Collection Metadata Solutions for Digital Library Applications

Linda L. Hill and Greg Janée
Alexandria Digital Library Project, University of California, Santa Barbara, 1205 Girvetz, Santa Barbara, CA 93106. E-mail: {lhill, gjanee}@alexandria.ucsb.edu

Ron Dolin
Computer Science Department, University of California, Santa Barbara, Santa Barbara, CA 93106. E-mail: rad@cs.ucsb.edu

James Frew
Donald Bren School of Environmental Science and Management, University of California, Santa Barbara, 6715 Ellison Hall, Santa Barbara, CA 93106. E-mail: frew@bren.ucsb.edu

Mary Larsgaard
Map and Imagery Laboratory, Davidson Library, University of California, Santa Barbara, Santa Barbara, CA 93106. E-mail: mary@library.ucsb.edu

Within a digital library, *collections* may range from an ad hoc set of objects that serve a temporary purpose to established library collections intended to persist through time. The objects in these collections vary widely, from library and data center holdings to pointers to real-world objects, such as geographic places, and the various metadata schemas that describe them. The key to integrated use of such a variety of collections in a digital library is *collection metadata* that represents the inherent and contextual characteristics of a collection. The Alexandria Digital Library (ADL) Project has designed and implemented collection metadata for several purposes: in XML form, the collection metadata "registers" the collection with the user interface client; in HTML form, it is used for user documentation; eventually, it will be used to describe the collection to network search agents; and it is used for internal collection management, including mapping the object metadata attributes to the common search parameters of the system.

Digital Library Collections

Digital libraries are more heterogeneous than traditional libraries have ever been in their objects and collections, their formats, their user communities, and in the services that they support. Traditional libraries have specialized in "some aspects of providing qualitative information, an aspect determined both by their own activity defining the jurisdiction and by the social and cultural context within which they work" (Abbott, 1989, p. 217). Digital libraries expand far beyond this to encompass what has been the domain of data centers and personal or group collections and into the contents of the objects heretofore treated only as "packages"—e.g., books and journal titles. The structural model of information sources that once could be considered as distinct primary (original works), secondary (indexing and abstracting services), and tertiary (encyclopedias) domains are merged in the new digital library environment. Also, digital libraries are coming into the work environment where contribution, access, and processing occurs—where "assembly of information into usable ideas" (Abbott, 1989, p. 244) takes place. One striking difference in the movement toward digital libraries is in the concept of *collections*.

In traditional libraries, collections are firmly associated with library holdings: a collection is a set of copies of materials that a library holds, just as a museum collection consists of the objects held by the museum. When libraries access information outside of their own holdings, they are conceptually accessing other collections and services. A library's overall collection of holdings can consist of a number of special collections, such as rare books or maps, which are given special treatment. A library's catalog is the index to the library's collections and contains the metadata for the items in those collections.

For a digital library, the concept of a collection is as broad as a dictionary definition: "a group of objects;" and the objects in a digital library are not necessarily physically owned by the library. A collection can be a set of metadata pointing to distributed resources or even (as with gazetteers or directories) to the real world. Indeed, any "bag of objects" could be a collection in a digital library. Such ad hoc collections might include sets of query results saved for future reference; objects from various collections selected for their relevance to a current project; metadata selected by an information retrieval filter or agent; or a directory of individuals or organizations.

Additionally, determining which object grouping will be treated as a collection is open to many interpretations, each being highly dependent on local circumstances and digital library policies. The determining characteristics of a collection *may* include topic coverage; format; geographic coverage; temporal coverage; pertinence to a particular study or project objective; source of origin; physical location; or source of funding support; but whether a collection *is* established on the basis of any of these criteria is a choice to be made (Marshall, 1998). Sets of objects with attributes in common are not necessarily best treated as separate collections. Map series and sets of aerial photographs from a particular flight, for example, can be described through parent/child records within a larger collection—the parent record ties all the child records together as a set rather than treating the set as a separate collection.

Substantial confusion remains about what a collection is in a digital library, what its characteristics are, and what the overall model of hierarchical sets of objects and collections is. In some sense, the term *collection* is used in so many different contexts in the digital library environment that it almost requires a modifier to qualify its meaning within the current discourse. The view of the collection structure presented to the users may not be the same view held by the database managers who may need to consider pieces of a collection on separate servers or in separate databases as separate collections. On the other hand, those who acquire *collection sets* to add to a larger collection often refer to these acquired sets as collections and may still think of them as subcollections, even though they are integrated into a larger collection.

Attempting to develop a general model of digital library collections raises several questions about the fundamental nature of collections. These questions are currently the subjects of vigorous debate within the Alexandria Digital Library technical community; we mention them here to place the collection metadata in a larger context.

- *Collections vs. objects:* To what extent is a collection similar to any other object manipulated by a digital library? To the extent that they are represented by the same kinds of metadata, should collections and objects be treated identically (e.g., should a query against the library return whole collections as well as individual objects)? Or is this similarity strained in the sense that it tends to preclude the development collection-oriented metadata and capabilities?
- *Collections vs. servers:* To what extent do the semantics of *collection* mandate a 1:1 correspondence between collections and online services? For example, does every collection require its own distinct online presence? Or should the library be free to associate multiple collections with the same online interface, and, in that case, must the interface support mechanisms to distinguish between collections?
- *Subcollections:* Can collections nest? To what extent is a collection's identity bound up in its context, such that a collection would fundamentally change by virtue of being included as part of another collection? And, do nested collections require special interfaces to navigate their nesting hierarchy?
- *Volatility and quality:* Should support be given to enable users to create their own ad hoc collections easily, or should the notion of collection retain some measure of the formality it carries in the traditional library community? Can ad hoc collections and professionally managed collections coexist in the digital library environment?

