

IACRL Spring Conference, April 13-14, 2000

BUILDING AN OUTREACH DIGITAL LIBRARY COLLECTION

Timothy W. Cole

Associate Professor & Interim Mathematics Librarian

Robert S. Allen

Assoc. Professor & Agricultural, Consumer & Environmental Sciences Librarian

John G. Schmitz

Manager, Agricultural Instructional Media Laboratory

University of Illinois at Urbana-Champaign

Abstract:

University extension services are finding the World Wide Web an increasingly useful tool in their efforts to fulfill outreach mission objectives. Documents aimed at off-campus users and previously published only in paper format are now being published in HTML or PDF and made available to end-users via the Web. The challenge for the academic library is to help organize and facilitate access to these online resources as it has for equivalent paper format resources in the past. This paper describes the results and lessons learned from an effort at the University of Illinois at Urbana-Champaign (UIUC) to build a prototype outreach information system to do just this for a collection of educational resources in the field of agriculture.

Funded in part by the Illinois Council on Food and Agricultural Research, the project is a collaborative effort of the UIUC Library and the Agricultural Instructional Media Lab of the UIUC College of Agricultural, Consumer and Environmental Sciences (ACES). The prototype system created as part of this project demonstrates technologies that can be used to organize and index a diverse collection of online text resources. The collection includes online versions of pamphlets, circulars, newsletters, handbooks, and technical reports published by ACES and the UIUC Cooperative Extension Services. Links to related state and federal governmental resources augment the collection of UIUC resources. Where copyright allows selected older resources held by the ACES Library are being scanned, OCR'd, and incorporated into the testbed. Metadata is created for each item indexed. The searchable index contains a combination of metadata and complete or partial full text of the source documents. Selected resources are being converted to XML to assess the potential of that format to enhance search functionality.

Introduction

Distance education and online extension services are becoming increasingly important and vital components of the work performed and services provided by today's land-grant universities. Researchers now routinely expect to conduct a large portion of their research online; citizens served want quicker, more convenient access to the resources of their publicly funded institutions; and legislatures are demanding maximum value for the dollar. Knowledge on demand, anywhere, anytime, has become a goal for institutions of higher education.

Simultaneously end-users of information are becoming increasingly aware of online search and retrieval systems. The advent and aggressive marketing of Web search engines that purport to return the best information available in response to complex natural language queries has introduced users to key architectural concepts of digital libraries. They have come to appreciate seamless access to information maintained on systems physically located in multiple remote locations. They are beginning to learn about the features that make for more interactive and effective search interfaces. They expect extensive, bi-directional linkage between online documents on related topics. The problem is that while digital library concepts are becoming increasingly known to the average user, real, full-fledged digital analogs to brick and mortar libraries are still to be built.

It is clear that for now expectations of Web searching exceed capability by a large margin. Recent estimates indicate that less than a quarter of the available Web resources is searchable by any particular Web search engine. Results returned by Web searches are often massive in number and of low quality, requiring careful checking of myriad links to find desired resources. A related limitation is that searches do not pinpoint where in a resource the desired information resides. Key digital library infrastructure technologies – e.g., rich and fully extensible markup schemas, standards and best practices for the creation of metadata, flexible, powerful, and robust full-text search and retrieval systems – still need to be better defined, developed, tested, and implemented.

Project Overview

In the fall of 1998, the Illinois Council for Food and Agricultural Research (C-FAR) funded a project at the University of Illinois at Urbana-Champaign (UIUC) to investigate technologies required to create and make accessible a digital library collection in the domain of agriculture and agricultural sciences. The project has been a collaborative effort of the UIUC Library and the Agricultural Instructional Media (AIM) Laboratory, a unit of the UIUC College of Agricultural, Consumer and Environmental Sciences (ACES). Augmented in the fall of 1999 with an equipment grant from the Intel Corporation, the project has created a small-scale prototype outreach information system. The expertise and technologies developed as part of this work will help inform and define systems and services that will be implemented in the new and expanded ACES Library, scheduled to be open before the fall semester of 2001.

