In 1957, Karl Schmidt, the renowned herpetologist of Chicago’s Field Museum, wrote in *Curator*, “Not everyone realizes how different the use of books may be in a museum from the familiar pattern of reading and note-taking in a public library. In a museum, a book may be... tested by reference to a specimen or a series of specimens drawn from the range and laid beside it.”

Schmidt’s comment on the use of scientific literature in natural history museums inspired the IMLS grant-funded project entitled “Linking Florida’s Natural Heritage: Science & Citizenry.” The Linking project was intended to use the specimen as the nexus for pulling together scientific data from museum specimen databases and library catalogs of scientific literature. The goals of the project were to integrate specimen records and bibliographic records about the same species; to create an interface equally easy for scientists, students and laymen to use; and to enhance bibliographic description to make it more usable in a taxonomic and environmental context.

### Project deliverables

Linking Florida’s Natural Heritage provides a single interface to disparate collections[1]. The original grant focused on linking two specimen collections from the Florida Museum of Natural History, ichthyology and herpetology, and four bibliographic files describing scientific literature. The goals of the project were to integrate specimen records and bibliographic records about the same species; to create an interface equally easy for scientists, students and laymen to use; and to enhance bibliographic description to make it more usable in a taxonomic and environmental context.

Three of these files were pre-existing: the FORMIS Ant Bibliography, the Archie Carr Sea Turtle Bibliography, and Everglades Online, a database of citations on the Everglades ecosystem produced by the Everglades Information Network. The fourth file, called Florida Environments Online, was created especially for the project. It initially consisted of 13,380 citations from eight research bibliographies compiled by scientists and state agencies in Florida. An online form was created to allow researchers and agency staff to add to the database and a workshop was held in January 2000 to train contributors in data entry.

To facilitate searching across different environmental citation databases, the project also developed a thesaurus of Florida environmental terms in cooperation with the Florida Geographic Board, the Florida Department of Environmental Protection, the South Florida Water Management District, the Florida Marine Research Institute, the US Geological Survey (USGS) and others. Terms were taken from ENVOC (a multilingual...
thesaurus of environmental terms), the Fish and Wildlife Reference Service thesaurus, Aquatic Sciences and Fisheries Abstracts, Fire Ecology Thesaurus, California Resources Agency Environmental Resources (CERES) Thesaurus, the Florida Natural Areas Inventory classification, the Florida Department of Transportation’s Florida Land Use, Cover and Forms Classification System, and the Federal Wetlands Classification. Gail Clement, who was at Florida International University during the grant period, compiled and harmonized the thesaurus database that now contains 4,376 terms in seven major subject categories. The thesaurus should be considered a work in progress and is available online[2].

Also as part of the project, a “core collection” of 200 seminal texts on species and ecosystems was selected by scientific experts throughout the state of Florida and these materials were digitized specifically for the Linking project. Catalog records for the materials were added to the Florida Environments Online database.

Finally, four educational modules were written and tested in local schools, using the Linking interface for lessons in taxonomy, invasive species, biodiversity, and library and museum collections.

Architecturally, the application is straightforward. A Z39.50 gateway product, OCLC’s SiteSearch, was used to create a user interface offering broadcast searching (now called “metasearch”) of the specimen and literature databases. The interface is shown in Figure 1. It is configured so that a user can search all of the databases at once, choose literature or specimens only, or select any combination of databases to search. Search results are aggregated by the database in which they were found, but displayed in a common labeled format.

The selected databases of literature citations were already in MARC format and accessible through library catalog systems. Because of this, implementing access to them was a relatively simple matter of establishing a Z39.50 connection to the databases. Records that were not in Z39.50-accessible systems were copied into a local library system with a Z39.50 server. The two specimen collections, however, were in Microsoft SQL Server databases at the Florida Museum of Natural History (FLMNH). Linking to these required local development of a Z39.50 server that could translate the SiteSearch client’s query into SQL queries and search the relational databases via ODBC. The tables constituting the result set were then mapped to MARC format records for display.