Collection Metadata

In one way or another, a spectrum of collections will be a feature of digital libraries. How can such a variety of collections be modeled to support both computer processing and human use? The key is *collection metadata*. Broadly speaking, there are two classes of collection metadata:

(1) *Inherent Metadata*—that is, information that can be derived through the computer analysis of the contents of any collection, such as:
- Temporal coverage (e.g., visualization of the time periods covered or publication dates)
- Types of items and the number of each
- Formats of items and the number of each
- Example of full metadata content
- Geospatial and image collections
  - Number of thumbnail/browse images available
  - Types of geospatial footprints (e.g., points, bounding boxes)
  - Geographic coverage of the information (map visualization based on latitude and longitude coordinates of items in the collection)

2. *Contextual Metadata*—that is, information supplied by the collection provider or collection maintainer that cannot otherwise be derived from the collection's contents, such as:
- Title
- Responsible party
- Scope and purpose
- Type of collection (digital items, offline items, gazetteer, etc.)
- Date (creation or latest update)
- Update frequency
- Metadata schema(s)

- Terms and conditions of use for the collection
- Contact(s) (name, position, e-mail, etc.)
- Special behaviors (e.g., search semantics) that the collection may exhibit or require in specific operational contexts (e.g., when accessed by a particular search engine)

For ad hoc collections, the metadata will consist of the *inherent* collection attributes, plus minimal *contextual* metadata, such as the date and time of creation, the query parameters (if it is a result set from a query), and identification of the collection's creator. If a user wants to "publish" a collection for use by others, then additional collection metadata, such as title, responsible party, and scope and purpose, are required to establish the context of the collection and provide contact information for the responsible party. Established library collections intended to persist through time will have more extensive contextual descriptions, which will be added by information specialists and digital library managers.

Collection metadata can be used for:

- *Collection "registration"* with the search and retrieval and client software that will provide access to it;
- *Network discovery* by providing information to network search agents about what the collection contains;
- *User documentation*—that is, information about the collection and the digital library interface to it;
- *Management* of the collection to provide a center point where information is stored (or referenced) pertaining to the collection.

The next section relates this work to other collection description activities. This is followed by a section describing the approach of the Alexandria Digital Library (ADL) (Alexandria Digital Library, 1998c) to collection metadata. First, the ADL system architecture and collection guidelines are briefly presented to place the discussion of the ADL collection metadata approach in context. Then the format and use of the ADL collection metadata for the purposes listed above are discussed in more detail. In an appendix, an example of the collection metadata for one of the ADL collections is included. ADL focuses on georeferenced information, such as maps, remote-sensing images, aerial photographs, and texts, so you will see that the geographic identity of the information content in terms of latitude and longitude coordinates (the *geographic footprint*) is a key feature of the ADL collection metadata structure.

Related Collection Metadata Activities

A form of collection metadata exists currently in the world of online bibliographic databases: DIALOG Bluesheets and similar user guides of other bibliographic databank services are versions of collection metadata designed primarily for user documentation. Bibliographic databanks began in the late 1960s and early 1970s with ORBIT by Systems Development Corporation (SDC) and DIALOG by Lockheed Missiles and Space Corporation, with Bibliographical Retrieval Services (BRS) added in 1977 (Lilley & Trice, 1989, p. 84). They each provided powerful, centralized searching software, large storage capacity, and transaction processing that could handle many simultaneous users. They can be considered the earlier solution to the digital library challenge of today. They started by loading files from various data producers, such as ERIC from the U.S. Office of Education, Engineering Index, Chemical Abstracts, and the National Library of Medicine files, all of which were essentially machine-readable files that were the byproduct of computerized print operations. Online bibliographic services provided access to 1094 bibliographic files in 1985 and to 2465 files in 1997 (Williams, 1998). For each file offered through the major database vendors, user guides are published in a standard format to inform users about the origin of the file; the identity and contact information for the owners/creators; the scope and coverage and update frequency; the search, sort, and print options; and so forth. Behind the scenes, the database producer's original file format and content is mapped and massaged into a databank service file with indexes, print formats, and attribute labels in the style applied to other files. The result is that the searcher can search across multiple files with a common approach and deal with a common approach to print and sort options.

In the traditional library world, "special collections" (e.g., archives of historical document sets) are most often described by collection-level metadata created by special collections departments using the USMARC Bibliographic format (U.S. Library of Congress, Network Development and MARC Standards Office, 1998). These collections often have a large number of one- or two-page items (e.g., correspondence, manuscripts, and pictures), making full description at the item level unrealistic. As funds allow, these departments provide more specific access to these collections by creating *finding aids*, generally hardcopy in nature. *Finding aids* are inventories, registers, indexes, or guides that are created to provide detailed information about specific archival collections and repositories. While the *finding aids* created may vary somewhat in style, their common purpose is to provide detailed descriptions of the content and intellectual organization of the collections. They often provide contextual information about a collection's provenance and the conditions under which it may be accessed or copied; biographical or organizational histories related to the collection; a note describing the scope and content of the collection; and progressively detailed descriptions of the parts or components of the collection, together with the corresponding call numbers, container numbers, or other means for researchers to identify and request the physical entities of interest to them.

The latest initiative in this area is the Encoded Archival Description (EAD), which is supported by the Society of American Archivists and the Library of Congress. The EAD

has been standardized into a platform-independent electronic format for publication and exchange, using an SGML (Standard Generalized Markup Language) Document Type Declaration (DTD) (U.S. Library of Congress, 1998; University of California, Berkeley, 1998). It used to be necessary for researchers to visit the special collections in person to consult the hardcopy finding aids. The development of electronic EAD-based finding aids has, for the first time, allowed the online distribution of these collection descriptions.

Within the networked information retrieval environment, the need has also been felt to provide documentation at the collection level to improve the performance of search engines. Stanford University and a consortium of network search engine vendors are collaborating on a project to improve the ability of network search engines ("metasearchers") to choose the appropriate collections for searching, to execute the queries at these sources using local search engines, and to merge the results of the queries from multiple sources into one ranked list to return to the user. The Stanford Protocol for Internet Retrieval and Search, known as "STARTS," has labeled these metasearcher tasks as the *source-metadata problem*, the *query-language problem*, and *the rank-merging problem*. The result is the STARTS protocol (Gravano, 1997). In STARTS, collection (source) metadata consists of two files: one containing the inherent metadata derived directly from the collection, including a complete word index for searching, and the other containing the contextual information providing ownership, coverage, and contact information. The STARTS collection metadata schema includes information about the attributes that the search engine can query, the rules applied to the creation of the indexes to those attributes, and the types of queries that the collection can accept. In addition to this information, STARTS also collects information particular to a search engine, independent of any collection, for the purpose of data fusion. The STARTS project currently deals only with text collections. A reference implementation of the STARTS protocol is being implemented at Cornell University (Lagoze, 1998).

