1-1-2006

Yale Leaf Morphology Digitization and Network Project

David Stern
Illinois State University, hstern2@ilstu.edu

Yale Leaf Morphology Digitization and Network Project

Abstract

This article describes a digitization project inspired by the innovative leaf morphology classification work of a faculty member in the Geology and Geophysics Department and the Peabody Museum at Yale University. We began our initiative by scanning the *Flora Fossilis Arctica*, a 7-volume fossil leaf identification tool covering various geological areas, published between 1868 and 1883. This classic paleobotany resource was digitized, creating tiff, pdf, and searchable pdf files. We are now converting the searchable pdf files into ASCII text, enhancing the raw data with metadata elements, placing this material on the web for searching and display; and linking this material to an existing set of preserved leaf plates, a locally created index of annotated article clippings, an online leaf morphology tutorial, and the published online literature. Many decisions must be made in terms of host platforms, mark-up standards, search and linking options, and preservation documentation. This article will outline our decision process as we explore the post-digitization dataset handling, which may prove instructive for others attempting to create and link locally digitized materials.

In the area of good management, there are the Four "D"s: Do, Delegate, Delay, and Drop. Following these principles ensures an effective use of limited time and resources. In the area of digitization projects we now define Four "D"s: Data, Design, Develop, and Deploy. These steps ensure a logical and well-designed implementation for a digitization project. Data requires brute computer power, automated scanning and OCR for items not born digital, and appropriate equipment and software. Design requires expertise in data structures and vision for proactively considering potential by-products. Develop requires collaboration skills with various population experts and the discipline to incorporate standards. Deploy requires time, personnel, and budget management and a dedication to assessment and revisions. We will discuss these stages in our project as we inch closer to an integrated digitization product.

History and Intention

The Peabody Museum connection

The initiative began with the identification and location of the library volumes that were spread across a variety of library locations. The item is: Flora fossilis arctica. Die fossile flora der polarlnder enthaltend die in Nordgrnland, auf der Melville-insel, im Banksland, am Mackenzie, Heer, Oswald, 1809-1883. and Flora fossilis arctica. Die fossile flora der polarlnder enthaltend die in Nordgrnland, auf der Melville-insel, im Banksland, am Mackenzie, in Island und in Spitzbergen entdeckten fossilen pflanzen von Dr. Oswald Heer ... Mit einem anhang ber verste Zrich, F. Schulthess, 1868-83.

The professor's own copy of the work was incomplete and damaged, and we located and gathered the missing official library copies for his immediate research needs. The condition of

the pieces ranged from slightly unstable to badly deteriorating and damaged. Many of the hand drawn leaf images were yellowing and becoming dangerously jeopardized as pages showing signs of tears and crinkles were falling out of the bound volumes. We decided that these rare volumes were prime candidates for preservation and/or reproduction.

The six distributed departmental science libraries had just initiated a limited preservation program with the assistance of the main library preservation staff. We were identifying and addressing simple repairs with local tip-ins and taping, having more significant repairs performed in the main library preservation department, sending simple rebinding to the commercial bindery, and we had just begun using the same commercial bindery for new "preservation photocopying" of materials that were beyond preservation. Our normal preservation photocopy product was a new bound paper copy. Images were often enhanced during this process, with the final product being an improved version rather than a simple facsimile. On occasion we would have the original item returned in a box for preservation as an archival item; usually due to important original colored plates, hand written annotations by historically significant scholars, or the rare nature of the original work.

In some cases the only way to maintain a working copy of these rare works is to unbind them and create new paper copies for intensive use. These can be delicate decisions - balancing the stewardship of the original items and the need for a more accessible manifestation of the contained information. These rare paleobotany materials provided a perfect opportunity for preservation photocopying.

In this instance there were many campus entities expressing interest in improved access options to this type of material, both in terms of immediate utility for the specific tool and as a long-term learning case. This material was seen as a logical dataset that could be used to develop a prototype digital library of interest to multiple units on campus and at the same time be immediately beneficial to worldwide scholars.

While the original intention was to scan the *Flora Fossilis Arctica* in order to preserve the disintegrating material (including drawings) for future researchers, we quickly determined that we could easily serve this rare data to a wider audience as image files mounted on the web. However, we were even more interested in taking the additional step of creating entirely new research tools and research possibilities by converting the static data and leaf images into dynamically manipulatable digitized data.

