

## **New Approaches for Mainstreaming Metadata In Digital Library Project Development and Management**

Jonathan Corson-Rikert

### **Introduction**

The common characterization of metadata as “data about data” provides a convenient shorthand definition but fails to convey the central role metadata plays in accessing and managing information resources today. The phrase “data about data” correctly implies the descriptive function of metadata but provides no indication of the many other roles metadata serves in the discovery, management, preservation, and even the presentation of digital content. As digital resources become more diverse and complex, and as the challenges of preserving content and maintaining access over time become more evident, simple notions of metadata no longer reflect the importance of metadata in planning, building and managing digital information resources.

Efforts such as the Dublin Core Metadata Initiative have increased awareness of the need for common metadata standards to enable discovery and comparison of content wherever it may exist online or in physical form.<sup>1</sup> At a minimum, digital collections created today can be expected to have metadata describing summary attributes of the entire collection and more detailed individual characteristics of every item within it. The set of metadata elements for a digital collection is typically based on the anticipated uses of the collection and in some cases, the anticipated needs of the

<sup>1</sup> Dublin Core Metadata Initiative, <http://www.dublincore.org/> (8 Mar. 2008).

digital system or institution, defining *a priori* the ways in which the information in the collection can be searched, retrieved, and accessed. Because creating metadata is typically time-consuming and thus costly, the highest return for a limited investment can be achieved by selecting elements that can be uniformly populated across an entire collection, ideally via automated or semi-automated processes. Metadata, in many cases, has often been created only as a means to make a collection discoverable online, and often only during the final phases of a project.

While this baseline form of descriptive metadata has been widely adopted and is widely understood, most collections fail to reach full potential for discovery and use under this simplistic model for metadata. First, many collections have important characteristics not adequately captured through standard Dublin Core metadata elements or even application profiles allowing additional qualifiers. And although many separate, well-developed standards such as MIX for still images address specific content types, collections often have additional characteristics that may not be easily represented via metadata standards oriented primarily toward documenting commonality of content rather than those attributes or combinations of attributes of a collection that make it unique.<sup>2</sup> Furthermore, collections are rarely so uniform that any single set of metadata elements will suffice, and a collection may benefit from interfaces to more than one metadata standard. For example, using multiple standards to allow discovery through a library's online public access catalog using MARC and to

<sup>2</sup> NISO Metadata for Images in XML Schema, <http://www.loc.gov/standards/mix/> (8 Mar. 2008).

harvest via a protocol such as the Open Archives Initiative (OAI) that supports simple Dublin Core (DC) is appropriate and perhaps necessary for most digital collections.<sup>3</sup>

As defined in the National Information Standards Organization publication, *Understanding Metadata*, “Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource.”<sup>4</sup> Emerging digital library standards such as the Metadata Encoding and Transmission Standard<sup>5</sup> (METS) and the Reference Model for an Open Archival Information System<sup>6</sup> (OAIS) reflect this broader role for metadata by documenting provenance as well as structural, technical, and administrative characteristics of a digital collection or repository as core components along with traditional descriptive metadata. These metadata standards are all managed in the context of a digital library “object,” which may also include software to support specialized display of items in the collection. For the application of such standards, metadata cannot be an afterthought addressed in the closing phases of digital library creation, but must be incorporated into the design from the earliest phases and may in fact govern all access to and administration of a project.

This chapter will describe two Cornell University Library projects developed in Mann Library; each illustrates a more central role for metadata in digital library development while taking very different technical approaches. The Cornell University Geospatial Information Repository (CUGIR) <<http://cugir.mannlib.cornell.edu>>

<sup>3</sup> Open Archives Initiative (OAI), <http://www.openarchives.org/> (8 Mar. 2008).

<sup>4</sup> National Information Standards Organization (NISO), “Understanding Metadata,” <http://www.niso.org/standards/resources/UnderstandingMetadata.pdf> (8 Mar. 2008).

<sup>5</sup> The Library of Congress, Metadata Encoding & Transmission Standard [METS] Official Web Site, <http://www.loc.gov/standards/mets/> (8 Mar. 2008).

<sup>6</sup> Consultative Committee for Space Data Systems, “Reference Model for an Open Archival Information System (OAIS),” <http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf> (8 Mar. 2008).

provides free public access to over 7,000 geospatial data files for New York State. Initiated in 1998 as a file-based data repository documented by Federal Geographic Data Committee (FGDC) metadata records, CUGIR has evolved through successive stages to require a more flexible and powerful metadata model.<sup>7</sup> The VIVO Virtual Life Sciences Library <<http://vivo.library.cornell.edu>> is a digital library project with very different characteristics, serving not as a repository of documents, images, or data resources, but as a rich index that cross-references the people, departments, laboratories, activities, equipment, publications, and events that collectively comprise the Cornell University New Life Science Initiative. As an index rather than a repository, VIVO is *entirely* a metadata resource, but as metadata it does not describe common attributes of homogeneous items, but rather the relationships among a diverse, open-ended set of “entities” that may represent abstract concepts, scientific databases, events, places, people, and institutions as well as more traditional library content such as journal articles, monographs, and images.

Despite their contrasting goals and structure, CUGIR and VIVO exhibit certain commonalities in how they address the assembly, indexing, storage, discovery, and delivery of information. Metadata is central to the success of both projects, and the requirements for metadata management have driven much of the workflow while contributing significantly to the overall value of the resulting online resources.

<sup>7</sup> Federal Geographic Data Committee (FGDC), <http://www.fgdc.gov> (8 Mar. 2008).