A *Z39.50 Profile for Access to Digital Collections* (U.S. Library of Congress, Z39.50 Maintenance Agency, 1996) was developed in 1996 for the purpose of providing the "semantics for navigating digital collections, to locate and retrieve objects of interest" (section 1.2). The profile contains, in section 2.2, a "Model of a Collection," including a discussion of why a collection differs from a database. An interesting aspect of this approach to collection and object description is that both functions are included within one profile by use of an attribute labeled "typeOfDescriptiveRecord", which designates whether the Descriptive Record pertains to a collection or to an object. This profile treats collections as *opaque* objects; that is, it describes the names, locations, and relationships of collections but not details about the content of those collections.

Each of these collection metadata developments serves its purposes within its particular application context and has informed the development of the ADL collection-metadata design. None of them is exactly what ADL needs to satisfy all of the roles of collection metadata for its georeferenced digital library system. The archival/EAD approach emphasizes contextual description at the collection level, providing item-level description only when it is practicable to do so. The STARTS protocol is specifically for text collections. Its model of inherent and contextual collection characteristics, however, has been adopted by ADL, with additional attributes but without the specificity of including the complete term list in the metadata. The user-documentation model of the online bibliographic databank services has been emulated by ADL with the addition of visualizations of geographic and temporal coverage. The unique contribution of the ADL approach is the role of the collection metadata in "registering" collections with the ADL system so that they can be accessed and in supporting all four roles described above for collection metadata with a single model. The following section describes the ADL system briefly and the structure of the ADL collection metadata applications.

Alexandria Digital Library and Its Collection Metadata

ADL is a research digital library project focussed on georeferenced/geospatial information; that is, both on the georeferenced aspects of all forms of information and on geospatial data types such as maps, aerial photographs, remote-sensing images, and data pertaining to particular geographic (Smith, 1996; Smith et al., 1996; Smith & Frew, 1995). In addition to fundamental research into digital library technologies, ADL has built prototype systems and released the third interface to its collections in the summer of 1998. This prototype has a Java-based client and a middleware layer that links the client to the underlying databases and collections. The collection metadata schema presented here is part of the prototype system and its development is in itself a prototyping activity.

Figure 1 is included to show what the ADL Java-based client looks like. Full object metadata display (the METADATA button in Figure 1) and access links (the ACCESS button) are not illustrated here. More information about the system and the interface is available through the ADL Homepage (Alexandria Digital Library, 1999b).

ADL Architecture

The ADL architecture, shown in Figure 2, is a three-level, client-server type architecture (Frew et al., 1998). At the highest level are client programs, which present the system to the user and maintain the state of the user's session with the system. (So far only one sample client has been developed, but more are anticipated.) At the lowest

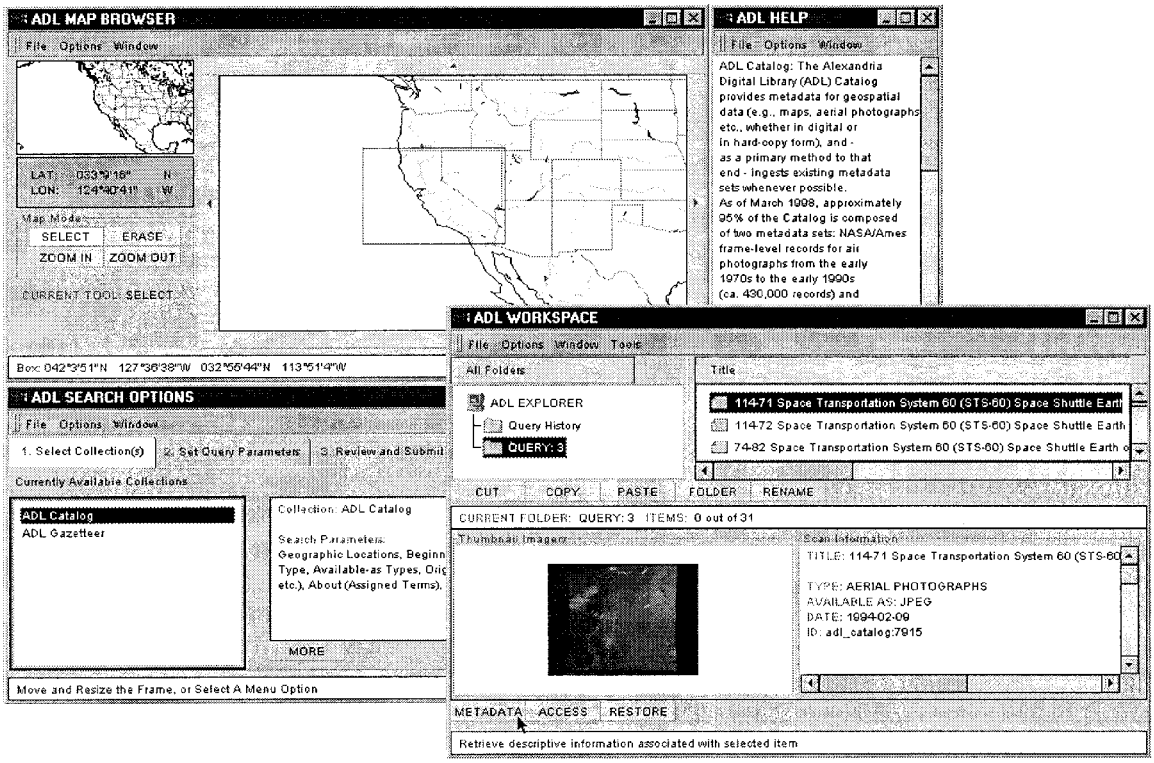

FIG. 1. Illustration of the ADL JiGi user interface.

level are collection servers, one per collection, which perform queries and retrieve metadata and holdings. Between those two levels is a middleware server, which hides collection-specific details from client programs and also handles such mundane tasks as access control and logging.