Professor Leo Hickey, Professor of Biology, Professor & Chairman of the Department of Geology & Geophysics, and Curator of Paleobotany at the Peabody Museum, expressed great enthusiasm for the enhanced capabilities of the described post-processing options. These new ways to search the data directly matched his long-term efforts to create authority records for duplicate discoveries of species identification. The Compendium Index of North American Mesozoic and Cenozoic Type Fossil Plants <**http://www.peabody.yale.edu/collections/pb/pb_compendium.html** >
was created for exactly this purpose.

The Compendium is a locally produced card file of 20,000 historically important fossil leaf morphologies and images and priority claims (from 232 literature references) -- a world renowned master file of leaf paleontology. This index provides two simultaneous services: it co-locates redundant discoveries over time and places these items within a unique descriptive morphology classification scheme. The index also provides chronological and geological elements for each specimen. The ability to search and create concordances would make recognizing additional duplicate discoveries and complementary records much easier. Tracing historical references to the appropriate first identifications within the *Flora Fossilis Arctica* would provide a rapid and automated way of reviewing the resulting citation pathways and identifying citation error patterns. The creation of geographic, epoch, and morphology indexes to the online *Flora Fossilis Arctica* would make reviewing and mining this complimentary data much more powerful.

In 1937 the late Princeton University paleobotanist Erling Dorf decided to compile the *Compendium Index of North American Mesozoic and Cenozoic Type Fossil Plants.* Dorf greatly facilitated the identification of fossil plants by arranging these cards into a unique set of numbered morphological categories (such as leaf shape and major venation type) that grouped like forms with one another regardless of their professed taxonomic assignments. During his lifetime Dorf was able to assemble over 10,000 cards from 140 paleobotanical references. After Dorf's death the *Compendium Index* was transferred to the Yale Peabody Museum's Division of Paleobotany. It presently covers fossil floras from North America, including Greenland, starting in the Triassic Period (240 million years ago) and extending to the Pleistocene or Ice Age, which ended about 10,000 years ago. Over 93 references have been added in the last 20 years, and the *Compendium Index* has grown from 10,000 cards to approximately 20,000 cards, with 9,881 entries from 235 references dating from 1866 to 2003. The *Compendium Index* is an invaluable resource that exists only at the Yale Peabody Museum.

This sample record from the *Compendium Index* shows the card layout and the fields into which it is organized. Each 8" x 10" card has an illustration and a description of a fossil plant species pasted on the front (on right, above) and reverse (on left), respectively. The front of the card also has the species name, its geological age, the formation and locality where it occurs, its status, and the citation of the reference where it was published

This digitization project would allow for the seamless linking of the original *Flora Fossilis Arctica* entries with the annotated *Compenium Index* material. Professor Hickey also proposed the linking of these digitized records with other related Peabody collections.

The first additional link is to the National Cleared Leaf Collection. **http://www.peabody.yale.edu/collections/pb/pb_clearedleaf.html**

This reference collection consists of over 6,500 cleared, stained and mounted extant leaves. While at the Smithsonian Institution Curator Professor Hickey began this collection in 1967 as part of his research on the systematic distribution of the leaf characters of the flowering plants in relation to the evolution of a group. This collection was transferred with him when he came to the Yale Peabody Museum as Director in 1982. The National Cleared Leaf Collection remains an integral part of research for national and international scientists.
The second collection link is to the Yale Paleobotany Collection and Online Catalog http://www.peabody.yale.edu/collections/pb/pbhist.html#pbtoday.
The Yale Peabody Museum's paleobotany collection numbers over 150,000 specimens, with 4,200 of these type and illustrated specimens. The collection is worldwide in scope, with approximately 75% of the collection from North America and the other 25% from the Arctic, Australia, Central American, Europe, Israel, Pakistan, Lebanon, South America and the West Indies. This collection is one of the most historically significant in the United States. Well before the establishment of the Museum in 1866, Yale's first geologist Benjamin Silliman assembled a teaching collection that included a substantial number of fossil plants. On Silliman's retirement in 1853, Yale purchased this collection, and many of its specimens remain among the holdings of the Paleobotany Division today.