## **The Evolution of CUGIR Metadata**

The CUGIR repository was started in 1998 in response to a burgeoning interest in spatial data display and analysis at Cornell and across New York State, an interest too often frustrated by difficulties in finding data in consistent formats, a lack of sufficient documentation to evaluate appropriateness for any intended use, and policy restrictions on access to data. CUGIR was unusual among early Web-based spatial data repositories in providing free, unrestricted access to data, and the librarians developing CUGIR placed a high priority on providing FGDC-compliant metadata records for all available datasets. The FGDC had in fact helped to establish CUGIR through a program of grants to establish state and regional clearinghouses for geographic data and metadata, and CUGIR still maintains a Z39.50 index of selected metadata fields using an Isite indexing profile created by the Clearinghouse For Networked Information Discovery and Retrieval and modified by the FGDC to support distributed searching from the FGDC central node as well as local searching at each participating repository.<sup>8</sup>

The first CUGIR implementation, illustrated in figure 1, followed a one-to-one correspondence model between the metadata records and the data itself, for those datasets that were published for each of the sixty-two counties in New York—each data theme represented in the collection was distributed by individual county, with an accompanying metadata record including county-specific place keywords and localized coordinate bounding boxes. In addition, as recommended by the FGDC, four

<sup>8</sup> Center for Networked Information Discovery and Retrieval (CNIDR) <http://www.cnidr.org/> (8 Mar. 2008).

individual metadata records differing only in syntax (HTML, SGML, XML, and text) were created for each dataset. As the collection grew in size and complexity—by including, for example, data distributed for each of the 962 New York 1:24,000-scale United States Geological Survey quadrangles,—the task of maintaining four metadata records for every individual dataset became unworkable, prompting the abandonment of the strict one-to-one correspondence model. The current CUGIR model maintains “core” metadata records for each data theme in each of the four FGDC-recommended formats, but does not replicate the records with appropriate place keyword and coordinate bounding box information for each individual dataset within a geographic series, which may include counties, quadrangle sheets, watersheds, or any other geographic units for which thematic data are published.

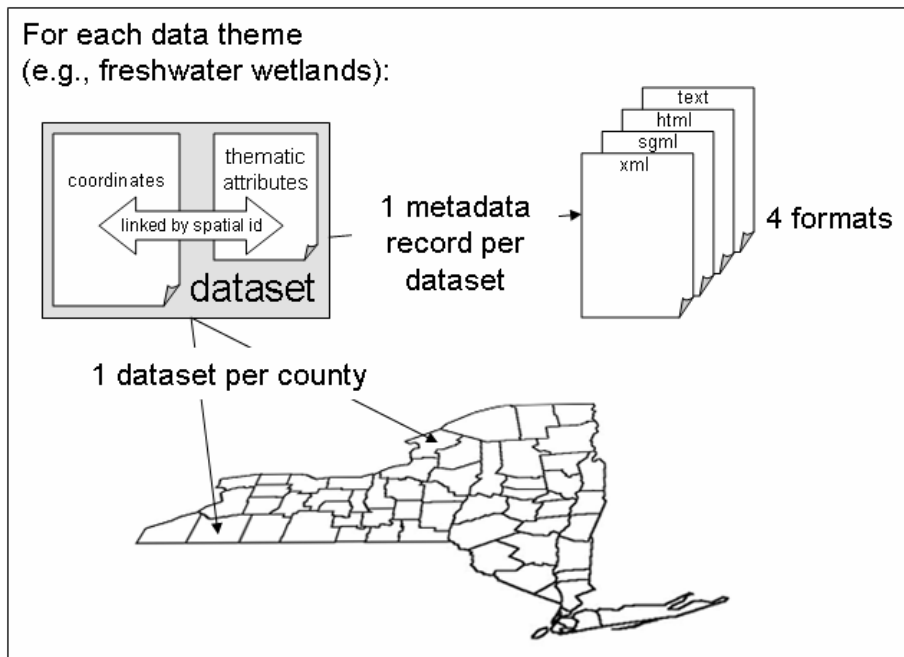


Figure 4.1. The 1:1 correspondence model linking metadata to spatial datasets

In 2001, Elaine Westbrooks, the CUGIR metadata librarian, and Adam Chandler, the Information Technology Librarian at Cornell's Olin Library, proposed and received internal library funding for a project to increase the visibility of the extensive collection of geographic metadata and accompanying data in CUGIR. Working with a student programmer, they converted the SGML-format FGDC metadata records into MARC format for inclusion in the Cornell and OCLC WorldCAT catalogs. Following a simplified version of the SODA (Smart Object, Dumb Archive) model proposed by Assistant Professor Michael Nelson, of the Old Dominion University Digital Library Research Group, they also extracted a set of Dublin Core elements for each geospatial metadata record in preparation for harvesting this subset of the FGDC metadata into Open Archives Initiative repositories.<sup>9</sup> The DC metadata elements for each dataset were assigned a common unique identifier and a persistent Uniform Resource Locator (URL) known as the "bucket." The DC metadata is displayed online as an abbreviated alternative to the lengthy FGDC metadata record, pointing directly to the HTML and XML versions of the FGDC metadata records, the dataset, and to alternative data formats such as shapefile or ArcExport.