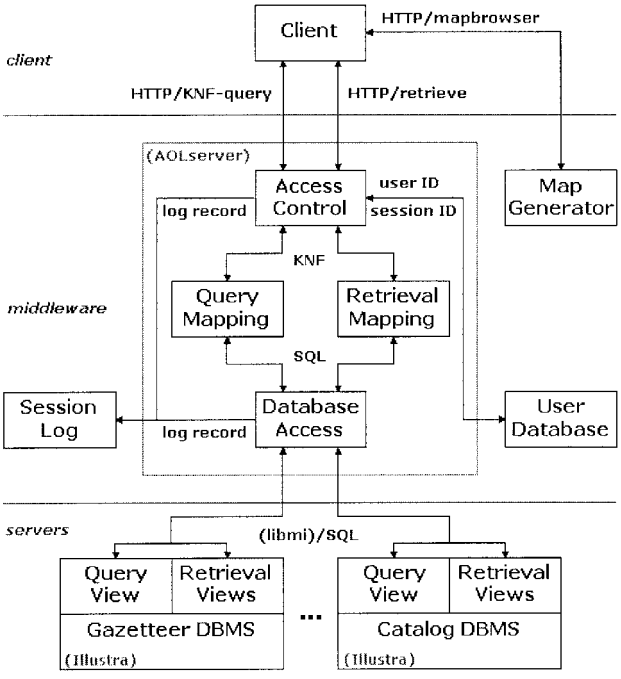

FIG. 2. Alexandria Digital Library system architecture.

The central and driving architectural features of ADL are the middleware server and the interfaces that the middleware server implements. The middleware server presents standard, collection-independent services to client programs for searching collections, retrieving collection and object metadata, and retrieving the collection objects themselves. Furthermore, with respect to searching collections, the middleware server defines standard *search buckets* ("Query Mapping" in Figure 2) that provide uniform, simplified search semantics across all collections (Frew et al., 1999). The search buckets that have been defined so far are shown in Figure 3.

These search buckets, and the operations that they support, represent a compromise between collection-specificity and uniformity. The buckets were designed to accommodate the metadata and search operations common to most collections, while providing sufficient expressive power to construct useful queries. (The ADL search bucket design is extensible to additional parameters and to *subbucket* disaggregations).

As stated earlier, the middleware server is the driving architectural feature in ADL; collection servers and client programs are subservient to the middleware. Collection servers are expected to implement the services that the middleware in turn provides to client programs, and in

Alexandria Digital Library Search Buckets

Geographic Locations - Latitude and longitude values showing the spatial location of the content of the collection objects.
Beginning and Ending Dates - Date ranges for coverage and/or publication.
Type - Categories for logical types of objects.
Format - Categories for the format in which the objects are available.
Originators - Personal and corporate names of those responsible for the creation or publication of the objects.
Topical (Assigned Terms) - Words from terms that were assigned to the objects by the catalogers from controlled vocabularies.
Topical (Freetext) - Words from all fields that will indicate what the object is about, including assigned terms.
Identifiers - Identification numbers/labels by which objects can be identified, such as ISSN, ISBN, CODEN, scene ids, and call numbers.

FIG. 3. Alexandria Digital Library search buckets.

particular, collection servers are expected to map their specific metadata schemes to the middleware search buckets for the purposes of searching and to the *retrieval mapping* for metadata display. Client programs are expected to use only the services that the middleware provides, and not to end-run the middleware and access collection servers directly.

ADL Collections

ADL has had two main collections from the beginning: the *ADL Catalog* and the *ADL Gazetteer*, which are shown in Figure 1 and illustrated in Figure 2. The architecture, however, allows ADL to accept multiple collections. These collections represent a range of collection types with very different metadata schemas for item description:

(1) Formal library collection of holdings using a metadata schema based the Federal Geographic Data Committee's (FGDC) Content Standard for Digital Geospatial Data (U.S. Federal Geographic Data Committee, 1995; U.S. Federal Geographic Data Committee, 1998) with MARC extensions; the *ADL Metadata Schema*;
(2) Gazetteer of placename references using a Gazetteer Content Standard developed by ADL (Alexandria Digital Library, 1999a);
(3) Index to bibliographic citations (GeoRef) (American Geological Institute, 1998) using a subset of USMARC (U.S. Library of Congress, Network Development and MARC Standards Office, 1998);
(4) Saved result sets from queries of ADL collections for classroom teaching purposes; and
(5) Research data collections described by metadata based on the ADL schema.

A collection is eligible to be *registered* with Alexandria Digital Library so that it is available through the ADL search and retrieval system if it meets these requirements:

(1) Each object in the collection must be associated with a *geographic footprint* (latitude and longitude coordinates of the area that the item is about).
(2) The metadata for objects in the collection are expected to contain the following attributes in addition to the geographic footprint. In most cases, all of these attributes are required, but there are some exceptions. This is the minimum set of attributes; most metadata will be much more descriptive.
   - a. Title
   - b. Date (publication and/or date of coverage)
   - c. Type (logical type of item)
   - d. Format(s) in which the item is available
   - e. Information or pointers (URLs) showing how to access the object.

(3) The collection itself is described by collection metadata (described below).
(4) For display purposes, the object metadata used in the collection must be formatted for full metadata display. That is, there has to be a full metadata report for each collection object for user viewing and printing on request ("Retrieval Mapping" in Figure 2).

The FGDC Content Standard is designed to describe digital geospatial datasets. ADL added attributes to the 1995 version to cover the characteristics of hardcopy maps also, and this became the ADL Metadata Schema used to describe objects in the ADL Catalog. The FGDC Content Standard and the ADL Metadata Schema are not designed to represent collections, except in the sense of single metadata records (parent records) that represent series such as the U.S. Geological Survey topographic map series.

ADL Collection Metadata Attributes

The ADL collection metadata schema contains elements of both the databank service user-documentation sheets and

the STARTS collection metadata and is informed by library and archive practices in describing collections. In addition, it provides for the particular search environment of ADL—namely, geospatially oriented attributes at the collection level and the particular mappings required by the ADL middleware and user-interface client. The ADL collection metadata is designed to contain all of the metadata needed by the middleware to add a collection to the ADL system, including the generation of user documentation. It is also a collection management tool in that it gathers together in one place all of the pertinent information about that collection. The portion of the Collection Metadata that describes the particular search capabilities for the collection is specific to a particular user interface and set of search programs. If, for example, a particular collection is made available through different interfaces, the collection information would remain the same but the search options would change.

The metadata structure used to describe the objects in a collection is referenced in the collection metadata. It is assumed that there will be a variety of item-level metadata schemas, each designed to most appropriately represent the type of data/information at hand. It is, of course, a gain in efficiency and user understanding if there are shared semantics and structure among various metadata schemas. However, there will always be differences among the structural components and expressions of metadata for various domains and types of objects—just as there will always be domain-specific language structures for expressing the concepts and relationships within a particular practice or topic area. A key function of collection metadata is to declare the metadata structure(s) used to describe the objects within it, with reference to complete documentation for that (those) structure(s) and to specific versions of the structures.

