia supports single or multiple subject focuses as well as both single or multiple institutional efforts. It is intended as community-ware and has proven itself to be of value in multi-institutional collaborations such as INFOMINE, NSDL, iVia, and, shortly, Data Fountains. User retrieval options are numerous for both fielded and full-text data and support both beginning and advanced searchers. iVia supports custom branding, interfaces, and data views for those accessing its collections. Numerous modes of content building are possible featuring varying levels of editorial review, styles of indexing, and divisions of labor. iVia is noteworthy because it saves resources and labor by integrating fully automated, semi-automated, and fully manual modes of record building. Resource discovery through various iVia Web crawlers and metadata generation through iVia classifiers (and other means) results in collections that require fewer resources and less expert labor to reach significant size. iVia emphasizes collaboration and empowers the librarian expert through the use of machine assistance.

References

Chakrabarti, S. (2003). *Mining the Web: Discovering knowledge from hypertext.* San Francisco: Morgan Kaufman.

Frank, E., & Paynter, G. W. (2004). Predicting Library of Congress classifications from Library of Congress subject headings. *Journal of the American Society for Information Science and Technology, 55*(3), 214–227.

Mason, J., Mitchell, S., Mooney, M., Reasoner, L., & Rodriguez, C. (2000). INFOMINE: Promising directions in virtual library development. *First Monday, 5*(6). Retrieved November 20, 2004, from http://www.firstmonday.dk/issues/issue5_6/mason/index.html.

Mitchell, S., Mooney, M., Mason, J., Paynter, G., Ruscheinski, J., & Kedzierski, A., et. al. (2003). iVia open source virtual library system. *D-Lib Magazine, 9*(1). Retrieved November 20, 2004, from http://www.dlib.org/dlib/january03/mitchell/01mitchell.html.