The best feature of the bucket model is that it removes the most volatile elements from metadata records—the names of the datasets and URLs for finding them—by leaving only the persistent bucket URL and a unique identifier for the dataset in the published or harvested metadata records. If we change the server, the

<sup>9</sup> Michael L. Nelson, Kurt Maly, "Buckets: Smart Objects for Digital Libraries," *Communications of the ACM* 44, no. 5 (2001): 60-62.

directory, or the name of a dataset then we only need to make changes to the bucket database; all of the FGDC and MARC records that are dispersed throughout CUGIR, the CUL OPAC, or OCLC FirstSearch remain unchanged. The bucket also provides flexibility for adding new services such as the online map preview now available for many CUGIR data themes, and makes these services immediately accessible not just from the CUGIR website but from the remote metadata records. A typical CUGIR bucket display is captured in Figure 4.2., showing both the dynamically-generated links to the dataset and metadata and the DC elements available for harvesting.


title	<b>Cayuga County Agricultural Districts</b>
data format	Arc Export
HTML metadata	<a href="http://cugir.mannlib.cornell.edu/Isite/CUGIR_METADATA/011/011aga.html">http://cugir.mannlib.cornell.edu/Isite/CUGIR_METADATA/011/011aga.html</a>
XML metadata	<a href="http://cugir.mannlib.cornell.edu/Isite/CUGIR_METADATA/011/011aga.xml">http://cugir.mannlib.cornell.edu/Isite/CUGIR_METADATA/011/011aga.xml</a>
MARC record	<a href="http://cugir2.mannlib.cornell.edu/userDir/bw47-cornell.edu/7.dat">http://cugir2.mannlib.cornell.edu/userDir/bw47-cornell.edu/7.dat</a>
data link	<a href="http://NY011ag05a.zip">NY011ag05a.zip</a>
map preview	 <a href="http://cugir2.mannlib.cornell.edu/buckets/Display.jsp?id=7&amp;action=map">http://cugir2.mannlib.cornell.edu/buckets/Display.jsp?id=7&amp;action=map</a>
<b>elements:</b>	
rights	Rights Access: 1) Please acknowledge New York State Agricultural Districts Mapping Program as the source of the data. 2) Recognize that the boundaries are not precise because they are a general representation of tax parcel boundaries, created at a scale of 1:24,000. The digital version is not a legal substitute for actual tax parcel information. 3) Districts are reviewed for renewal every eight years, from the date of formation, found in the field "Created" in the attribute table. It is important to check the certification date, found in the field "Certified" of the attribute table. It indicates the currentness of this version of the digital data. The district data may have been outdated since the publication of this digital version. 4) This data is not to be resold in any form.
rights	Access Constraints: none.
relation	Mode of Access: World Wide Web.
relation	System Requirements: Some files require desktop Geographic information Systems (GIS) software such as MAPInfo, ARC/Info, ArcView, or Adobe Acrobat Reader, for storing, modifying, querying, analyzing, and displaying various forms of geospatial data on Windows, MAC or UNIX platforms. Additionally, some files require desktop extraction utilities such as Winzip to handle compressed or archived files.
language	eng
identifier	<a href="http://cugir2.mannlib.cornell.edu/buckets/Display.jsp?id=7">http://cugir2.mannlib.cornell.edu/buckets/Display.jsp?id=7</a>
description	Title from metadata.
description	These files are geographic attribute data for New York State (NYS) Agricultural District Boundaries for each county. They are ArcInfo (version 7.2.1) export files, in polygon format, derived from the scanned version of New York State (NYS) Agricultural District boundaries. Original maps were individual districts drafted at a scale of 1:24,000. the ArcInfo files are any districts within a NYS county boundary. The data depict private land areas placed in a district under the protection of NYS Agricultural District Law.
coverage	a W0764514 W0761557 N0434228 N0423708
coverage	n-us-ny
coverage	Scale 1:24,000 Universal Transverse Mercator W076 45'14"--W076 15'57"N043 42'28"--N042 37'08"
publisher	Cornell Insitute for Resource Information Systems,
publisher	Ithaca, NY :

Figure 4.2. CUGIR bucket #7 – Agricultural Districts Record for Cayuga County



The database tables that were developed to store DC elements and to manage metadata conversion became the core of a more complete relational database management system (RDBMS) developed to drive CUGIR. The database provides simpler searching of the collection and additional functionality, including a local copy of the USGS Geographic Names Information System Gazetteer<sup>10</sup> and Internet-based mapping.<sup>11</sup> In addition to tables identifying the unique geographic units (e.g., county, quadrangle, watershed), data formats (e.g., Arc Export, Shapefile, CAD), and thematic content (e.g. soils, freshwater wetlands, agriculture districts) of each dataset and its corresponding online persistent bucket, new tables were added to store information about data partners, subject-based groupings of data, and the aggregate series of geographic units (e.g., counties, quadrangle sheets, watersheds) by which datasets are published. These tables reflect the primary breakdowns by which we anticipated users would want to find and access CUGIR datasets.

The conversion to a relational database provided more control over the CUGIR interface. The use of “buckets”, as a persistent URL at which to gather the most pertinent information the user needs to know about a dataset, has improved access to metadata and data while removing many of the maintenance problems associated with dataset name and URL changes. However, neither effort solved the administrative challenge of storing and maintaining thousands of individual FGDC metadata files. Further impetus to store information within some form of management system has come from the desire to provide similar maintenance advantages for all metadata

<sup>10</sup> United States Geological Survey (USGS), Geographic Names Information System (GNIS), <http://geonames.usgs.gov/domestic/> (8 Mar. 2008).

<sup>11</sup> Jaime Martindale, “Cornell University Library Serves GIS Resources on the Web,” *ArcNews* 25, no. 2 (2004): 21.

elements and to provide better integration of metadata with the CUGIR website. Users have requested the ability to query by multiple dates, including dates when the data were added to CUGIR, and by less- commonly referenced metadata elements such as map projection and horizontal datum. While each new requested element can be extracted from stored metadata records and made available via the website database on a piecemeal basis, the more general need to consolidate repetitive text and streamline updates across the entire CUGIR metadata collection has argued for a more comprehensive solution.

The CUGIR strategy for migration away from separate FGDC metadata records has been heavily influenced by the approach taken by the FGDC in connection with the federal GeoSpatial One-Stop (GOS) Initiative.<sup>12</sup> Geospatial One-Stop is an e-government initiative sponsored by the Federal Office of Management and Budget to make it easier, faster, and less expensive for all levels of government and the public to access geospatial information. The GOS Portal is one component of the GOS Initiative <<http://www.geodata.gov/>> that allows participants to search and retrieve geospatial data, make maps, or publish data.