The collection has seen unparalleled growth with the addition of 2 orphaned collections: The New York Botanical Garden Collection and a substantial part of the Princeton University paleobotanical collections. These holdings include material that formed the basis of the research of many of the founders of American paleobotany, including J.S. Newberry, Leo Lesquereux, E.W. Berry, W.M. Fontaine, Lester Ward and Arthur Hollick. In June 2002 the Paleobotany Division moved its entire collection of over 150,000 fossil plants into the new state-of-the art facilities of in the Class of 1954 Environmental Science Center. In support of this move, the Division of Paleobotany was awarded a grant of $365,346 from the National Science Foundation to help purchase and install the mobile compact storage system and hire personnel to assist in the

move, reorganization and electronic cataloging of the paleobotany collections. This grant project was completed in October 31, 2004.

The Yale Paleobotany - Online Catalog is located at <http://george.peabody.yale.edu/pb/>. The online holdings of the Paleobotany collection contains all type specimens, and approximately 50 percent of the non-type catalogued material.

These new integrated search and link options may produce new research areas and will certainly change the historical record for these fossil discoveries. These links will also complement Professor Hickey's creative use of the Manual of Leaf Architecture as a learning tutorial as it also becomes integrated into the network.

The Manual of Leaf Architecture (Morphological Description and Characterization of Dicotyledonous and Net-veined Monocotyledonous Angiosperms.)

http://www.peabody.yale.edu/collections/pb/pb_mla.html

The main goal of the The Manual of Leaf Architecture is to define and illustrate for the reader an unambiguous and standard set of terms for describing leaf form and venation, particularly of dicots, and also to provide a template and set of instructions that show how descriptive information can be entered into a standardized database of fossil and extant leaves.

Future explorations

During our conversations we also identified related digitization and preservation projects that are logical extensions of these explorations. In the future we anticipate developing joint Institute for Library and Museum Services (IMLS) grants for:

(a) the digitization of the *Compendium Index* - (see above). This would address preservation, enhanced searchability, and web-based distribution of this teaching and research material.

(b) the conservation of the National Cleared Leaf Collection - the largest collection of its type, and one of the most carefully described collections.

and

(c) geo-referencing of all specimens for teaching and research purposes based upon the BioGeoMancer tool set. http://www.biogeomancer.org/

One complication for this next phase may be obtaining permissions for the clipped information on the cards; we will need to obtain publisher permission for the images and text for approximately 123 references. We hope this will not be a problem, but rather a technicality, as the material is already recognized as an important research tool by the community.

Other Interested Partners

The Library and the Peabody Museum also share an interest in developing seamless linking to a broader range of campus resources. There is a Collections Collaborative grant for exactly this type of cooperative information discovery between the library and other campus resources, and we will soon submit a proposal for matching funds. The principle investigator for this grant and the author both sit on the campus Integrated Access Council that attempts to create such campus-wide collaborations.

One particular Peabody staff member, Reed Beamon, has a complimentary interest in both automatic geo-referencing software and integrating international tools into the rapidly developing paleontological portal scholarly network
http://www.paleoportal.org/

The Bridgeport National Bindery http://www.bnbindery.com/ has an interest in developing enhanced digitization capabilities to complement their commercial bindery and preservation photocopy operations. They have recognized the drop in traditional paper binding as ever-larger numbers of journals migrate from paper to online products. The bindery also recognized the market for large-scale preservation photocopying and has supplemented their tiff digitization and printing operation with an on-demand printing service based upon either their stored PDF materials or publisher provided digital resources. The bindery serves as a national on-demand clearinghouse printer for selected collections and publishers.

******** INSERT FIGURE 1: creating text data from images **********

Their equipment has the ability to generate searchable pdf files as well, but we were the first client to request such by-products. In a casual conversation at a local library conference we decided to explore the conversion of simple images into searchable and somewhat manipulatible files. We requested that the bindery perform Step 2 of Figure 1. We recognized that searchable PDF files still had limited capabilities compared to fully marked-up text files, and the bindery was interested in our explorations of this next level of metadata encoding, searching, and linking. We intend to perform Step 3 and Step 4 of Figure 1 in-house.

The library itself was interested in creating and exploring the power of searchable OCR material, offering knowledge management opportunities far beyond the limitations imposed by the searchable PDF files. During the process we discovered that the bindery could not provide all pages as searchable PDF; pages with color illustrations were excluded from the process due to software limitations, and we will need to complete Step 2 locally for a portion of the material.