The initial task of the project was to identify items representative of library materials used in conjunction with outreach programs and appropriate for inclusion in a digital library serving such programs. Having identified the different classes of materials that were to be included in our prototype, the next step was to investigate and test technologies and processes that would support the acquisition and organization of these materials so as to facilitate access and use. Particular attention was paid to metadata generation and the design and implementation of a full-text search and retrieval system. Along the way, issues were discussed with document publishers – both to help us better understand trends in publishing of these resources and to help them better understand the ways in which the materials they produce may be used in the future. The end-result to date, a small-scale prototype of a digital library / outreach information system, serves as a testbed for experimentation with and demonstration of the technologies involved.

Testbed Collections

The process of building a collection for an online digital library differs in several respects from the process used to build a print-based collection. While much of the intellectual content is similar, items in a virtual collection are typically widely dispersed, often on servers not under the control of the library (raising significant archiving, availability, and index building concerns). Content is more dynamic and the discrete units and formats of content may vary significantly (also impacting on archiving and indexing). Intellectual property and rights issues must be handled differently (and in ways not yet entirely clear).

An academic outreach digital library collection will include not only items identifiable as books, journals, pamphlets, and reports, but also individual web pages, entire websites, courseware and other software application objects, database objects, even portals and other access points to dynamic and possibly overlapping collections (raising “appropriate copy” issues¹). Some items will have analogs in paper collections, some will not. Even traditional items such as books may be accessible and/or viewable in different ways (e.g., handbooks may be retrievable piecemeal and may include multimedia extensions or links to addenda and commentary added after initial publication).

For some time to come (at least), formats, and associated display quality and flexibility, will vary. While XML eventually may supplant the more limited HTML specification as the Web format of choice for text, systems will need to support both formats for the foreseeable future. PDF and other high-quality proprietary formats will continue to be used. Some older materials will be available only as scanned images.

To create our prototype collection we focused on examples that were relatively static, primarily textual, and available free of charge (to limit intellectual property issues). We included online versions of ACES-published handbooks and guides; State of Illinois statistical reports, Extension Service circulars, fact sheets, and newsletters; Experiment Station serials, digests and program reports; etc. Documents in HTML, XML, and PDF formats are included in the prototype. Several of the PDF items we included were created by scanning older reports and then converting using Adobe Acrobat. (Acrobat’s built-in OCR tool was used to generate unstructured full-text for such scanned documents.)

An illustrative compendium, organized by publishing institution or unit, was created to provide simple browse access to the prototype collection of online materials. This browse interface provides access not only to specific items published by UIUC organizations, but also links to websites of other organizations that provide related services and/or similar, relevant collections. **Figure 1** shows the range of publishers represented in our prototype system browse interface.

Such a virtual, distributed collection model raises numerous issues. What defines adequate archiving of materials in such a distributed collection and how is such a level of archiving assured? In the event of overlaps among the virtual collections of the linked sites, how is the user assured of getting access to the most appropriate copy? (E.g., if an item is available both under an institutionally licensed arrangement and as a pay-per-view item directly from the publisher, how does the system make sure the user is directed to the institutionally licensed copy?) How is an index to such a distributed collection built and maintained?



Figure 1 - From the Prototype Website

While the appropriate copy and long-term archiving issues raised by this model have not yet been fully addressed in our project, we have designed and implemented a search and retrieval interface to a subset of the content available from the browse interface. This allowed us to explore digital library search and retrieval issues in greater depth.

Approaches used to index the items in the collection subset ranged from full capture and reformatting of documents (e.g., the ACES-published *Illinois Agronomy Handbook* was captured and transformed into XML) to representing items in the index using metadata only. Realistic resource budgeting and the distributed and dynamic nature of digital collections argue for some compromise approach, relying on an index containing both metadata for each item and full or partial text content of items when available (and when

permissible to capture). Exposed structure of the full text can be used to enhance search and retrieval as much as allowed by the markup schemas used.

Metadata Issues

The metadata record for a digital object typically contains both descriptive and functional information. The item is described in a manner to facilitate discovery and assist in collection maintenance. Metadata may include technical information about how an item may be used, as well as information about conditions of use (e.g., license fees, who is allowed to use which parts of the object for which purposes, etc.). Metadata also will frequently contain version information, information about how the item relates to other online (and sometimes paper) objects, and/or information about the multiple ways in which an item (or portions of an item) may be accessed and viewed. Taken to an extreme, metadata records describing items in a digital library collection could routinely exceed in size the text being described – especially given that online items are frequently smaller and more discrete than print counterparts.