Until recently, the FGDC had relied on the federated network of Z39.50 servers of the National Geospatial Data Clearinghouse <<http://clearinghouse3.fgdc.gov/>> to permit users to search for distributed geospatial data via their central website. The frequent downtime of individual servers had affected both system performance and the consistency of search results. With the GOS initiative, the FGDC has now established a

<sup>12</sup> National Spatial Data Infrastructure (NSDI), GeoSpatial One-Stop, <http://gos2.geodata.gov/wps/portal/gos> (8 Mar. 2008).

central metadata repository supporting a range of options for individual repositories to submit metadata to this central registry, including interactive forms entry, harvesting from existing Z39.50 servers, OAI harvesting, or uploading from a directory of XML files.

Prior to April 2005, SGML versions of the metadata records had to be created and maintained in order to be indexed using Isite. Now, participation in the GOS metadata repository has allowed CUGIR to designate the XML format of FGDC records as the definitive format. The use of a Java XML parser allows the CUGIR team to easily:

1. Make global changes to the metadata such as correcting systematic errors or adding the new ISO topic keywords proposed by GOS<sup>13</sup>;
2. Update individual metadata elements such as the data provider's contact telephone; and
3. Extract more metadata elements into the CUGIR relational database to support additional query options.

After any necessary processing, a new XML record is written out to a directory that will be harvested by the central FGDC repository. This eliminates CUGIR dependency on the legacy SGML, HTML, and text metadata files, which will not be updated but instead will be generated from the revised XML files using XSL style sheets.

<sup>13</sup> Federal Geographic Data Committee (FGDC), "FGDC/ISO Metadata Standard Harmonization," <http://www.fgdc.gov/metadata/us-national-profile-iso19115/archive> (8 Mar. 2008).

This interim model for managing CUGIR metadata will also allow improved presentation of CUGIR metadata. Users are often intimidated by FGDC metadata, which can be lengthy, complex, and confusing for readers. The use of multiple XSL stylesheets will allow the presentation of an initial “lite” version of the most basic metadata—in a format less intimidating for users—along with options to display more detail in stages up to the full metadata record as recommended by the FGDC.<sup>14</sup>

The common one-to-one correspondence model between data and metadata files in repositories such as CUGIR may also prove limiting in the face of new modes for distributing the data. Geographic datasets have traditionally been distributed by relatively small geographic areas such as individual municipalities, counties, or USGS quadrangle sheets in order to limit download times and minimize user storage requirements. With commercial GIS software and database solutions now incorporating more efficient models for spatial query and data retrieval, repositories are beginning to support requests for downloading data by custom geographic areas – a popular feature for users whose area of interest frequently spans more than one county, USGS quadrangle sheet, or watershed. An exemplar of this customization is the Pennsylvania Spatial Data Access <<http://www.pasda.psu.edu/>>, which currently allows users to download data from the 2000 Census for any arbitrary combination of census tracts, accompanied by a custom selection of demographic attribute values—effectively offering millions of combinations of spatial location and demographic attribute information from a single statewide spatial and statistical data archive.<sup>15</sup>

<sup>14</sup> Federal Geographic Data Committee, “Metadata Presentation via XML and XSL” (updated version of the FGDC page no longer online), <http://www.dot.ca.gov/hq/tsip/gis/datalibrary/metaxml.html> (15 Mar. 2008).

<sup>15</sup> Pennsylvania Spatial Data Access (PASDA), <http://www.pasda.psu.edu/> (8 Mar. 2008).

Customizing data on demand also creates opportunities for adding value to geospatial metadata through more precise specification of attribute information, documenting actual bounding coordinates, projection, file format, and coordinate transformations and including appropriately filtered place and theme keywords based on a user's selected geographic and thematic areas of interest. However, it will be an ambitious task to continue beyond the CUGIR interim model described above, and extend the range of metadata stored in a database to encompass a more complete element set and eliminate the redundant specification of elements common across multiple datasets. Certain elements may need to vary independently between records, even if the current contents are identical; cross-dependencies among elements may not yet be understood well enough. Commercial geospatial systems, such as ESRI's ArcCatalog or Intergraph's SMM, appear to offer attractive out-of-the-box solutions for managing metadata and accessing datasets via metadata records, but may not be able to provide the flexibility of the CUGIR delivery system, the persistence of the CUGIR buckets, or the support for the multiple output metadata formats (MARC and DC as well as FGDC). The CUGIR team plans to evaluate off-the-shelf metadata solutions, XML databases, relational databases that can store blocks of XML, and more traditional relational database approaches as part of an overall strategy to carry the repository forward and continue to meet our users' needs for well-organized, easily accessible geospatial data.

In summary, the management of metadata is tightly bound with the entire CUGIR project, providing operational access to the information that is needed to drive new features, and to ensure that users receive optimal documentation about the data they download. Metadata elements that were once stored and searched independently

of the repository data have become key components that are integral to all interaction with the repository. The scope of metadata includes administrative, structural, preservation, and descriptive information that is central to the organization, presentation, and remote discovery of CUGIR data via library catalogs and national metadata repositories.

### **The Virtual Life Sciences Library**

VIVO, the Virtual Life Sciences Library <<http://vivo.library.cornell.edu/>>, serves as a curated index to life sciences research, transcending campus, college, and department boundaries to provide an integrated view of the life sciences at Cornell. VIVO is the creation of the Life Sciences Working Group in Cornell University Library, charged in early 2003 to develop an integrated Web presence for library resources and services relevant to life sciences, as well as to address additional goals including improved instruction opportunities for librarians and for patrons. Because Cornell University is geographically distributed, as well as academically diverse, the committee recognized the need to transcend individual services and staff expertise in ten unit libraries to create a sense of “our library” within the life sciences community. Since many of the relevant resources were already online, it seemed more appropriate to aggregate and index them, rather than duplicate existing content. Because the concept of a “resource” can be quite varied in type and scope, from individual electronic databases to full semester courses and faculty research profiles, the group looked for a solution that would also be open-ended.