Converting the searchable pdf files into marked-up ASCII would allow for more sophisticated searching, linking, and post-search repurposing of the original data. This effort would require

us to determine logical host platform(s) that could store and serve this large amount of data, and also provide linking and seamless integration to the relevant non-bibliographic material located across the international scholarly networks.

The library is currently developing a rescue repository for holding material in a variety of media types, and our digitized *Flora Fossilis Arctica* material could be temporarily housed in this repository for short-term preservation. The Yale University Library is just beginning to develop a fully functional institutional repository (IR) which will address the long-term support issues such as life-cycle considerations and service interfaces. For the moment our embryonic FEDORA-based IR implementation will not serve as an adequate service platform, and we will need to explore alternative platforms.

We will now review the information and progress we have made in the Four D areas.

Technical Details

1. DATA

As stated previously, the majority of our raw data creation process began with a commercial digitization of the paper material into image and searchable image files. The process of locally extracting ASCII searchable text and creating metadata enhancement began with an exploration of full-text collections and mark-up options in use at other digital repositories. Lists of digital libraries are found at locations such as http://sunsite.berkeley.edu/Collections/ and http://www.academicinfo.net/digital.html. This review process quickly became a study in metadata standards, communications syntax, and research community vocabularies. The details of the technical requirements for mark-up are discussed later in the Design portion of this article.

In terms of subject-based initiatives, we discovered a number of collaborative paleobotany networks with pointers to various servers, but there was no federated searching or tightly coupled system in place. There is an active community of researchers and museum technologists working on this problem. http://www.nhm.ac.uk/hosted_sites/paleonet/

The simplest full text search option would be created by loading the searchable PDF files onto a networked Adobe Acrobat server. As this provides minimal extended metadata creation and storage benefits, and unknown levels of embedded link possibilities, we chose to not use this approach to create our final product. We are briefly testing search and retrieval using the Acrobat software on a dedicated workstation in order to be prepared for future initiatives where this might be an adequate solution. We also briefly explored the Acrobat software application that creates ASCII text from searchable PDF documents. This ability to generate simple flat ASCII files will be used to complete our color page conversion task and might be useful for other projects in the future.

However, we chose to explore implementing more robust text file creation methods in order to provide expanded searching and linking possibilities. Our embryonic Fedora architecture will ultimately provide very powerful search and delivery options, but we need to explore more immediate options until service modules are developed. So we began to investigate other

ASCII-based possibilities.

There are a number of other ways to convert our searchable PDF material into ASCII text (for eventual mark-up). There are a number of low-cost http://www.verypdf.com/pdf2txt/pdf2txt.htm and free open source programs that extract plain text from pdf source material. Many such tools can be located by performing a web search for "PDF and ASCII and conversion".

Another possibility is to use slightly more powerful ingest software that is part of various institutional repository software toolkits. We are exploring the use of the commercial VITAL software suite and related tools such as the ELATED and FEZ software found at http:// www.fedora.info/tools/ which produce both raw ASCII text and some level of metadata (descriptive technical information about the file itself, but not the content).

We have found that generating the raw data is the easiest part of the process. Structuring and marking the data for manipulation requires far more effort, resources, creativity, and planning.

2. DESIGN

In terms of a search and delivery system, a low-level development option is to load flat ASCII (without any metadata elements) into a text application, using software such as OpenText Livelink http://www.opentext.com/2/products/pro-km/pro-km-ll-libraries.htm. This flat text could be enhanced through the introduction of field-delimited elements serving as embed metadata. URL links could be created using standard OpenURL syntax. The use of fields requires some sort of schema for consistent tagging. There are a variety of metadata tagging approaches, with higher complexity introduced as more sophisticated searching is desired, especially across media types, networks, and terminologies.

Manipulating coded metadata within a complex environment involves creating an overall database scheme, declaring descriptive element sets for all materials and possible actions, and incorporating standard ontologies.

A Document Type Definition, or DTD, serves as a declaration of *element types* (e.g authors) and their attributes (e.g. names) for a particular class of documents. It is most often used as a framework for XML coding and searching digitized library material. http://www.w3.org/TR/REC-html40/intro/sgmltut.html#h-3.3

In this case, since the schema must reference the Darwin Core, a non-bibliographic biology schema, as well as handle the document-based MODS (bibliographic), and PREMIS (preservation) elements, we chose to define the material in a more general Resource Description Framework (RDF) schema rather than the more restrictive document-oriented DTD.