Such extreme investment is rarely warranted and even less often practical. In order to generate enough metadata at a low enough cost, a combination of techniques must be employed. In addition to manually generating parts of the metadata records, digital library systems also rely on existing metadata when available (e.g., MARC records describing print versions of the same content) and on auto-generation algorithms that extract metadata from the document itself. This latter approach is especially useful when dealing with collections containing well-encoded XML documents.

The definition of what is “enough” metadata remains fluid. For our project we began with the Dublin Core (DC) metadata element set². Though there remains some variation in how this schema is implemented by different projects and work to further refine the schema continues, this element set is becoming something of a lowest common denominator for many bibliographic database projects. The selection of DC elements we use and the purpose for which we use each element is shown in **Table 1**.

While a good starting point and a good schema to insure interoperability, our experience indicates a need for at least some additional granularity. In conjunction with another text-based digital library project underway at the UIUC Library³, we developed a supplemental set of metadata elements to be used in combination with DC elements. Some of these elements were designed to require specific attributes. The list of supplemental elements and each element's associated attributes is given in **Table 2**. This list is extensive and elements are intended to be used selectively – as most appropriate for a given application.

This two-part metadata schema has been implemented in our prototype system through the use of Web forms. Authorized users may add items to the searchable collection by entering available metadata on the form and submitting it to the webserver. The metadata is captured from the submitted form, and when available, the document itself is downloaded over the Web to our webserver for indexing and to allow additional metadata extraction. For HTML and XML documents additional metadata auto-generated from the source document may include information such as external links, document title and author information, author affiliation information, etc. – depending on the richness of the original markup. The metadata is then stored in an SQL-compliant database.

Table 1 -- Dublin Core Elements Used

Dublin Core Elements	Use in the Testbed
Coverage	Time or Place which constrain the context of the object.
Creator	The author of a work, both individual and corporate. May include affiliations.
Date	Any dates relevant to the object, such as paper publication date, date accepted for publication, date it was made available online, etc.
Description	Contains abstract or equivalent, if available.
Format	This will be the MIME type and/or a size in bytes of the object (if known).
Identifier	Any identifiers which can be used to uniquely refer to the object, such as DOI, internal accession numbers, PIDs, etc.
Language	A code taken from RFC 1766 which identifies the language of the item, e.g., EN for English.
Publisher	The organization name of the entity which published the article. May include address.
Relation	References to related resources. These can include external resources, such as cited or citing works, alternate forms of the current work such as PDF, HTML, and XML, or parts of the current work, such as figures or tables.
Rights	Statement of copyright ownership for the work. May include CCC.
Source	The paper publication from which this digital item was derived, including information such as Journal Title, Volume, Page Number, etc. For an item available in multiple alternate formats/types, this field may also contain reference to authoritative original.
Subject	Contains a key word, phrase, or code from a controlled vocabulary that describes the item.
Title	The title of the work as it appears in the work.
Type	The type of the object, e.g.: collection, dataset, event, image, interactive resource, model, party, physical object, place, service, software, sound, or text. May also include a subtype.
<i>Contributor</i>	<i>This is the only Dublin Core Element not currently in use in our testbed.</i>

Metadata is searched along with full-text content. Item-specific metadata records can be extracted from the database and displayed individually. Such records are encoded consistent with the Resource Description Framework (RDF) recommendation of the W3 Consortium⁴, which in turn is an XML schema. DC elements are repeated as appropriate, contained within RDF container elements (e.g., BAG, SEQ, ALT).