Museum specimen databases have some commonalties, but are not standardized to the extent that libraries have standardized on MARC. To facilitate the integration of additional specimen databases, we decided to specify a standard view that participating museums would be required to provide. (In relational database technology, a “view” is a special organization of data, created as needed from the base tables.) The view was based on a set of metadata elements for specimen collections developed by the ZBIG (the Z39.50 Biological Implementers Group out of the Museum of Natural History, University of Kansas) known as the “Darwin Core.” For the Linking project, the elements of interest included the specimen identifier and location, taxonomic information, and the collector, date, and place collected.

Although architecturally straightforward, the application is complicated by incompatibilities between the library and museum data. Information elements unique to one category or the other, or elements that were not central to our purpose, were not a problem. For example, we made no attempt to reconcile the form or content of author names and collector names although we are aware that collectors are often authors of articles related to their taxonomic interests. However, we felt that it was critical for species names and geographical location information to be cross-searchable in all of the included databases. Different practices in library and museum communities made this difficult.

Species names

Traditionally, taxonomists who work with museum specimens have used a binomial nomenclature.
refined by Linnaeus. It consists of a two-part Latin name formed by appending a specific epithet to the genus. “The advantage of scientific names is less their stability than that they are the same no matter what the user’s language and that changes in them are governed, in principle, by an internationally adopted Code” (Walker and Moore, n.d.). Thus, scientists around the world know that *Grus americana* is the scientific name for the bird commonly known as the Whooping crane. What somewhat complicates the picture is that some species have been reclassified and renamed and this can lead to multiple synonyms (previously valid names). While taxonomists try to keep up with all the scientific name changes, specimen records may or may not be updated as names evolve. It should also be noted that current technological advances in determining relatedness of species through genetic mapping might eventually disrupt the hierarchical relationships implied in the species epithet. Nonetheless, the names will likely be retained for their value in identification.

Unlike museums, libraries follow the Library of Congress and use common names in their cataloging records. This is clearly an attempt to serve a broader audience of library users than taxonomically literate scientists. However, common names differ even more than scientific names, often depending on the language of the speaker and where the specimen was collected. Thus, a species may have multiple common names and multiple scientific names. The first conundrum of the project was how to search simultaneously museum records containing scientific names, but no common names, and library records containing common names, but no scientific names.

The initial approach was to add scientific names to the library citations, using the Integrated Taxonomic Information System (ITIS). ITIS is a multi-agency project intended to provide a taxonomic authority for the US biological species spectrum. The ITIS database[3] can be searched by common name or scientific name and returns the full taxonomic hierarchy for the species. Common names found in bibliographic records were searched in ITIS and the taxonomic information was cut-and-pasted into the MARC 754 field (Added entry – Taxonomic Identification).

At that time the 754 was defined to contain a repeatable subfield “a” for the taxonomy and a subfield “2” for the taxonomic authority. The example given in the USMARC documentation was:

```
754 ##
$aPlantae (Kingdom)
$aSpermatophyta (Phylum)
```

The Linking project adapted this to a format more easily cut-and-pasted from ITIS, and added common names and synonyms. A populated 754 field looked like this:

```
754 ##
$aAngiospermae (Class)
$aDicotyledoneae (Subclass)
$aRosales (Order)
$aRosaceae (Family)
$aRosa (Genus)
$asetigera (Species)
$agramosuta (Variety).
$s2 [code for Lyman David Benson’s Plant Classification]
```