Once all the appropriate biology and bibliographic elements of the RDF scheme are identified, one must build a comprehensive RFD that describes the elements, values, and functions of the digitized material. Additional complexity in the structure will allow for additional functionality. MIT, UC San Diego, and the NSDL are among those exploring the development

and documentation of novel functionalities using enhanced metadata and smart search-and-link agents.

What are these technical requirements?

***** See Figure 2: Technical components *******

RDF stands for Resource Description Framework, and it is a framework for describing and interchanging metadata. Think of this as a schema for metadata and a syntax that makes sharing across communities possible. This metadata can describe a Resource (document) with associated Properties (e.g. author, title) and values (author names). http://www.w3.org/RDF/

In many cases the RDF will need to assimilate multiple schemas containing various aspects of resources. In our instance the schema will need to describe the bibliographic information, the preservation information, the biological information, and information about the possible actions of the materials and the system.

While RDF creates a syntax for Properties, and XML is the mark-up language that can handle these element types and values, one needs a definition of these Properties for a specific community. These vocabularies are developed by research communities. Dublin Core http://dublincore.org/ and the Library of Congress MODS http://www.loc.gov/standards/mods/ and its associated METS containers provide examples of well-known schema for bibliographic element sets.

For the *Flora Fossilis Arctica* full-text material the RDF would be created based upon the following standards, outlined in Figure 2.:

The Metadata Object Description Schema (MODS) is an XML schema intended to carry selected data from existing MARC 21 records as well as to enable the creation of original resource description records. This bibliographic-oriented schema handles data such as author, title, journal, doi, and publication date for various media types. http://www.loc.gov/standards/mods/

The Darwin Core 2 is a specification of data concepts and structure intended to support the retrieval and integration of primary data that documents the occurrence of organisms in space and time and the occurrence of organisms in biological collections. The resulting set of data element definitions are designed to support the sharing and integration of primary biodiversity data, providing a minimal set of data elements required to share the information (eg. species name, geographic coordinates, sample dates and source collection). http://darwincore.calacademy.org/

A secondary function of the Darwin Core is to enable a user to discover the contents of biological collections. Because biological collections are diverse collections, the Darwin

Core supports the search and retrieval of descriptive information in relatively simple ways. Extensions to the basic schema are being developed for expanded uses (e.g. geospatial, curatorial, paleontological, and microbial extensions to the Darwin Core 2. http:// darwincore.calacademy.org/Extensions/

For our purposes, the immediate concerns are the definitions of the elements of the paleontological data extension
http://darwincore.calacademy.org/Extensions/PaleoExtension/PaleoElementDefs
and the schema for these definitions
http://digir.net/schema/conceptual/darwin/extension/paleontology/1.0/paleontologyWithDiGIRv1.3.xsd

Our coding will use various elements such as genus-species, first discovery authority genus-species (for duplicate discoveries), geological location, epoch date, collection location, discovery date, pinnate/palmate leaf structure, and "newly discovered and mined data from our database" to create links between records and among networked servers. A challenge will be to allow for dynamically created data (e.g. newly discovered duplicate discovery records) to be saved and linked to previously existing material.

PREMIS, the preservation metadata schema will be used to store important version information. This calibration and format metadata will assist in future migrations and for the dynamic creation of alternative versions based upon the official item of record.
http://www.oclc.org/research/projects/pmwg/ Examples of information stored within the PREMIS schema are technical data (e.g. physical characteristics such as tiff or pdf format: the appropriate software versions and resolutions), file names, size, ingest pathway and certification data, risk assessments, derivative entities (embed PDFs as secondary images) and relationships between materials. Once again, a challenge will be to allow for dynamically created data to be saved and linked to previously existing material.

The RDF structure must also allow the service to define possibilities and access agents capable of performing actions based upon the properties and values of materials within the system. One would want to see options for searching, combining or limiting, and linking materials based upon their characteristics. Imagine declaring that genus-species values can be linked to discoverer authority records; hand-drawn leaf images can be linked to the digitized specimen leaves; any item can be linked to the tutorial; or any entry can be linked to locally created annotations. This enhanced service requires rules governing connections between elements across the entire network of resources. This portion of the project will require programming of context sensitive scripts.