Indexing and Search of Full-Text

The creation of metadata as described helps to normalize and facilitate search of our digital library sample collection. We also index full text content of the items to maximum extent possible. Since this searchable index is never used to regenerate the original source document (rather the original source URL is always provided), any non-essential tagging is discarded. For XML and HTML items, we preserve document structure information useful for item discovery. Text data contained in HTML title tags, anchor tags, and appropriate meta tags are indexed both individually and as part of the document full-text. XML encoding that makes use of selected SGML tag names defined in the ISO 12083 Book and Article DTDs⁵ is also preserved and indexed. **Table 3** is a list of the ISO 12083 elements that are indexed and/or extracted automatically for inclusion in the metadata.

Table 2 -- UIUC-specific Metadata Elements & Attributes

Locally Defined Metadata Elements	Attributes	Use in the Testbed
Abstract		The text of the abstract for the item. Contained in dc:Description.
Alternate	type	Reference to an alternate form of the current work. Contained in dc:Relation.
author_info		An individual author name and affiliation, or a corporate author name. Contained in dc:Creator and external.
author_name		An individual author's name, typically 'last_name, first_name.' Contained in author_info.
Ccc		Copyright Clearance Code for the item. Contained in dc:Rights.
controlled_term	authority	A controlled subject term which describes the item. Contained in dc:Subject.
Copyright		Copyright statement for the work. Contained in dc:Rights.
Date	event	An important date in the history of the object. Contained in dc>Date.
External	type	A reference to an external document or object related to the current work. Contained in dc:Relation.
first_page		The page number in the paper source at which the current work starts. Contained in publication.
Identifier	scheme	Any type of formal identifying number or code. Can be used for identifying the current work or the journal, such as DOI or PII, or CODEN or ISSN. Contained in dc:Identifier, publication, & external.
Issue		The issue number of the journal from which the current work is derived. Is contained in publication.
item_title		The title of the work as it appears in the work. Contained in dc:Title.
Language		A code taken from RFC 1766 which identifies the language of the object, such as 'EN' for English. Is contained in dc:Language.
last_updated		The date on which this metadata record was last updated. Is contained in preparation.
Link	mime, role	URL of a related entity. Contained in external, alternate, and part.
Mime		The MIME type of the item. Contained in dc:Format.
organization_name		Contains the name of a corporate author or a publisher of the current work. May include address information. Contained in dc:Publisher and author_info.
Pagination		Complete pagination of the item as it occurs in the source; may include discontinuous ranges such as '45-47, 48, 50-53.' Contained in publication.
Part	type	Reference to an otherwise standalone object that makes up a part of the current work. Contained in dc:Relation.
Place		Name of a geographic place that describes a coverage area for the current work. Is contained in dc:Coverage.
Preparation		Contains information about the preparation of this metadata record, such as the preparer's name, email, and date of last update.
preparer_email		The email address of the person responsible for the last update of this metadata record. Contained in preparation.

Table 2 -- UIUC-specific Metadata Elements & Attributes (cont.)

Locally Defined Metadata Elements	Attributes	Use in the Testbed
preparer_name		Name of the person responsible for the last update of this metadata record. Is contained in preparation.
Publication	type	Information about a publication which is the source of the current work. May include journal name, volume, first page, etc. Contained in dc:Source and external.
publication_date		Date of publication of a related item. Contained in publication.
publication_title		The title of a publication, such as a journal title. Contained in publication.
publication_title_abbreviati on		Common abbreviations for a publication's title, such 'J. Appl. Phy.' for the 'Journal of Applied Physics.' Contained in publication.
Size		The file size of the current work in bytes. Contained in dc:Format.
Subtype		May be used in conjunction with type to more fully describe the type of an object. Is contained in dc:Type.
Time		A time period which describes the temporal coverage for the current work. Contained in dc:Coverage.
Title		The complete displayable title or description for some related work or some part of the current work. This is commonly the same text as would be seen for a citation, and may include authors, journal title, page, etc. or the text used for a figure or table caption. Contained in external and part.
Type		Type of the object such as image or text. Contained in dc:Type.
Volume		The volume number or volume identifier for the publication in which a work is published. Contained in publication.
volume_title		The official title of the individual volume in a series in which the work is published. Contained in publication.