For each ADL collection, the mappings from the object metadata attributes to the ADL system's *search buckets*, *scan elements*, and *full metadata display* are identified through the collection metadata. These constructs support these functions:

- *Search buckets* provide common search parameters across various collections with a small number of top-level buckets (see Figure 3).
- *Scan elements* are the attributes displayed in the brief one-line display of items in result sets. They currently consist of title, date, type category, type of available format, and the short name of the collection.
- The *full metadata display* provides a labeled, and perhaps standardized, presentation of an object's complete metadata.

Collection metadata also references the controlled vocabulary lists for domain values, thesauri, subject heading lists, classification schedules, and so forth, that are used to represent the values for particular object metadata attributes. These references facilitate the use of the vocabulary tools in the user interface: If they are in digital form and available through the network, they can be used as pick lists or word navigation tools; if they are not online, users can be referred to them as a source of information about the terminology used.

The attributes of the ADL Collection Metadata are listed in the Appendix 1. Each attribute is described with three attribute characteristics:

(1) *Type* {text, text & image, text & link, integer, relation, compound}
(2) *Domain* {free text, e-mail text, phone text, List reference, numeric range, URL, map, tree, date}
(3) *Source of value* {inherent, contextual}

The *Type* characteristic defines the type of data to be entered for the attribute. ADL's attributes include the usual text and integer entries as well as entries that include images to visualize collection characteristics, entries that provide hypertext links to additional information, and entries that are tabular displays of collection data (these latter are called "relations"). A "compound" attribute is a group of attributes.

The *Domain* of an attribute can be defined as "freetext", or as a specific type of text that can be computer checked to some degree for quality control. The *Domain* can also be limited to the values from a particular "List"—in other words, values may only be selected from a controlled vocabulary. If the attribute is of the type "integer," the *Domain* can limit the values to a valid numerical range. For "Search Type," the domain list specifies whether the search operation is based on geographic footprints ("map"), text matching ("text"), hierarchical categories ("tree"), or dates ("date").

The *Source* of the attribute value is designated as either inherent or contextual. Inherent attributes are generated from the collection data through automated processes; contextual attributes are supplied by collection developers/maintainers.

Visualizations of collection characteristics are used currently for two aspects of collection description: the geographic and temporal coverage of the collection objects. Examples of these graphics are shown in Figures 4 and 5 (the HTML versions are in color (Alexandria Digital Library, 1998c)). These visualizations are created by computer programs and can be applied to subsets of the collection; for example, the geographic coverage of the various types of objects in a collection can be separately visualized.

In Figure 4, longitude degrees are displayed along the horizontal axis and latitude degrees along the vertical axis. The world is tiled into a 10-degree grid. The color for a tile (shown in shades of gray in Figure 4) reflect the number of items whose footprints overlap the tile; the lighter the shade, the more items in the collection for that area.

In Figure 5, years are shown along the horizontal axis from 1890 to 1998. The vertical axis is a logarithmic scale for number of items in the collection. For each year, a bar

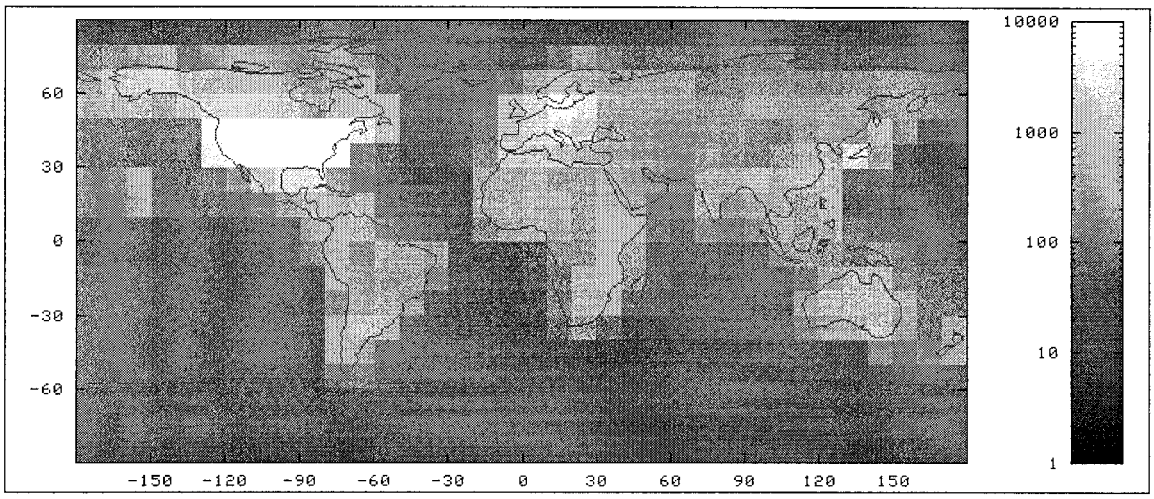

FIG. 4. Spatial Coverage of the ADL Catalog (as of April 20, 1998).

indicates the number of items whose date range overlaps the year. The line above the bars represents the number of items per decade.

An example of ADL collection metadata is included as an Appendix. This example illustrates how the attributes are used to describe a particular collection.

ADL Collection Metadata Applications

This section describes the uses of the ADL collection metadata in more detail for

- *Collection Registration*—identifying the collection to the middleware
- *Network Discovery*—using collection metadata to "advertise" collections to network search engines
- User Documentation
- *Management* of digital library collections

Collection Registration

An XML format for the ADL Collection Metadata has been implemented (Alexandria Digital Library, 1998a) that dynamically provides the middleware (and through it, the user-interface client) with the following (subset of the attributes in the Collection Metadata):

Name of collection
Short name of collection
Collection ID
Description (for client help window)
Buckets that are populated for the collection, description of the buckets and the types of values in it for this collection, and list of values for controlled-list buckets

This XML version of the collection metadata allows the ADL system to be completely flexible about adding—and removing—collections. For each collection that is available, the client presents the set of search buckets that are appropriate for the collection and the object-type categories used by the collection. The client knows how to display the object metadata for evaluation and the access information the user needs to obtain the information object. All of this is possible through the collection metadata.