A key design question is how much we choose to develop a relational database containing pre-created and embedded metadata information (e.g. URLs for each leaf plate image) compared to having our SFX resolver create context-sensitive dynamic links at the point of need. The requirement for hand-coding of all linkable metadata elements raises serious concerns about the

scalability and redundancy of information within our growing database. Certainly, the ability to make this database a node on a much larger network will mean we need to emphasize the resolver approach for connecting to external information resources. This is an important aspect of our integration and extension plans for the second phase of our linkage to the larger research network.

While this list of schema relates to *Flora Fossilis Arctica,* the other related tools will also require similar RDF schema and coding. Among the many elements to be considered, a few important examples include:

The Compendium Index will require coding for genus-species identification and local annotations.

The Cleared Leaf Collection will require that the specimen records contain information such as technical preparation details and collection holdings metadata.

The Manual of Leaf Architecture will require metadata about concepts such as naming standards, image descriptions, and tutorial options.

How do these schema relate to XML? XML describes and handles the content. It has tags, which may equate to both text data and other descriptive tags. XML distinguishes between *element types* (say an image) and *elements* (an individual image); however, the order in which elements appear in an XML document is significant and meaningful. This can create complexities in searching and retrieval without the overarching structure provided by either RDF or DTD definitions and declarations.
http://www.xml.com/

Let us provide a few potentially problematic raw XML issues. In the first case, multiple instances of "date" elements occur in a single entry: publication date, discovery date (or first and redundant discovery dates), and epoch date. These dates can be distinguished within the RDF or DTD. A second example is where there are multiple instances of "location" elements: specimen collection location, discovery location (or even multiple/redundant discovery locations). A similar problem may exist within the citation information, as there are at least two authors: the author of the original entry and the author of cited references. Complex datasets can require very detailed RDF or DTD definitions for the many related elements.

We intend to bring in subject specialists and computer science personnel to finalize our final schema – based upon our preliminary RDF specifications.

3. DEVELOP

There are a number of considerations that must be addressed in order to move from the Design into the Develop stage of a digitization project. While we are early in this process, we have discovered that many of these decisions are based upon logistics.

Where will you find storage space? What type of backup security will you want? What interfaces exist - and will you need to develop or customize them (for both ingest and search functions)? What technology will you need in order to provide seamless links to other tools and networks? (Are there communications standards such as OAI for harvesting metadata and OpenURL for broadcast searching?) Will you need to create API smart agents for enhanced discovery and post-processing across these networks? How will you handle non-traditional multi-media material such as images, maps, raw data sets, and citations? How do you assimilate new linking options such as citation tracking and Related Records; keywords passed to WWW search engines; and reaching in and out of researcher personal bibliographic and knowledge management tools? How do you interact with new partners to provide deep linking into free and proprietary collections (e.g. museums)? All of these questions require ongoing environmental scans of current industry practices and standards.

Of course there are the usual budget and management concerns. In some cases these exploration costs are handled as part of short-term implementation projects, but in other cases these costs are addressed as permanent reallocations of personnel and as continuing equipment support premiums.

Our two most important development efforts are (1) designing and building the RDF schema and (2) identifying the appropriate XML authoring tools for both manual and automated coding of the raw ASCII data.

Building the RDF schema will involve a number of researchers, metadata librarians, computer programmers, and public service librarians. We will need to understand the desired element descriptions, linkages, and networks in order to develop an expandable framework with extensible possibilities for linking specimen data, personal information, literature tools, and teaching materials. Our hope is to make this tool a part of our campus Sakai teaching and research portal. We are now involving a number of in-house librarians and researchers as a team to create an outline and timeline for this preliminary schema work. We imagine hiring outside consultants to review and suggest appropriate networking standards for the completion of a final RDF schema.

In terms of identifying the appropriate XML authoring tools for both manual and automated coding of the raw ASCII data, we have begun a review of the many XML authoring tools on the market and in the public domain. We may be required to utilize multiple authoring tools. One tool would be a relatively simple XML authoring application such as XMETAL http://www.xmetal.com/index.x which would be used for basic text mark-up by hired graduate students in the discipline. We would probably want a more expandable XML authoring tool in order to satisfy our desire for on-the-fly modifications and tests for developing new fields, relationships, actions, outside service links, etc. This more sophisticated mark-up would also allow us to explore creative handling of specimen images, images from the full-text volumes, the assimilation of abstracting and indexing services, and links to the leaf morphology tutorial.