Experience with this data exposed problems with the way the MARC field was defined. Because subfields did not map directly to taxonomic levels, it was difficult to construct a search for genus name or species name only. Also, embedding labels in the data removed all flexibility from display. The most common form of display, the genus and species name together (e.g. “Rosa setigera”) required extensive parsing to create. These problems led us to submit a proposal to the USMARC Advisory Group in 2000 that was eventually amended and approved in 2001. The currently defined 754 field has a repeatable subfield “c” used to label the taxonomic level in the following subfield “a”. The example given in the MARC21 documentation is now:

```
754 ##
$ckingdom
$aPlantae
$cphyllum
$aSpermatophyta
```

```
Potential populations are forecast by matching environmental parameters supporting known populations to geographic areas with similar characteristics.

In the case of museum specimens, locations are recorded when the specimen is collected. Today, the widespread use of global positioning system units (GPS) leads to accurate location information. However, historic locations are often open to interpretation and accuracy is usually qualified in terms of reliability. Museums usually record a hierarchical structured location: region, country, state or department, county or province.

This may or may not be accompanied by a fuller textual description of the collection site, e.g. “ca. 1/3 mi. w. of NW 98th St., and ca. 2 mi. w. of I-75 and Gainesville, S6, T10S, R19E.” Work being done to convert descriptions into spatial footprints is described under “Future plans reflect current initiatives.”

In the case of marine collections, the location can also include drainage, e.g. Lower St Johns River. Thus, geographic data associated with a Gulf sturgeon could be: North America (Region), USA (Country), Florida (State), Putnam (County), Apalachicola (Drainage).

In libraries, geographic information is given in the 651 field (Subject Added Entry – Geographic Name), but it is not given in a similar hierarchical format. A work on Gulf sturgeon in the Suwannee was assigned the subject heading “Gulf Sturgeon – Suwannee River (Ga. and Fla.)”. While specificity is high, a search for the state “Florida” would not retrieve this record, nor is the heading related to a county.

In order to provide more consistency between the bibliographic and specimen data, a hierarchical form of name was added to the MARC records in the 752 field (Added Entry – Hierarchical Place Name). This field is most heavily used for places related to newspaper circulation area, but it is defined more generally to hold any “hierarchical form of place name that is related to a particular attribute of the described item, e.g. the place of publication for a rare book” (MARC21 Concise Bibliographic). In the Linking project, it was used to provide hierarchical geographic information related to the content of the text. For example, the book on Gulf sturgeon in the Apalachicola River was assigned the heading:

752:: |a United States |b Florida |c Franklin |d Apalachicola River

Beyond the place names in the 752, additional geographical information was added to bibliographic records in order to anticipate their use in a GIS environment. We manually looked up the place names in the USGS Geographic Names Information System database to obtain latitude.

---

**Geographic locations**

In addition to the scientific name, the collecting site is critical to a specimen’s scientific value. The first specimen of a species sets the loci on the earth’s surface for future questions on species distribution, possible genetic lineages, future species viability, and other important vectors.
and longitude coordinates, which were entered into the MARC 034 field (Coded Cartographic Mathematical Data). We also entered standardized forms of county names and hydrological names taken from the Federal Information Processing Standard (FIPS) and the USGS Hydrologic Unit Codes (HUC), federal government standards maintained by the USGS.

The 651 field (Subject added entry – Geographic name) seemed the most appropriate field for the HUC and FIPS data. The field is defined so that the authority for the data is given in coded form in subfield 2, but only if the source is listed in the MARC Code List. At the beginning of the project, the MARC Code List did not have codes for either HUC or FIPS, so we put the source of the term in a subfield x, with a second indicator “4” indicating, somewhat inaccurately, that the source of the term was unknown.

651: 4: |a Monroe |z 12087 |x FIPS
651: 4: |a Apalachicola |z 3130011 |x HUC

An advantage of this approach was that the data displayed in our system as a common subject heading (Monroe – 12087 – FIPS) making it clear to any geographically literate researcher that this was a FIPS code available for searching. Later, at our request, the Library of Congress added HUC and FIPS authorities to the MARC Code List. The same data are now coded:

651: 7: |a Monroe |z 12087 |2 ceeus
651: 7: |a Apalachicola |z 3130011 |2 huc