Table 3 -- ISO 12083 Book & Article Tags Indexed

Element Name	Contains
front, pubfront	Publication front matter, including publisher information, publication date, preface, foreword, etc.
title within titlegrp	Item title information.
title within part, chapter, section, subsect#	Title of a part of the item.
fname, sname, & aff within author within authgrp	Name and affiliation of personal author of the item.
corpauth within authgrp	Name of corporate author of the item.
biblist, citation	Information about related items cited.
index, glossary	Internal index and glossary references.

Making use of explicit document content models when available provides a richer, albeit potentially less even, search index and allows for a more functional search interface.

Figure 2 shows the prototype system's current metadata-only search interface. In this interface, full-text content if present at all is assumed to be undifferentiated for searching purposes. It is only searched simultaneously with metadata. Fielded searching can only be done against available metadata fields, which should be present for all items.

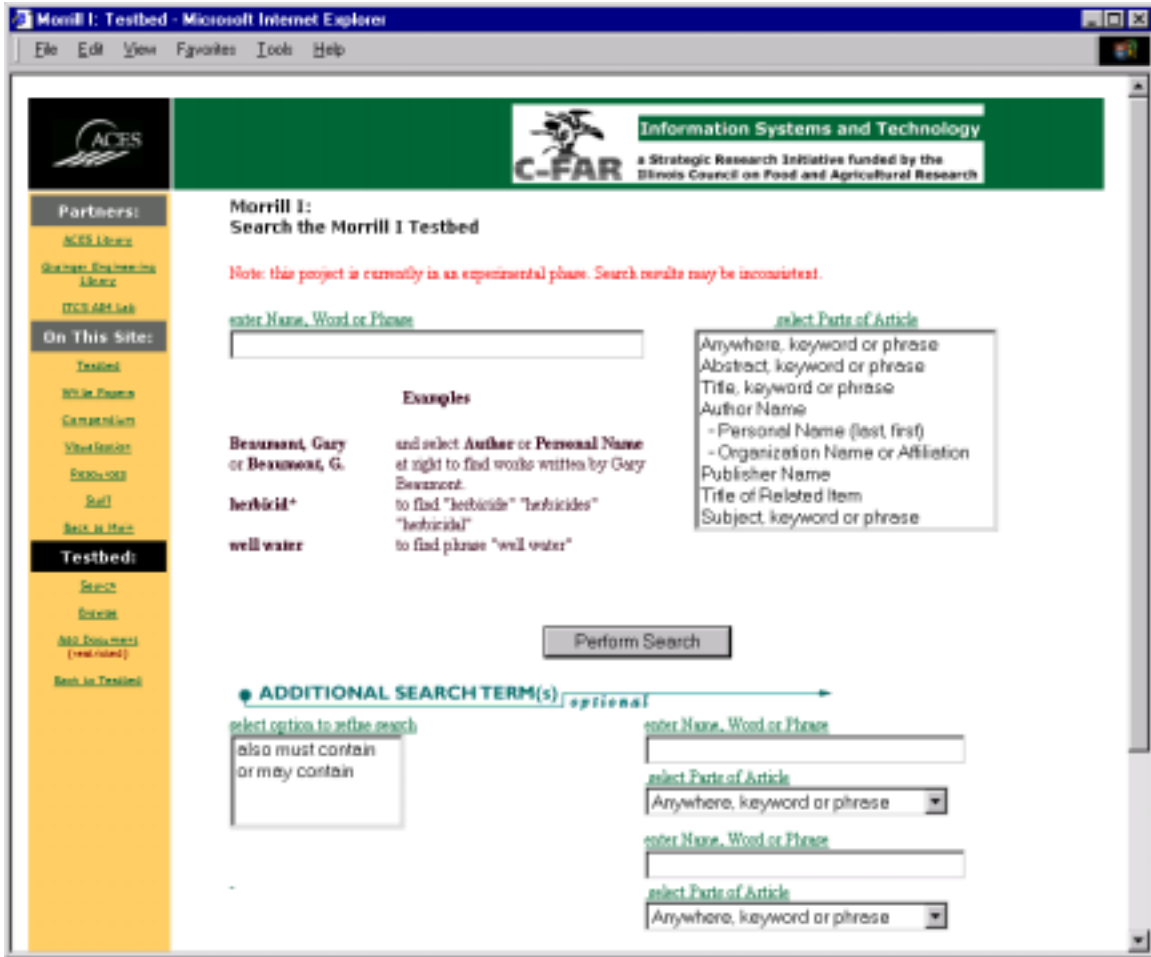


Figure 2 – Metadata Search Interface

Figure 3 shows a similar search interface that assumes the structure of the document itself can be searched. Obviously searches of specific fields in the document will not find documents for which such structural information is unavailable. (Note – the search interface shown in **Figure 3** suppresses the searching of some metadata fields searchable in the interface shown in **Figure 2**.)