Examples of these more powerful XML authoring tools, found as portions of more sophisticated XML manipulation suites, are:

XMLSpy http://www.xmlspy.com/
Altova suite http://www.altova.com/
<oXygen/> http://www.oxygenxml.com/

Other areas we intend to develop in phase 2 are the incorporation of the previously mentioned geo-referencing capabilities provided by Yale's own Reed Beamon and seamless linking to additional nodes on the rapidly developing paleontological scholarly network http://www.paleoportal.org/

4. DEPLOY

The Deploy stage also requires the consideration of many logistical issues.

How quickly can you migrate your material into your chosen platform? Can you find the staff, computer time, and expertise required to ingest the data into the chosen platform? How often will you provide backup files, and do you require 24-hour recovery service? Who will provide updates for the seamless links to other tools and networks? Will you need to update and create new smart agents for enhanced discovery and post-processing across these networks? How will you identify and adequately handle new multi-media materials, new linking options, and enhanced access to personalized researcher tools (e.g. RefWorks and EndNote) as they become available? How do you continue to interact with new partners to provide deep linking into additional collections?

Again, there are the usual budget and management concerns. These additional and recurring costs must be addressed through the permanent reallocation of both personnel and operating funds. In some cases you may be creating new and shared services with other organizations which will require budget and staffing commitments and new funding structures.

Our two most important deployment efforts are (1) creating the initial test bed for exploration by researchers, and (2) mounting the initial marked-up data on a service platform for distributed access.

We have begun creating the initial test bed of selected raw ASCII material from the *Flora Fossilis Arctica* using our local Adobe Acrobat software on a stand-alone workstation. This will allow researchers to understand the power of searching across this newly digitized material. We intend to quickly mount a marked-up portion of the same text via the web to demonstrate enhanced field searching and embedded links to outside material. We believe that even the initial advantages of searching and linking of marked-up data will generate a groundswell of excitement among the paleobotany research community, and will also serve as a powerful demonstration of the worth of such future developments to other campus researchers and librarians.

Before we choose a final interface, we are beginning to explore the advantages offered when searching the data using XML Path Language (XPath), a language for addressing parts of an XML document. http://www.w3.org/TR/xpath

In terms of platforms, as stated previously, our Fedora http://www.fedora.info/about/ platform will soon allow us to handle the lifecycle concerns of the data and also the content description with many more layers of functionality than other document-oriented repository tools. We hope to investigate the data creation, metadata development, and integration issues from within our local repository regardless of whether we mount the material with an outside service.

Our initial consideration of institutional repositories such as DSpace would have provided a workable search engine, but would have limited our ability to handle non-bibliographic material such as locally created scholar tutorials and images, and would not have addressed many issues related to the preservation and repurposing of accumulated data for future teaching and research.

Other platform options we continue to consider are mounting our data on a local campus service for researcher support (Research@Yale), or asking an existing full-text service organization such as the University of Virginia or commercial e-book companies to help us develop the infrastructure and perhaps to eventually host our material.

The eventual networked environment will appear as in Figure 3.

******** INSERT FIGURE 3: The integrated network **********

For information about Fedora capabilities see
http://xml.coverpages.org/ni2005-03-18-a.html

which lists the following articles:

- "Fedora: An Architecture for Complex Objects and their Relationships." By Carl Lagoze, Sandy Payette, Edwin Shin, and Chris Wilper (Computing and Information Science, Cornell University). Draft online. Also forthcoming in *Journal of Digital Libraries*, Special Issue on Complex Objects, Springer 2005. [PDF, cache]
- "The Fedora Project: An Open-source Digital Object Repository Management System." By Thornton Staples (Digital Library Research and Development, University of Virginia), Ross Wayland (Digital Library Research and Development, University of Virginia), and Sandra Payette (Computing and Information Science, Cornell University). In *D-Lib Magazine* Volume 9, Number 4 (April 2003).
- "Fedora 2.0: A Powerful Open-Source Solution for Digital Repositories." Summary report in *D-Lib Magazine* Volume 11, Number 3 (March 2005), 'In Brief'.