Network Discovery

Although ADL manages a few collections (gazetteer, catalog, etc.), there are many tens of thousands of collec-

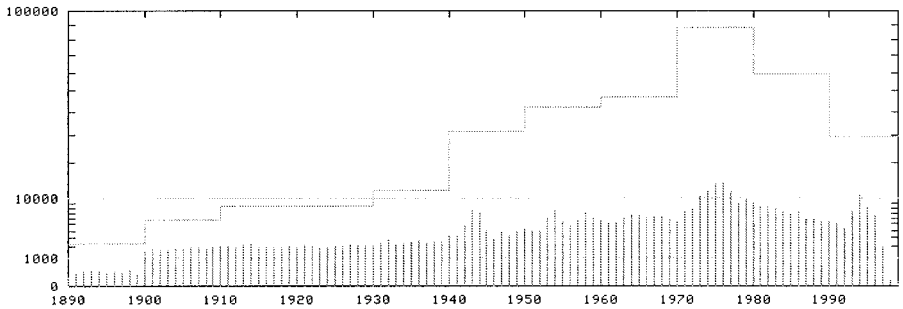

FIG. 5. Temporal Coverage of the ADL Catalog (as of April 20, 1998).

tions distributed across the Internet. The problem of locating the "best" subset of these collections to search directly is one of resource discovery (Bowman, Danzig, Manber, & Schwartz, 1994b). The role of collection metadata in this environment is to represent various dimensions of the collections so that potentially relevant resources can be identified as an intermediate step to further searching at the object metadata level.

The brute-force method of searching the Internet, such as AltaVista (AltaVista, 1999), gathers and indexes "all" documents (i.e., information objects) on the network. STARTS (Gravano, Chang, Garcia-Molina, & Paepcke, 1996), on the other hand, first gathers metadata about collections and then selects a small set of collections to search directly. STARTS is more scalable than the brute-force method because it gathers collection-wide metadata rather than every individual document.

Other resource discovery frameworks exist. A good example is Harvest (Bowman, Danzig, Hardy, Manber, & Schwartz, 1994a), which provides a mechanism for the transfer of collection metadata via "SOIF" records. Research done in connection with the Alexandria Digital Library Project has developed this approach further by building the Pharos architecture (Dolin, Agrawal, El Abbadi, & Dillon, 1997). Pharos partitions hierarchically structured collection metadata across intermediate servers and extends the collection discovery model to nontextual collection attributes, including geographic location and temporal coverage. This architecture, based on classification schemes, appears to be highly scalable.

A next step in the development of ADL collection metadata model is to build on previous work in networked resource discovery and further develop the ADL collection metadata model for that purpose.

User Documentation

When a user connects to the ADL through the Java user-interface client, the collections that are available for searching are presented (as described above). Figure 1 shows, in the lower left-hand corner, the two collections available when this screen shot was captured. The HTML versions of the collection metadata are also made available to users through a link to full user documentation for each collection (Alexandria Digital Library, 1998c). Users can consult this collection metadata display to find out more about the collection including:

- An overview of the scope and purpose and subject coverage of the collection;
- An overview of geographic and temporal coverage of the collection;
- The types of objects in the collection, how many of each there are, and their geographic and temporal coverage;
- The search buckets available for the collection and details about the contents of the buckets and the search operations that can be done for each of them;
- The ownership and contact points, terms and conditions, and update frequency of the collection; and
- The detailed mapping of the object metadata scheme or schemas to the search buckets and the display templates.

Some of this information is useful to owners of other collections who want to add new collections to the ADL system, because mappings that have been done before can be adopted or adapted to the new collection.

The example of collection metadata in Appendix 2 (which is slightly abbreviated) illustrates what the user documentation for an ADL collection provides.

Management

The creation and management of collection metadata is automated to the fullest extent possible in ADL. Inherent metadata is created by two sets of scripts and programs that are run periodically. The first set computes or extracts the required data from the ADL collections and processes it into simple, raw forms. The second set converts this raw data into any of several external forms. Contextual metadata is created as a set of text files under the control of a configuration management system. A set of scripts combines these files into any of several collection metadata versions. Currently, an HTML version with inline GIF images is created for user documentation (Alexandria Digital Library, 1998c) and an XML version is created for collection registration with the user interface according to a template (the Document Type Declaration (DTD)) (Alexandria Digital Library, 1998a).

As a tool for collection management, collection metadata also provides a central point for collection documentation. The contextual information included, such as the mapping tables from underlying metadata to the search buckets and the full metadata display, can be easily consulted when it is linked to the collection metadata. Attributes useful only for internal management, such as notes about the acquisition and processing of the metadata and data, could also be added to the schema. These local-management attributes would not be displayed in the user documentation, but would be accessible to the management staff.

Summary

It probably is impossible to arrive at a definition of the concept of *collection* in the digital library environment that everyone will agree to as an operational definition. There are just too many ways that collections can be designed or come into existence, for short or extended time periods. The decision to call some set of objects a *collection* and to treat it as such in a digital library system is purely dependent on local and contemporary circumstances and on local digital library policies. Therefore, the construct of *collection metadata* needs to be an integral part of the digital library architecture. Through collection metadata, collections can

explain themselves in terms of inherent and contextual characteristics in such a way that they can be registered with specific search and retrieval services, be informative to network search engines, be understandable to end users, and be manageable. The Alexandria Digital Library has implemented a highly automated version of collection metadata that currently supports collection registration, user documentation, and collection management for its georeferenced collections. Future integration of this collection metadata model into networked discovery systems is planned.

Appendix 1: ADL Collection Metadata
(Lists of domain values are not included here)