Further motivation for VIVO came from a sense that students were finding the proliferation of library-based search tools confusing, and that the Google search engine

was in fact the most comfortable model for searching, if not always the best for finding or organizing results.<sup>16</sup> Neil McLean, Pro Vice-Chancellor, E-Learning and Information Services Macquarie University, Sydney, Australia and Clifford Lynch, Director of the Coalition for Networked Information (CNI), published a white paper in 2003 that describes the problem as follows:

There is growing acceptance that simply making resources available on the network without an additional layer of services may not be very effective. There are some clear reasons for this, arising from the characteristics of the current generation of network resources. In general, many of these characteristics flow from the fact that resources are made available at interfaces with very low levels of interconnectedness between them. This in turn puts the burden of interconnection back on the user, and it means that in many cases the potential value of interconnection is not realized.<sup>17</sup>

The Life Sciences Working Group set out to craft an online information service that offers the simplicity of Google, increases the library's web presence, and most importantly, highlights the interconnections among all of the stakeholders in a vibrant academic research community.

### **VIVO Antecedents**

Early designs for VIVO drew most directly on the ABC Ontology<sup>18</sup>, a framework developed by the Harmony Project<sup>19</sup> as a model for metadata interoperability among disparate digital library and museum collections with strong temporal or spatial components. The ABC Ontology itself is derived, in part, from concepts articulated in

<sup>16</sup> Google, <http://www.google.com> (8 Mar. 2008).

<sup>17</sup> Neil McLean and Clifford Lynch (2003), "Interoperability between Information and Learning Environments – Bridging the Gap," [http://www.imslobal.org/DLims\\_white\\_paper\\_publicdraft\\_1.pdf](http://www.imslobal.org/DLims_white_paper_publicdraft_1.pdf) (8 Mar. 2008).

<sup>18</sup> Carl Lagoze and Jane Hunter, "The ABC Ontology and Model," [http://metadata.net/harmony/JODI\\_Final.pdf](http://metadata.net/harmony/JODI_Final.pdf) (8 Mar. 2008).

<sup>19</sup> Harmony Project [home page], <http://www.ilrt.bris.ac.uk/discovery/harmony/> (8 Mar. 2008).

the 1997 IFLA Functional Requirements for Bibliographic Records (FRBR) report – best known for elucidating the concepts of an intellectual *work*, the *expressions* of that work in written, musical, or artistic form; the *manifestations* of each expression through editing, translation, or alternative media formats; and finally the individual physical books and other *items* that form a library collection.<sup>20</sup> The ABC Ontology incorporates many elements of the FRBR work/expression/manifestation/item model while adding concepts and explicit relationships to encode the actions that produce each successive state, along with other events affecting the ownership, location, condition, or annotation of a work in any of its forms.

We first implemented an abridged version of the ABC model with Java servlets and Java Server Pages, using a MySQL database persistence layer developed to support a number of Web-based projects at Mann Library, including CUGIR. To allow ongoing flexibility for VIVO, a single database table is used to store all of the resources or assets to be indexed, called “entities.” Each entity has only two required core attributes, a name and an assigned type. Optional attributes include a URL (the link to the “real” resource, at least as it is represented on the Web), a short description, a long description, a thumbnail image, a citation, and date fields allowing support for events and news releases. When additional attributes are needed, they are constructed not as individual text or numeric values but as relationship options from one type of entity to either the same or different type of entity, whether abstract or physical—e.g., a faculty member may have a relationship as participant in a research area, and a seminar may

<sup>20</sup> International Federation of Library Associations and Institutions, *Functional Requirements for Bibliographic Records* (FRBR), <http://www.ifla.org/VII/s13/frbr/frbr.pdf> (8 Mar. 2008).



have a relationship to a speaker or to the room where it will be held. Some of these optional relationships are used rather infrequently; faculty members, for example, have a very diverse set of connections throughout an academic community.

The screenshot shows the VIVO virtual life sciences library interface. At the top, there is a search bar with the word "fossil" entered and a "Go" button. To the right of the search bar is the Cornell University Library logo. Below the search bar is a navigation menu with links for Home, Education & Training, People, Research Tools, Facilities, Jobs, Index, About, and What's New. The main content area displays the profile of Amy R. McCune, a Cornell faculty member. It includes a photo of her, her name, and a list of related entities such as Ecology and Evolutionary Biology (EEB), Zoology, and the Cornell University Museum of Vertebrates (CUMV). There is also a section for recent articles and a list of keywords.

VIVO  
virtual life sciences library

Search fossil Go  
Advanced Search | Search Tips

Cornell University  
Library

Home Education & Training People Research Tools Facilities Jobs Index About What's New

VIVO > You last searched for: *fossil* as a whole word

McCune, Amy R. | Cornell faculty member | *Ecology & Evolutionary Biology profile*

faculty member in:

- Ecology and Evolutionary Biology (EEB) | Cornell department | *EEB web page* | *Ecology and Evolutionary Biology courses*

member of graduate field:

- Ecology and Evolutionary Biology | graduate field | *field web page* | *The Field of Ecology and Evolutionary Biology web page*
- Zoology | graduate field | *field web page* | *Graduate Field of Zoology web page*

affiliated with research unit:

- Cornell University Museum of Vertebrates (CUMV) | museum collection | *CUMV web page*

author of recent article:

- Two classes of deleterious recessive alleles in a natural population of zebrafish, *Danio rerio* | recent journal article | *via Ingenta Select*

**Keywords:** evolutionary biology, ichthyology, paleobiology, systematics

"I am interested in the evolution of biological diversity, particularly in fishes. Research in my lab focuses on one or more of the following areas of interest: (1) speciation, through study of the evolution of species flocks of fishes, (2) the evolution of phenotypic characters in a phylogenetic context, and (3) study of the origin of variation on which natural selection can act. Fishes studied include cichlids, needlefishes and their relatives, danios, swordtails, and basal actinopterygian (rayfined) fishes, both living and **fossil**."