The tradeoff between the two search approaches becomes a tradeoff between precision and recall. More precise searches can be done by specifying document fields to search, but some recall is sacrificed since not all the documents expose the same level of structural detail. The decision of which approach is better for a particular collection

should be based on the characteristics of the collection. How rich is the metadata? What percentage of the documents lack basic structural markup?

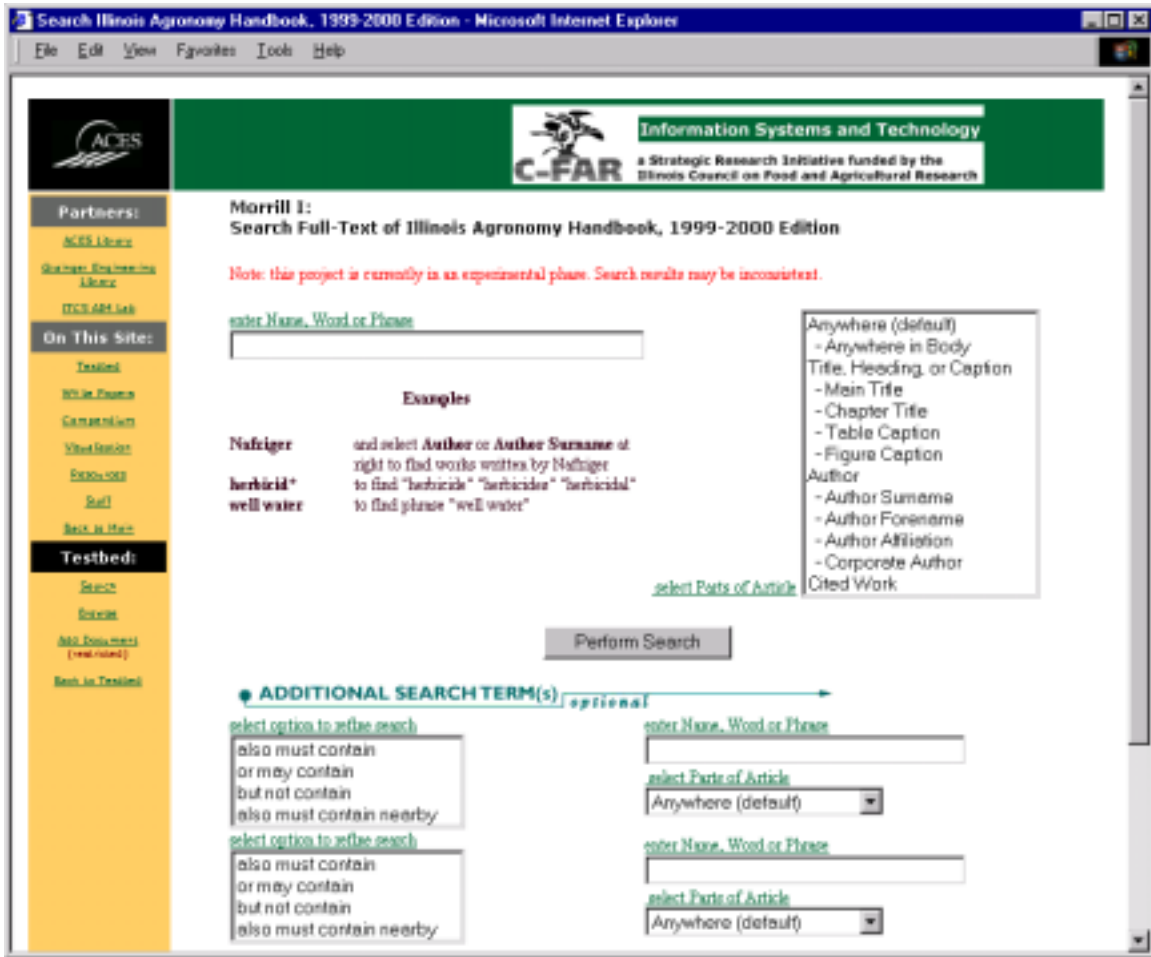


Figure 3 – Full-Text Search Interface Assuming Extensive Full-Text Structure

Discussions with document publishers indicate that there remains a wide variation in the standardization and rigor of markup approaches being used today. In general current techniques employed in this domain are not sophisticated. Some publishers of outreach materials are making an effort to markup HTML consistently enough to allow certain content structures to be inferred, at least within their own organization (e.g., authors are always shown in H3, italics). However our experience indicates that it is expensive to make use of such implicit “understood” practices in even a modest sized digital library collection. There are simply too many different publishers involved and consistency over time is too unpredictable.

More explicit mechanisms such as available with XML or through the use of the HTML class attributes should be employed, though as of yet such techniques are not being widely or consistently used in this domain. Further work to develop XML schemas and/or markup best practice standards for outreach and extension publications remains to be done. Assuming such schemas and standards are developed, then the expectation as time passes will be that it will become easier for users to find newer,

better encoded items, and harder for them to find older documents for which full-content and/or well differentiated content is not available. This is appropriate in an information system where more current publications can be considered more important, but arguably may not be appropriate in other settings. In other situations, additional retrospective work may be required to insure all items in the collection are equally represented in the index. Alternatively, the decision may be to index to the lowest common denominator – i.e., to build search indexes and interfaces based on the minimum markup standards represented in the collection.

Conclusion

