Int J Digit Libr (2005) 5: 266–274 / Digital Object Identifier (DOI) 10.1007/s00799-004-0094-8

Digital library information-technology infrastructures

Yannis Ioannidis[1], **David Maier**[2], **Serge Abiteboul**[3], **Peter Buneman**[4], **Susan Davidson**[5], **Edward Fox**[6], **Alon Halevy**[7], **Craig Knoblock**[8], **Fausto Rabitti**[9], **Hans Schek**[10], **Gerhard Weikum**[11]

[1] University of Athens, Dept. of Informatics and Telecommunications, Athens, Greece; e-mail: yannis@di.uoa.gr
[2] Oregon Health & Science University – OGI School of Science & Eng., Dept. of Computer Science and Engineering, Beaverton, OR, USA; e-mail: maier@cse.ogi.edu
[3] INRIA-Futurs and LRI, Parc Club Orsay-University, Orsay Cedex, France; e-mail: serge.abiteboul@inria.fr
[4] University of Edinburgh, School of Informatics, Edinburgh, UK; e-mail: peter@cis.upenn.edu
[5] University of Pennsylvania, Dept. of Computer and Information Science, Philadelphia, PA, USA; e-mail: susan@cis.upenn.edu
[6] Virginia Polytechnic Inst. & State University, Dept. of Computer Science, Blacksburg, VA, USA; e-mail: fox@vt.edu
[7] University of Washington, Dept. of Computer Science and Engineering, Seattle, WA, USA; e-mail: alon@cs.washington.edu
[8] Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA; e-mail: knoblock@isi.edu
[9] ISTI – CNR, Networked Multimedia Information Systems, Pisa, Italy; e-mail: Fausto.Rabitti@isti.cnr.it
[10] ETH Zurich, Computer Science Dept., Zurich, Switzerland; e-mail: schek@inf.ethz.ch
[11] Max-Planck Institute of Computer Science, Stuhlsatzenhausweg 85, Saarbruecken, Germany
e-mail: weikum@mpi-sb.mpg.de

Published online: 22 July 2005 – © Springer-Verlag 2005

Abstract. This paper charts a research agenda on systems-oriented issues in digital libraries. It focuses on the most central and generic system issues, including system architecture, user-level functionality, and the overall operational environment. With respect to user-level functionality, in particular, it abstracts the overall information lifecycle in digital libraries to five major stages and identifies key research problems that require solution in each stage. Finally, it recommends an explicit set of activities that would help achieve the research goals outlined and identifies several dimensions along which progress of the digital library field can be evaluated.

Keywords: Digital libraries – System infrastructure – System architecture – Digital library information management – Digital library research agenda

1 Introduction

Many digital library (DL) systems have been developed Over the past several years. Each system is typically built from scratch and develops its own techniques, focusing on a specific type of information or services and addressing the needs of a specific community or domain. However, the future of digital libraries goes beyond what these past efforts indicate individually. Furthermore, the traditional management of plain text will make way for that of enriched documents with embedded knowledge.

Digital libraries can become the universal knowledge repositories and communication conduits of the future, a common vehicle by which everyone will access, discuss, evaluate, and enhance information of all forms.

For the full potential to be realized, DL development must move *from an art to a science*. Advanced information-technology infrastructures of DLs should be created that will lead to the following:

1. Unifying and comprehensive theories and frameworks across the lifecycle of DL information.
2. Interoperable multimodal and multilingual services and integrated content management ranging from the personal to the global, for the specialist and the general population.

It is time for *generic* DL technology to be developed and incorporated into industrial-strength digital library management systems (DLMSs), offering advanced functionality through reliable and extensible services.

The remainder of this paper identifies the key ingredients of a future DL infrastructure and the research required to provide the corresponding functionality. Section 1 provides two scenarios from different domains that would benefit from any major advances in the field. Section 2 gives a general framework of the field and then analyzes in detail research developments needed to bring DL technology to the next level. Section 3 lists a range of activities to achieve the research goals mentioned above. Section 4 charts the expected evolution of digital library technology across various dimensions and concludes the paper.

This paper follows from the report [14] of a DELOS-NSF working group that met three times between June and November 2002.