Title (Name by which the collection is known)
Type:text
Domain: freetext
Source of value: contextual
Short Title (Abbreviated collection name)
Type: text
Domain: freetext
Source of value: contextual
Responsible Party (The organization or person who is responsible for the collection, as it exits in ADL)
Type: text
Domain: freetext
Source of value: contextual
Scope & Purpose (Summary of the intentions for developing the collection or intended scope of the collection)
Type: text
Domain: freetext
Source of value: contextual
Subject Coverage (Descriptive statement of major subject coverage)
Type: text
Domain: freetext
Source of value: contextual
Type of Collection (Category of collection type)
Type: text
Domain: **List 1** (e.g., digital holdings, gazetteer data)
Source of value: contextual
Relationship to Other Collections (Description of the relationships to related collections)
Type: text
Domain: freetext
Source of value: contextual or captured as part of collection registration and management
Date of Collection (Date that the collection was established and/or last updated in the form of a note)
Type: text
Domain: freetext
Source of value: contextual or captured as part of collection registration and management
Update Frequency
Type: text
Domain: **List 2** (e.g., irregular, monthly)
Source of value: contextual
Total Number of Entries (the total number of entries as of a specified date)
Type: text
Domain: freetext
Source of value: inherent
Overview of Collection Characteristics
Type: compound
Spatial Coverage (Map display of geographic coverage of the collection. This display can be both overall and by type of item.)
Type: image & text
Domain: freetext
Source of value: inherent
Temporal Coverage (Date ranges and number of objects in each range. This display can be both overall and by type of item.)
Type: image & text
Domain: freetext
Source of value: inherent
Type of Items (Logical categories of types of objects and the number of each type.)
Type: relation
Thesaurus of terms or Authority List (Name of type categorization scheme)
Type: text
Domain: **List 3** (e.g., ADL Feature Type Thesaurus)
Source of value: contextual
Type Name (Category of object)
Type: text
Domain: From Thesaurus or List (e.g., aerial photograph)
Source of value: inherent
Number of Objects by Type (Number in the collection of the type)
Type: integer
Domain: ≥ 0
Source of value: inherent
Formats (Formats in which the objects are available)
Type: relation
Thesaurus of terms of Authority List (Name of format categorization scheme)
Type: text
Domain: **List 4** (e.g., MIME types)
Source of value: contextual
Format Name (Category of object)
Type: text
Domain: From Thesaurus or List (e.g., JPEG, GIF, HTML)
Source of value: inherent
Number of Objects by Format (Number in the collection of the type)
Type: integer
Domain: ≥ 0
Source of value: inherent
Types of Footprints (Number of points and bounding boxes, etc.)
Type: relation
Footprint Type
Type: text
Domain: point, bounding box, polygon, line
Source of value: inherent
Number of Footprints by Type
Type: integer
Domain: ≥ 0
Source of value: inherent
Search Options in ADL
Type: relation
Name of Searchable Attribute
Type: text
Domain: **List 5** (ADL search buckets)
Source of value: contextual (ADL)
Search Type
Type: text
Domain: map, text, tree, date
Source of value: contextual (ADL)
Types of Matching Supported
Type: text
Domain: **List 6** (e.g., Boolean, adjacency, mathematical)
Source of value: contextual (ADL)
Special Treatment (Stopwords, date handling, punctuation handling,

stemming, etc.)
Type: text
Domain: freetext
Source of value: contextual

Representation Scheme (Term sets or classification systems used, geographic location system used, or link to source of description of schemes and systems and systems used to represent the attribute)
Type: text
Domain: URL, freetext
Source of value: contextual

Metadata Schema (Name and version of metadata schema used to describe objects in the collection and link to fuller description)
Type: text & link
Domain: URL, freetext
Source of value: contextual

Metadata Mapping (Link to a file that shows the mapping from the metadata attributes to the search buckets and the full metadata display.)
Type: text & link
Domain: URL, freetext
Source of value: contextual

Sample Metadata Display (A sample ADL XML metadata record)
Type: text
Domain: freetext
Source of value: contextual (ADL)

Terms and Conditions (Terms and conditions of use of the collection)
Type: text
Domain: freetext
Source of value: contextual

Contacts (Names & positions and contact information; may be repeated)
Type: compound

Name
Type: text
Domain: freetext
Source of value: contextual

Position
Type: text
Domain: freetext
Source of value: contextual

Email
Type: text
Domain: email text
Source of value: contextual

Phone
Type: text
Domain: phone text
Source of value: contextual

Contact Note (Description of the relationship of the person to the collection)
Type: text
Domain: freetext
Source of value: contextual

Appendix 2: Collection Metadata Example

Metadata for the Alexandria Digital Library Catalog Collection

Title: Alexandria Digital Library Catalog Collection

Short title: ADL Catalog

Responsible party: Map and Imagery Laboratory, Davidson Library, University of California, Santa Barbara

Scope/Purpose: The Alexandria Digital Library (ADL) Catalog provides metadata for geospatial data (e.g., maps, aerial photographs, etc., whether in digital or in hard-copy form), and—as a primary method to that end—ingests existing metadata sets whenever possible. One major collection component is the Geodex sheet-level records for mainly topographic map series (ca. 335,000 records). Coverage is worldwide but primarily concentrated in the southern California area.

Type of collection: digital holdings, offline holdings

Relation to other collections: The ADL Catalog is completely separate from the catalog of the UCSB Davidson Library.

Date of the collection (creation or latest update): Latest update: unknown

Update frequency: irregular

Total number of entries:

332,116 items as of April 20, 1998

Overview of collection characteristics:

Spatial coverage of the ADL Catalog (see FIG. 4)

Temporal coverage of the ADL Catalog (see FIG. 5)

Types of items:

Number of each type in the collection (as of March 24, 1998):

Aerial photographs	639
Remote-sensing images	320
Maps	323,446
Photographs from space	2,832
Databases	318
Texts	3

Formats:

Number of each available-as type in the collection (as of March 24, 1998):

Online

Application: AICOV	19
Application: ARCE	333
Application: ARCGRID	17
Application: ARCVECTOR	53
Application: BIL	190
Application: BIP	22
Application: ERDAS LAN	8
Application: IPW	45
Image: DOQ	4,204
Image: GIF	4
Image: MPEG	1
Image: JPEG	3,342
Image: TIFF	1,687
Text: ASCII	5
Text: HTML	15
Text: PS	3

Offline: PAPER 20377

Search options:
Alexandria Digital Library Interface:

Name of searchable attribute	Search type	Types of matching	Special treatment	Representation Scheme
Map location (coordinates)	map	contains overlaps		latitude/longitude bounding boxes
Type	tree	exact match OR operation for multiples		See Type list above
Format	tree	exact match OR operation for multiples		See Format list above
Originator (personal and corporate authors, publishers)	text	Boolean ANY (OR) ALL (AND) PHRASE		AACR2 entry standards
Identifier (URLs, control numbers, ISBNs, ISSNs, CODENs, call numbers, etc.)	text	exact match (case sensitive)	Identifiers contain all of the original punctuation and each is treated as a "word"	None
Freetext (words from titles, abstracts, purpose statements and platform and sensor names as well as the list for Freetext-Assigned)	text	Boolean ANY (OR) ALL (AND) PHRASE		None
Freetext-Assigned (words from theme, chronological, stratum, and geographic name keywords and project names.)	text	Boolean ANY (OR) ALL (AND) PHRASE		Library of Congress Subject Headings, with some exceptions
Date: beginning date and ending date range (currently date of content only)	date	≥beginning date ≤ending date	All dates are expressed as beginning and ending date ranges and expanded (if necessary) to the following form: yyyymmdd	None

Metadata schema: ADL schema, based on the FGDC Content Standard with MARC extension for non-digital objects.
Metadata attributes mapped to searchable attributes and display template: (URL link to the mapping table)
Sample metadata entry: (omitted)

Acknowledgments

Funding from NSF, DARPA, and NASA under NSF IR94-11330 supports this work. We would also like to thank the members of the Implementation Team of the Alexandria Digital Library Project, all of whom contributed to the design and implementation of the collection model and collection metadata described in this paper. The comments of reviewers are also gratefully acknowledged.