The ease with which information can be mounted on the World Wide Web belies the challenge and expense of organizing that information and making it easy to find. While the project described here demonstrates the potential of the technologies available, it also illustrates how time-consuming and labor-intensive the process of constructing a digital library is. To deal with the diversity and amount of relevant online information being generated every day, a balance must be struck between the labor-intensive work of manually creating metadata and the equally labor-intensive work of using high-fidelity mark-up schemas to adequately expose document structure. Achieving this balance will require that university libraries work even more closely than they have in the past with publishers of information, both within the institution and outside of it. It also will require that the agricultural extension and outreach community work diligently to develop standards and best practices for both markup and metadata creation, and that we continue to study the information-seeking behaviors of our end-users so as to better inform the development of those standards and best practices. Small prototype testbed projects such as the one described will help us to better investigate both the technologies and the use of those technologies by end-users in practical contexts.

The good news is that we have an idea technically of how to meet the challenges ahead. The bad news is that it's going to be hard work and it's going to require an extensive investment of time and resources.

¹ For further discussion about the appropriate copy issue see: **Caplan, Priscilla and Dale Flecker (1999)**. *CHOOSING THE APPROPRIATE COPY: Report of a discussion of options for selecting among multiples copies of an electronic journal article* [online]. Available: <http://www.niso.org/DLFarch.html> [10 March 2000].

² **The Dublin Core Metadata Initiative (1998)**. *The Dublin Core: A Simple Content Description Model for Electronic Resources* [online]. Available: <http://purl.oclc.org/dc/> [29 February 2000].

³ **Schatz, B., et al. (1999)**. “Federated Search of Scientific Literature: A Retrospective on the Illinois Digital Library Project.” *IEEE Computer*, 32(2), 51-59. Also: **Arms, William Y., et al. (1999)**. “The D-Lib Test Suite: Testbeds for Digital Libraries Research,” *D-Lib Magazine*, 5(2) [online]. Available: <http://www.dlib.org/dlib/february99/arms/02overview.html> [13 March 2000].

⁴ **W3C: World Wide Web Consortium (1999)**. *Resource Description Framework (RDF) Model and Syntax Specification: W3C Recommendation 22 February 1999* [online]. Available: <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/> [29 February 2000].

⁵ **Kennedy, Dianne (1999)**. *ISO 12083 Information* [online]. Available: <http://www.xmlxperts.com/12083.htm> [13 March 2000].