2 Scenarios

This section presents two scenarios that motivate the research issues in the following discussion. The first focuses on cultural materials, while the second is from the biomedical domain.

2.1 Cultural heritage preservation

Alberto is a graduate student investigating comparative cultural heritage preservation styles throughout Europe. His particular interest is cultural artifacts. He is building a knowledge base that includes tables of holdings and acquisition rates for the many museums, government properties, and places of worship across the continent. He uses automatic feeds of history and tourism magazine articles on new exhibitions and collections, as well as auction catalogs, which arrive as multimedia documents and are semiautomatically analyzed and cataloged. He is integrating his knowledge base with those of a consortium of researchers in three other countries, who are working on regional dialects, 3D digitization of threatened cultural sites, and tracking oral history projects.

After annotating some newly acquired articles at his office, Alberto goes to his first lecture, on historic preservation law. A multimedia version of class activities arrives in real time on his laptop. Part way through he asks his computer to compare some cases brought under these laws (using links from the lecture-note feed) with artifacts and sites in his knowledge base. He uses the matches to annotate the lecture notes, publishes those annotations to the rest of the course, and later uses them to illustrate a point during class discussion.

After lunch, Alberto leads a discussion section for a course on ancient European history. Today's topic is Etruscan civilization. He has drawn on the Perseus Digital Library [5] for Etruscan inscriptions to illustrate the Etruscan language and for images of Etruscan pottery. His own annotations connect translations of the inscriptions with passages on Etruscan religion in the class's reading assignment. Other annotations link images of Etruscan and Greek vases of similar styles. Small groups of students use these links to debate which culture was borrowing from which. They make notes, some shared with the section, some with Alberto, and some with the whole class.

Later, Alberto works on the DL maintained jointly by the research consortium. It was easy to generate the joint library from their individual collections as they were all implemented on digital library management systems (DLMSs) conforming to Open Digital Library protocols [13]. Alberto set up the joint library using a declarative specification tool that let him indicate users, rules for constructing a joint catalog, workflow for reviewing and releasing resources, and additional components for rating, distributed annotation, and recommendation. He examines a mapping from an ontology of languages and dialects to an ontology of cultural groups that a collaborator built using ontology-matching software. Once the mapping is complete, it will be used to automatically generate public virtual tours of sites captured in 3D, with artifacts, documents, and avatars relating oral histories from the appropriate cultures placed within the scene.

That evening, Alberto is working on a paper relating the relative growth rates of private and public collections in different countries with regulations on foreign sales of cultural artifacts. He and his coauthors are using an array of tools to extract and index information from exhibition catalogs, museum reports, auction results, and their own and other DLs. They can then query and analyze this information to test hypotheses about the effect of particular policies on acquisition rates.

Many parts of this scenario are not supported by current technology. Support for automatic acquisition of resources is limited and imprecise. Real-time construction of collections via capture of lectures and conferences is expensive, where it exists at all. Personal annotation of resources is seldom supported within individual DLs, much less across two or more. Fine-grained control of sharing of resources is cumbersome to configure. It is not easy for a nonspecialist to create a DL; declarative construction lies in the future. Tools for ontology matching are still in their infancy. Information extraction from unstructured and semistructured sources often requires intensive manual effort.

2.2 Biomedical research

Alicia, a biologist, is studying neuroblastoma, a malignant cancer that develops in the nervous systems of young children and may be linked to rearrangements on human chromosome 1. Alicia built a small data repository with her own experimental findings but wishes to view and share this information in a larger context. In particular, she wants to:

a. Access data repositories of protein-product and metabolic-pathway information to understand how the rearranged sequence translates into modified gene regulation;
b. Find analogous diseases in other species;
c. Scour the biomedical literature for information on neuroblastoma;
d. View gene-expression data correlated to a 3D brain atlas showing neurophysiological development over time;
e. Augment her personal repository with this external information;
f. Record how and from where she obtained the information and from where the source databases obtained their information; and

g. Place annotations on her database and the source databases that her colleagues can view.