The second indicator “7” designates that the source of the term is given in subfield 2. Unfortunately, because the FIPS county codes are defined in a document titled Counties and equivalent entities of the United States, its possessions, and associated areas, the code assigned to the FIPS codes was “ceus”, a designation that nobody would know. A better code would have been “fips55”, which uniquely identifies the county codes among FIPS. We are planning to submit a change request to the Library of Congress.

Current status

At this point in time, nearly seven years later, Florida Environments Online has grown from some 13,140 records at the end of the grant period to over 26,000 records in December 2003. Citations from other databases have been ingested, although no effort has been made to identify or remove duplicates, and indexing/cataloging idiosyncrasies of contributing agencies have been accommodated. Florida records from the Aquatic, Wetland and Invasive Plant Information Retrieval System (APIRS) (5,000 records) were added in 2003.

The digitized “core collection” has now grown far beyond the original 200 texts. The first 15 volumes of the Bulletin of the Florida Museum of Natural History are also available, as are some 338 publications and 126 maps of the Florida Geological Survey. Projects are underway to digitize the remaining Florida Geological Survey series, and to add the major series written by staff of the University of Florida’s Institute of Food and Agricultural Services and the Howard T. Odum Center for Wetlands.

The number of databases available for broadcast search has also grown with the addition of one bibliographic database and six museum specimen databases. The new bibliographic file is the South West Florida Environmental Documents Collection covering Estero Bay, Charlotte Harbor, and the Caloosahatchee River. Specimen databases added include the mammalogy and herbarium collections of the Florida Museum of Natural History, the ichthyology collection from the Florida Marine Research Institute, the bird specimen collection from the Tall Timbers Research Station, and the bryophytes and lichens collection and vascular plants collection from Camp Blanding.

Usage statistics

Project Web site. Use of the project Web site remains fairly constant with about 4,000 accesses a year from 2000 through 2003. The total number of searches recorded on the Z39.50 server is 20,278. Of particular interest is the use by institution. Access from the ten public universities in Florida are coded separately, all other users are lumped into an unknown category. Over the four years, the highest number of uses came from the unknown category, followed by use from the University of Florida and beginning in 2002 from Florida Gulf Coast University. Figure 2 shows usage of the site between the years 2000 and 2004.

At this point in time, we have no data on the querying methods used by searchers using this interface nor do we know if the project is fulfilling their research or educational imperatives. As the statistic gathering becomes stabilized, we will probably explore some Web based surveys to determine how people are using this system and what suggestions they might have for improving it. We do feel that usage could be higher, given the utility of the application, and intend to pursue several avenues of promotion, including placing notices in agency publications, adding links to related Web sites, and publicizing the availability of the educational modules.
Core collections. More than 200 digitized full text documents associated with the project were moved to a new server in 2002. Thus, consistent use statistics are available for 2002 and 2003 only. However, they are of interest because of the phenomenal increase in use in one year. In 2002, tables of contents of core collection documents were displayed 5,425 times; in 2003, this increased to 58,168 times. In 2003, the table of contents from Bulletin #3 of the Florida Geological Survey, “Miocene gastropods and scaphopods of the Choctawhatchee formation of Florida,” by W.C. Mansfield was viewed 1,101 times. In second place was Ida Cresap’s The history of Florida agriculture with 829 viewings.


Future plans reflect current initiatives

Linking Florida’s Natural Heritage is a stable application offering unique functionality related to Florida’s species and ecosystems. It could remain as it is for some time, with most of our efforts directed to increasing awareness and use. At the same time, there are both technical and functional reasons to re-evaluate the architecture and content.

The mapping table approach to relating common and scientific names requires the maintenance of a static local thesaurus. The way it is implemented, as an “under the covers” process, means that unsuspecting users may get unanticipated or even incorrect results. The implementation also restricts its use to find common names for specimens. A better approach would make use of new developments such as uBIO, the taxonomic name server being developed by Cathy Norton and Dave Remsen at Woods Hole. This initiative is creating a concordance of common and scientific names that is tied into the major taxonomic work being carried on by ITIS, the Global Biodiversity Information Facility (GBIF), Species 2000, and others.

Integrative software is being developed so that individual projects such as Linking Florida’s Natural Heritage can make use of the powerful concatenating, taxonomic name server to help users formulate their queries. This would obviate the local maintenance problem, and add assurance that all identified name variations were included in user queries.