[find any recent co-authors from Biosis or PubMed](#)

More search options for **fossil**:

Figure 4.3. An individual entity display in VIVO, showing links to related entities

An important feature of VIVO is that the creation of a relationship from one entity to another also establishes an inverse relationship to the original entity. This bidirectional structure inherently emphasizes cross-relationships among the content—departments to faculty, faculty to courses and publications, equipment to laboratories, or online resources to an intended audience—while providing navigation paths for

browsing across the virtual community of the life sciences. The VIVO database structure also supports simple inferencing operations that utilize the relationships among entities. For example, figure 4 illustrates how the relationships among entities can be used to aggregate all the publications of affiliated faculty, and to assemble dynamic lists of departmental publications.

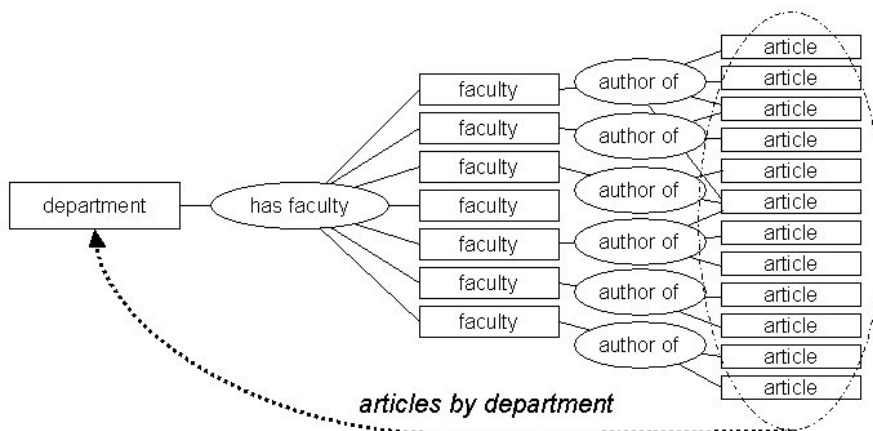


Figure 4.4 Leveraging relationships to report articles by department

Entity-relationship models of this type are not unique. The term "ontology" originates in the field of philosophy as the study and characterization of what things exist in the universe, but the fields of computer and information science have adopted the term to refer more specifically to models of entities and their relationships or interactions,<sup>21</sup> or “a formal representation of technical concepts and their interrelations in a form that supports domain knowledge.”<sup>22</sup> We used the Protégé<sup>23</sup>

<sup>21</sup> Robert Stevens, Carole A. Goble, and Sean Bechhofer, “Ontology-based Knowledge Representation for Bioinformatics,” <http://www.cs.man.ac.uk/~stevens/papers/briefings-ontology.pdf> (8 Mar. 2008).

<sup>22</sup> Jet Propulsion Laboratory, California Institute of Technology, “Semantic Web for Earth and Environmental Terminology (SWEET),” <http://sweet.jpl.nasa.gov/ontology/> (8 Mar. 2008).

<sup>23</sup> Stanford Medical Informatics, Protégé Project, <http://protege.stanford.edu> (8 Mar. 2008).

ontology editor from Stanford University to create a prototype in the W3C Consortium's OWL<sup>24</sup> format, but the working ontology behind VIVO is a local implementation in Java and MySQL to allow direct Web-based editing and display of the contents populating the ontology, not just the ontology itself. Most users will not in fact be aware of the ontology except indirectly through the fluid cross-navigation VIVO offers.

VIVO's separation of the ontology (entity types and relationships connecting them) from the content (the instances) allows further flexibility to modify the ontology without disrupting existing relationships among data entities. After six months of development, VIVO's original ABC-derived ontology was extended to incorporate additional concepts from the AKT (Advanced Knowledge Technologies) reference ontology, a data model developed in the United Kingdom to support analysis of research collaborations among an association of academic computer science departments.<sup>25</sup> The AKT ontology is based on many of the same underlying temporal, spatial, and event relationships as the ABC Ontology while adding additional entity types (*classes*) and relationships (*properties*) that are directly relevant to university research, including institutional and intellectual provenance relationships.

<sup>24</sup> World Wide Web Consortium, "OWL Web Ontology Language Overview," <http://www.w3.org/TR/owl-features/> (8 Mar. 2008).

<sup>25</sup> Advanced Knowledge Technologies, "AKT Reference Ontology," <http://www.aktors.org/publications/ontology/> (8 Mar. 2008).

## VIVO as a Form of Metadata

Vivo can be considered to be metadata at both a macro and a micro level. At the micro level, every entity's relationship to another entity can be compared to the familiar attribute-value pairs of Dublin Core metadata. For example, a journal article in VIVO is defined to have a relationship called "has author" to a person, much as that journal article described in a Dublin Core metadata record would have a "creator" element populated with a value, the author's name. VIVO's <entity><relationship><entity> triplet includes an explicit designation of the subject of the relationship that would be implicit in a set of DC metadata attribute-value pairs associated with the article.

Figure 5 is a screen shot that demonstrates how VIVO displays and groups the results of a simple search. The results from a search for "fossil," highlighted in bold, are displayed in context under four broad categories: people, activities, organizations, and publications. Although none of the names or titles under the people category contains the word "fossil", the term is found within the longer descriptions that are accessed by clicking on a person's name; most of the remaining entries do show the term highlighted in their titles. For additional information, the user can link directly to the person's own website using the URL at the right of each entry. Sorting by type allows users to focus on particular results of interest—grants or courses, for example--without having to first scan each result to determine what it is.