References

Abbott, A. (1989). The system of professions: An essay on the division of expert labor. Chicago: University of Chicago Press.

Alexandria Digital Library. (1998a). ADL collection metadata document type definition (DTD). http://www.alexandria.ucsb.edu/docs/metadata/ADL-collection-metadata.dtd.

Alexandria Digital Library. (1998b). Enter ADL page. http://www.alexandria.ucsb.edu/adljigi/. Click on the "Collection Metadata" link.

Alexandria Digital Library. (1999a). ADL digital gazetteer development information. http://www.alexandria.ucsb.edu/gazetteer.

Alexandria Digital Library. (1999b). Homepage. http://www.alexandria.ucsb.edu.

AltaVista. (1999). http://altavista.com/av/content/about.htm.

American Geological Institute. (1998). GeoRef. A description of this file can be found at http://library.dialog.com/bluesheets/html/bl0089.html.

Bowman, C.M., Danzig, P.B., Hardy, D.R., Manber, U., & Schwartz, M.F. (1994a). The Harvest information discovery and access system. Proceedings of the Second International World-Wide Web Conference (http://harvest.cs.colorado.edu, pp. 763–771). Chicago, IL.

Bowman, C.M., Danzig, P.B., Manber, U., & Schwartz, M.E. (1994b). Scalable Internet resource discovery: Research problems and approaches. CACM, 37(8), 98–107.

Dolin, R., Agrawal, D., El Abbadi, A., & Dillon, L. (1997). Pharos: A scalable distributed architecture for locating heterogeneous information sources. Proceedings of the 6th International Conference on Information and Knowledge Management (CIKM '97). Las Vegas, NV.

Frew, J., Freeston, M., Hill, L., Janee, G., Larsgaard, M. & Zheng, Q. (1999). Generic query metadata for geospatial digital libraries. Proceed-

ings of the Third IEEE Meta-Data Conference (Meta-Data '99), April 6–7, 1999, Bethesda, MD, sponsored by IEEE, NOAA, Raytheon ITSS Corp., and NIMA. http://computer.org/conferen/proceed/meta/1999/papers/55/jfrew.htm.

Frew, J., Freeston, M., Freitas, N., Hill, L., Janee, G., Lovette, K., Nideffer, R., Smith, T., & Zheng, Q. (1998). The Alexandria Digital Library Architecture. In C. Nikolaou & C. Stephanidis (Eds.), Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries (ECDL '98), Heraklion, Crete, Greece, Sept. 1998 (pp. 61–73). Berlin: Springer-Verlag, http://www.springer.de/comp/lncs/index.html.

Gravano, L., Chang, K., Garcia-Molina, H., Lagoze, C., Paepcke, A. (1997). STARTS : Stanford protocol proposal for Internet retrieval and search. Web site http://www-db.stanford.edu/~gravano/starts.html; Stanford University, Digital Library Project, January 19, 1997.

Gravano, L., Chang, K., Garcia-Molina, H., & Paepcke, A. (1996). STARTS: Stanford protocol proposal for Internet retrieval and search. Working Paper SIDL-WP-1996-0043: Computer Science Department, Stanford University, July 1996.

Lagoze, C. (1998). STARTS: Stanford protocol proposal for Internet search and retrieval; reference implementation. http://www2.cs.cornell.edu/lagoze/starts/starts_reference.htm.

Lilley, D.B., & Trice, R.W. (1989). A history of information science, 1945–1985. San Diego, CA: Academic Press.

Marshall, C.C. (1998). Making metadata: A study of metadata creation for a mixed physical-digital collection. In I. Witten, R. Akscyn, & I. Frank M. Shipman (Eds.), Proceedings of the Third ACM Conference on Digital Libraries (DL '98), June 23–26, 1998, Pittsburgh, PA (pp. 162–171). New York: ACM.

Smith, T.R. (1996). A brief update on the Alexandria Digital Library Project: Constructing a digital library for geographically referenced materials. D-Lib Magazine (March 1996). http://www.dlib.org/dlib/july96/new/07smith.html.

Smith, T.R., Andresen, D., Carver, L., Dolin, R., Fischer, C., Frew, J., Goodchild, M., Ibarra, O., Kemp, R. B., Kothuri, R., Larsgaard, M., Manjunath, B.S., Nebert, D., Simpson, J., Wells, A., Yang, T., & Zheng, Q. (1996). A digital library for geographically referenced materials. Computer (IEEE), 29(5), 54–60.

Smith, T.R., & Frew, J. (1995). Alexandria Digital Library. Communications of the ACM, 38(4), 61–62.

U.S. Federal Geographic Data Committee. (1995). Content standard for digital geospatial metadata workbook. Washington, DC: FGDC, March 24, 1995.

U.S. Federal Geographic Data Committee. (1998). Content standard for digital geospatial metadata. http://fgdc.er.usgs.gov/Metadata/Cont-Stan.html.

U.S. Library of Congress. (1998). Encoded archival description (EAD) DTD. http://lcweb.loc.gov/ead/.

U.S. Library of Congress, Network Development and MARC Standards Office. (1998). MARC standards. http://lcweb.loc.gov/marc/.

U.S. Library of Congress, Z39.50 Maintenance Agency. (1996). Z39.50 Profile for access to digital collections. Final draft for review, May 3, 1996. http://lcweb.loc.gov/z3950/agency/profiles/collections.html.

University of California Berkeley. (1998). EAD @ UC Berkeley. http://sunsite.berkeley.edu/ead/main.html.

Williams, M.E. (1998). The state of databases today. Gale directory of databases (pp. xvii–xxix). Detroit, MI: Gale Research.