Another architectural problem is that the way results (citations and specimen records) are returned as HTML makes it impossible to use the data in a geographic information systems (GIS) environment. For many researchers, the ability to plot a selected set of specimens on a map by their collection location would greatly enhance the value of the system. In order to do this, however, the result set would optimally be formatted in a tabular form and is made available as a downloadable file. The nature of the SiteSearch application precludes our supplying results in this way.

Because of the research utility in applying GIS functionality to bibliographic works as well as natural history data, future enhancements of the project would facilitate spatial analysis of both forms of information. Two current projects: one
from the museum world and one from the library world are of particular interest.

Recently, a new site developed by Tulane University’s GEOLocate[6] attempts to georeference specimens using textual descriptions. GEOLocate software is available for free download or it may be used through the Web interface. The description mentioned in this paper, i.e. “ca. 1/3 mi. w. of NW 98th St., and ca. 2 mi. w. of I-75 and Gainesville, S6, T10S, R19E.” produced a latitude of 29.65139 and a longitude of –82.3749051520174.

The Perseus Digital Library[7] being developed at Tufts University creates spatial visualizations of document texts. While the current collaborations relate to mapping classical works, the functionality could be applied to natural history publications as well. Essentially, place names in a text would be displayed on a map and the user can shift between the text and the document.

Ideally, shifts between maps, texts, and specimen data should all be supported.

A major architectural redesign would be required to fix the limitations of either the local taxonomic/scientific name thesaurus or the lack of spatial analysis. If such a redesign were undertaken, we would probably also reconsider whether the application is best implemented through Z39.50 metasearching or whether another model is more appropriate.

The original work of the Z39.50 Biological Implementers Group (ZBIG) also used Z39.50 as the communications protocol for integrating natural history collection data. Limitations they encountered included the following.

- Complicated protocol specification means a very steep learning curve for developers.
- Protocol not well understood by network administrators, and hence they are reluctant to open the necessary network port (even though Z39.50 is far less likely to allow a security breach than HTTP).
- Conceptual schemas are not defined with a formal language such as XML Schema.
- Limited support for XML and Unicode (although this has improved greatly over the last couple of years) (Relationship of DiGIR to Species Analyst, 2003).

While Linking Florida’s natural heritage encountered few problems using Z39.50 to integrate bibliographic collections, individual customized views had to be created for each specimen collection and the difficulties noted above occurred in the Linking project as well.

One option is to investigate the SRW/SRU services developed as a kind of “Z39.50 Lite” by the Z39.50 International Next Generation (ZING) initiative. Another alternative is a protocol developed specifically for exchanging natural history collection data, Distributed Generic Information Retrieval (DiGIR). This protocol is “based entirely on the use of XML documents for messaging between clients and data providers, with a data transport mechanism that was predominantly based on HTTP. DiGIR is designed from the ground up to offer the same capabilities as Z39.50 except using simpler technologies and a more formal specification for description of information resources.” (Relationship of DiGIR to Species Analyst, 2003).

DiGIR uses for its transport syntax an XML schema based on the Darwin Core called Access to Biological Collection Data (ABCD). It appears that the DiGIR protocol could be adapted to work with XML metadata for bibliographic works.

The globalization and digitization of natural history information is leading to scientific analysis and research that was previously unimaginable. In the seven years since the Linking project was initiated, the cutting edge technologies of 1997 have rapidly given way to XML, harvesters and portals. New technologies in turn raise expectations for functionality. We prophesize that eventually, the rematrixing of metadata will lead to infinite integrations of text and other forms of data. One of the greatest challenges now is to be aware of digital project applications that already exist, and to conceive and design a system architecture that allow their integration to provide new functionalities.

Notes
1 Available at: palmm.fcla.edu/fnh/
2 Available at: palmm.fcla.edu/fnh/thesauri/feol2/
3 Available at: www.itis.usda.gov
4 Available at: palmm.fcla.edu/fnh/matrix/T/
5 Available at: www.ubio.org
6 Available at: www.museum.tulane.edu/geolocate/default.aspx
7 Available at: www.perseus.tufts.edu/

References
Relationship of DiGIR to Species Analyst (2003), available at: http://speciesanalyst.net/docs/digir/