**PEOPLE:**

▣ **Cornell faculty** (5)

- Cisne, John L. | Cornell faculty member | [Earth & Atmospheric Sciences profile](#)
- Howarth, Robert W. | David R. Atkinson Professor | [Ecology & Evolutionary Biology profile](#)
- Likens, Gene E. | Cornell adjunct faculty member | [Institute of Ecosystem Studies profile](#)
- McCune, Amy R. | Cornell faculty member | [Ecology & Evolutionary Biology profile](#)
- Nixon, Kevin C. | Cornell faculty member | [Plant Biology profile](#)

▣ **non-Cornell faculty** (1)

- Lovett, Gary M. | external research scientist | [Institute of Ecosystem Studies profile](#)

**ACTIVITIES:**

▣ **Cornell semester course** (2)

- Human Biology and Evolution | 3 credit course | [ANTHR 275 / BIOEE 275 / NS 275](#)
- Plant Evolution and the **Fossil** Record | 3 credit course | [BIOPL 448](#)

▣ **research grant** (2)

- DESSERTATION RESEARCH: E. HERMSEN: **FOSSIL** HISTORY OF THE SAXIFRAGACEAE SENSU STRICTO AND THEIR WOODY RELATIVES, CRETACEOUS TO PLEISTOCENE | [Research Grant](#)
- EXTREMELY DIVERSE **FOSSIL** FLORAS FROM THE PALEOGENE OF PATAGONIA, ARGENTINA: IMPLICATIONS FOR ORIGINS OF HIGH PLANT AND INSECT DIVERSITY IN SOUTH AMERICA | [Research Grant](#)

**ORGANIZATIONS:**

▣ **research program, unit, or center** (1)

- Invertebrate **Fossil** and Recent Mollusk Collections (at the Paleontological Research Institution) | [museum collection](#) | [PRI web page](#)

**PUBLICATIONS:**

▣ **news release** (1)

- Climate change and growing vehicle traffic are increasing nitrogen pollution in nation's coastal waters, Cornell-led study finds | [Cornell news release](#) | [02/19/2005 news release](#)

**recent journal article** ▣ (6)

▣ More search options for **fossil**:

Figure 4.5. VIVO display of search results grouped by type

At the macro level, since none of the actual articles, databases, or (of course) people indexed in VIVO are actually stored in VIVO as a repository, the whole of VIVO can be considered metadata. From this perspective, VIVO provides much of the traditional functionality of metadata to enable the discovery of data and facilitate access to it. Unlike metadata which has been abstracted and removed into static lists of elements, however, VIVO maintains live, multi-directional connections among its entities to allow fully dynamic interaction and traversal of its structure. A user's view is

not restricted to a single browse or search interface, and individual search results are not displayed in isolation, but linked directly to any and all associated resources, thereby providing significantly more context for users as they explore the life sciences at Cornell.

While the VIVO approach has many advantages for the user over more statically defined metadata, long-term sustainability will be a challenge. As an index of current activities rather than a repository, VIVO has value only if populated with a continuous flow of new and updated information. While the bulk of content to date has been gleaned manually by systematically traversing Cornell websites to find people, organizations, activities, and laboratories active in the life sciences, certain content areas such as recent publications can be harvested on a regular basis from Biosis<sup>26</sup> and PubMed.<sup>27</sup> To remain viable, our long-term strategy must include harvesting time-sensitive content from existing but isolated central databases of students, employees, courses, news releases, events, and the like. We have recently incorporated imports of active research projects from the Cornell Office of Sponsored Programs data warehouse, and Mann Library is participating in a campus-wide initiative to develop and register Web services in support of data sharing including a calendar of events.

The value of any metadata structure or schema will be limited if it cannot interoperate—through query or transformation—with metadata encoded in other schemas. This challenge was recognized and addressed by the ABC Ontology project,

<sup>26</sup> The Thomson Corporation, Biosis, <http://www.biosis.org/> (8 Mar. 2008).

<sup>27</sup> National Center for Biotechnology Information (NCBI), Pubmed, <http://www.ncbi.nlm.nih.gov/sites/entrez> (8 Mar. 2008).

and more currently by the SIMILE project at MIT.<sup>28</sup> Querying metadata across multiple formats or converting metadata from one schema to another requires more than simply establishing the format and syntax for exchange. There must be some agreement on the meaning or the semantics of metadata to assure that one source uses definitions compatible with another. While such concepts as the title of a publication seem straightforward, confusion can still arise over multiple titles, the inclusion of initial articles, and sort order. For this reason, current best practices in both metadata schemas and ontology design include explicit references to external standards, such as Dublin Core, from which terminology has been derived. With such references, it becomes much simpler to confirm a common definition of title, creator, publisher, or date when comparing values defined in different ontology frameworks. Human evaluation and reconciliation of terminology is very expensive, however, and more formal tools for encoding the meaning of terminology used in metadata schemas and ontologies are being developed, both to support preservation of complex, interrelated data, and to promote data preservation in neutral preservation formats and data exchange.

The AKT project addressed the issue of access to the data encoded in their ontology by optimizing search and retrieval access to the <object><relationship><object> triples. The Resource Description Framework (RDF) standard, expressible as XML and in alternative compact notations, also provides a mechanism for encoding object hierarchies and their property relationships, and is one

<sup>28</sup> Semantic Interoperability of Metadata and Information in unLike Environments (SIMILE), <http://simile.mit.edu> (8 Mar. 2008).

obvious choice for export of data from systems such as AKT or VIVO into a neutral format with sufficient standardized structure such that most information could be retained.<sup>29</sup>

A more immediate requirement is to make VIVO a resource for query and harvesting by other websites and other metadata archives such as the National Science Digital Library (NSDL).<sup>30</sup> In effect we need to extend VIVO to “speak” the languages of several common metadata schemas including Dublin Core, to support queries for specific types of content, and to adequately identify the content returned from a query. Internal crosswalks can be built from the VIVO ontology, producing XML-encoded content for OAI harvesting. If demand is sufficient, the same functionality could also be implemented as RSS feeds or a Web service. This would enable other applications at Cornell, for example, to retrieve and display VIVO entities such as event listings within their own interface in real time.