Several difficulties confront Alicia. Although some standard data repositories exist (e.g., metabolic-pathway information), Alicia may not know where they are. Furthermore, dozens of more specialized sites are relevant to her interests. Each week brings new sources for particular species whose genomic structures have been deciphered. Even though searching a source this week turned up nothing, a search next week may turn up new information. Furthermore, the sources Alicia consults use different ontologies, vocabularies, and scales, complicating her data-integration problem. In addition, integration must be performed across multimodal data, with time and spatial dimensions, and appropriate display techniques provided. Augmenting her personal repository without locally replicating everything is problematic because the information sources lack a reliable, multigranularity addressing scheme for referencing their elements. Recording sources and methods used to obtain information is a largely manual and error-prone task, and few sources detail where their information was obtained. While Alicia could modify her own data repository to hold annotations, she cannot do so on source databases she does not control. Finally, all her warehoused and annotated data must be kept current in the face of updates to her own findings and revisions to the information sources she consults.

3 Research issues

These scenarios have requirements for sophisticated functionality far beyond anything supported in current digital library systems. Underlying such functionality is a host of technical problems in information management calling for advanced research. The topic of information-technology infrastructures is broad. We focus on the most central and generic system issues, including system architecture, user-level functionality, and the overall operational environment.

3.1 System architecture

First-generation DL systems were largely "from scratch," sometimes incorporating existing components (such as indexing engines), but with limited modularity. Such one-off solutions were reasonable for examining issues and usage, such as handling different kinds of content and organizations, deciding what functionality a DL should possess, and determining which interfaces users find most appropriate. Performance, maintainability, robustness, and scalability were secondary concerns while the basic nature of DLs was still being determined.

Not surprisingly, first-generation implementations are not necessarily reusable, easy to install, customize, and configure, or amenable to distribution. However, any DLMS intended for widespread use needs these attributes. Having gained basic knowledge on organization, function, and interaction for DLs, it is time for second-generation DL implementations based on DLMSs with a more deliberate and performant architecture. Such an architecture should establish externally visible capabilities and also a standard set of internal interfaces and protocols. In this way, a collection of interoperating components from different sources can be assembled into DL implementations meeting a variety of needs. We have seen encouraging steps in this direction, in particular the Open Archives Initiative (OAI) (`http://www.openarchives.org`) [21] for the metadata provisioning-gathering interface and the Open Digital Library framework (ODL) (`http://oai.dlib.vt.edu/odl/`) [13], which defines component interfaces for functions such as searching, metadata union, and document-reviewing workflow.

A component-based approach to DLMS architecture allows tailoring of individual DLs through component selection and replacement. Distributed implementations of DLs should be easier to obtain since components can run as independent processes on different machines. Further, components are a more logical unit of reuse than a single, monolithic implementation. Components also provide an alternative pathway to DL federation and scalability. Some current approaches to federating DLs wrap an entire DL (say through a request translator), which can be awkward and brittle in the face of user-interface changes. Federating DLs in this manner views a DL as a collection of all its functions, then must split out particular capabilities in the wrapper to combine with equivalent functionalities from other DLs. It makes more sense to us to federate DLs through peering of components with the same functionality, for example, the metadata manager bridging directly to another metadata manager. We also posit that component-based architectures are more scalable than monolithic architectures. The component corresponding to the particular performance challenge can be upgraded, replicated, or distributed, with minimal modification elsewhere in the system.

A component-based approach also provides help with heterogeneity issues. Heterogeneity can be in capabilities, content types, and search mechanisms. A DL without the need for a particular capability (e.g., a recommender service) can omit components for that service. Similarly, a DL requiring a specialized capability (e.g., metadata extraction for a particular class of documents) need only customize one or two components. Further research in component architectures should address combined search and ranking.

Once a rich base of components is available, the next challenge is packaging and deployment: tools for DLMS construction using these components that provide for specification, selection, installation, configuration, and operation. One should be able to craft solutions that are matched in costs to the DL capabilities needed. It should be possible to put up a basic DL quickly and inexpensively on a single computer, but also to construct a DL with custom functionality, high availability, and a replicated, distributed architecture in the same framework.