The second goal for making VIVO accessible as a metadata service is to be able to respond to an external query in a more complex fashion, encompassing the full context of each entity retrieved as it would be displayed within the native VIVO interface. For example, in the case of a person, it would be useful to display the full range of activities, affiliations, publications, and other associations that he or she may have with other VIVO entities, along with pointers to each of the related entities. As an example of this model, the National Agriculture Library's thesaurus Web service is capable of responding to a user query with a list of not just the matched thesaurus terms, but the full set of broader terms, narrower terms, related terms, “see also”

<sup>29</sup> National Science Digital Library (NSDL) [home page], <http://nsdl.org/> (8 Mar. 2008).

<sup>30</sup> World Wide Web Consortium, “Resource Description Framework (RDF)” <http://www.w3.org/RDF/> (8 Mar. 2008).



terms, and so forth.<sup>31</sup> More complex responses in the form of sets of entities and their interrelationships could allow remote applications such as the NSDL to reproduce much of the VIVO interactive experience without duplicating VIVO metadata internally, allowing users much more capability to view related content from multiple sources in a more coherent fashion than unstructured Web search engines.

Ironically, one area where VIVO falls short is in integrating library content—the project was initiated in order to look beyond library resources, but it may need to circle back and incorporate more seamless access to the library catalog, to licensed bibliographic databases, and other resources now available largely through independent interfaces. The seamless integration envisioned by McLean and Lynch has not yet been fully achieved.

Web service models would theoretically make it possible for VIVO to operate solely as a request broker – receiving incoming requests and launching federated searches on its own in real time to find answers from a set of known Web services at Cornell or elsewhere, while retaining only minimal data in order to deliver perpetually current information. However, we expect to emphasize a harvesting model for the foreseeable future, in large part because the value added by seamlessly “connecting the dots” among entities is the unique feature and probably the biggest asset of VIVO. For example, it might one day be possible to retrieve a faculty member’s recent publications, current research projects, the courses he or she teaches, and any upcoming seminars from separate Web services in real time, but not the coherent secondary and follow-up information about co-authors, research sponsors, related

<sup>31</sup> National Agricultural Library, NAL Agricultural Thesaurus, <http://agclass.nal.usda.gov/agt/agt.shtml> (8 Mar. 2008).

curriculum, or other speakers in the same seminar series. These secondary connections are not likely to be as easily discovered and assembled on the fly, and a persistence layer for these interconnections provided by services such as VIVO will probably be the only reliable way to deliver this functionality in the near future, given the limitations and instability of a widely distributed system. The ongoing effort by the federal GeoSpatial One-Stop (GOS) Initiative to migrate the FGDC's flawed and unreliable model of distributed Z39.50 searching into a more stable and reliable centrally harvested metadata repository is a case in point.

### **Conclusion**

Although CUGIR and VIVO are very different forms of digital library resources serving different communities and different needs, they share certain common principles and illustrate common challenges for online repositories in general. Both systems rely heavily on relational databases to manage all aspects of the repository, from fundamental data organization to presentation and delivery. These databases have both been set up to model the objects and interrelationships inherent in the respective content domains of geospatial data and a university research community. We have found that if the data model reflects the underlying information content at a suitable level of granularity, then the necessary functionality to support data entry, management, searching, and retrieval will emerge naturally. With VIVO, the model is more diverse and extensible, while CUGIR targets a narrower application and supports a large traditional body of content. Both systems internalize their metadata and use the knowledge inherent in the metadata as the core operating information for the entire project. Every level of project functionality depends on some aspect of metadata, and

organizing and maintaining metadata is a central management task for the entire project.

It has been useful for these projects to distinguish between internal and external use of metadata, and also to deliver content to the outside world based on the internal metadata model, rather than seeing metadata simply as an abstraction or by-product of the “real” content. We are importing formerly file-based metadata into the CUGIR database in order to use the information to better serve our users, and to help keep CUGIR and the metadata harvested into the Geospatial One-Stop portal as current as possible. We anticipate major improvements in workflow as we incrementally manage more and more of the CUGIR metadata through a central project database. By having core metadata elements linked via a database, we will be able to deliver metadata that is more location-, time-, and attribute-specific with each download of geospatial data.

VIVO will, in many ways, serve as a test case for the metadata system of the future – a rich, flexible, but powerful structure that can integrate a wide variety of metadata into a coherent structure and deliver it via a very simple searching and browsing interface. Time will tell whether the model can be sustained, and how well it will integrate with other models for metadata management, but so far the structure has proven resilient, and user response has been positive.

The ongoing development of CUGIR and VIVO relies on the library’s role as information organizer, serving as a custodian of books and data with the propensity to add value to anything we acquire. To fulfill that very role requires the full range of expertise in the library and a constant effort to identify changing best practices. Metadata librarians are adapting their cataloging and controlled vocabulary development skills to the new challenges of ontology design and maintenance, and

they are using their project experience to ensure that metadata systems interoperate as much as possible. Collection specialists are rethinking traditional subject guides as they develop more complex online resources, and public services librarians are working with information technology staff toward a goal of improved common functionality across diverse digital library platforms within and beyond Cornell. In combining these skills and applying them to the stewardship of physical and electronic resources, libraries are continuing to contribute significant value to the knowledge within and beyond the university community.