Research in DLMS architectures must track developments in other areas. The component-based approach dovetails with Web services [1, 12]. DL components can function as Web services in larger information systems. Conversely, a DLMS can incorporate external Web services. In particular, a DLMS should permit a Web service as a content source. Developments in the Semantic Web [4, 10] are also relevant. Semantically tagged content has great utility for DLs, allowing more targeted searching, easier determination of documents' relevance, and better information discovery. Finally, the effect of Computational Grids [9] on DL architectures needs to be explored. Is massive computational power useful in DLs, such as for advanced feature extraction and classification techniques? Conversely, can DL technology catalog and search the vast data resources resident on a grid?

3.2 User-level functionality

Relative to user-level functionality, we abstract the overall information life cycle in a DLMS to the following major stages:

A. The user interacts with the DLMS to express some information need;
B. The DLMS processes the user input and passes it to the underlying storage systems;
C. The requested information is accessed and retrieved;
D. The information collected is transformed, cleaned, integrated, ranked, formatted appropriately, and presented to the user;
E. The user selects, organizes, and enriches the information collected (possibly from multiple requests) for the task at hand.

The main questions are grouped below based on the five stages above. Clearly, some topics raise problems in multiple stages.

A. User interaction

The information life cycle begins with a user requesting information believed to reside in the underlying DL. Alberto has traversed this stage several times with the Perseus Digital Library, preparing for his discussion section and searching his personal collection to identify the effect of policies on acquisition rates. User-system interaction should be as simple as possible, to serve a large audience, but should permit the expression of sophisticated information needs when required. We group the underlying technical problems to meet this goal into three broad areas.

- **Languages and interfaces**: The term "language" here includes all forms of communication with a DL, textual, visual, or other. Languages developed in other fields (e.g., databases) are relatively narrow and cannot capture the full range of requests to DLs. For example, current language technology does not allow Alicia to easily express a request for gene regulation information associated with DNA sequences present in scientific articles discussing neuroblastoma, as it combines keyword-style text retrieval with associated selections from structured repositories. New languages must be expressive enough for users to pose their most sophisticated needs succinctly. Although formal textual languages may be the first results, the development of visual or even natural-looking languages with equivalent expressive power will enhance ease of use. Furthermore, given the increasing importance of semistructured data in capturing digital documents, special attention should be paid to semistructured data models, such as XML, and corresponding query mechanisms.
- **Paradigms**: The traditional approach to DL interaction draws on conventions of related fields, such as databases and information retrieval: A user poses an information request and an answer is generated and returned. In most cases, this process is isolated from other similar requests. For the next generation of DLMSs, other interaction paradigms must be supported to make the DL experience more productive and appealing. Personalized interaction and access is critical given the diversity of backgrounds, needs, and preferences of DL users. For example, the Perseus Digital Library should provide different answers for the same request about the Etruscans depending on whether it is made by Alberto (a researcher on the topic), the members of his discussion section (undergraduate students in the area), or Alicia (an interested member of the pubic in this context). User profiles should affect system behavior, including the interaction context of the user (session history), the location of the user, the time, etc. A promising direction for personalization is ontology-based interactions and request interpretation, where ontologies capture personal or group frameworks within which requests are posed. Another paradigm shift is needed towards relaxing strictness and preciseness levels of interactions. Formulation of vague queries and uncertain representations of documents should be the rule, as searching in a DL is more an exploratory than interrogatory task. Intimately related to the direction above is providing relevance feedback or other interaction mechanisms to support implicit query reformulation, e.g., relaxation. Another significant paradigm for user-DL interaction is query subscription, where data transfer is initiated by the DLMS itself based on standing requests that the user registers.
- **Tools**: Independent of the expressiveness of interaction languages or the richness of interaction paradigms, the supporting tools are the main

determinants of effective user-system interactions. Despite progress in user interfaces, interaction with information is poor in general. We seek easy-to-use tools that have low cost of entry for new users, provide instant gratification for all users, and allow incremental exploration of the available information space. Without this capability, interacting with a DL cannot compete with actual library visits and will serve the initiated few.

B. Analysis

Analyzing and evaluating an information request is the heart of the information life cycle. The sophistication and effectiveness of this stage affects the value the user obtains from interacting with the system. Furthermore, some of the interaction paradigms above generate additional challenges in processing complicated, context-sensitive sessions. DLMSs must address the problems outlined below.

- **Similarity**: Since most data in DLs are not structured and lack precision, what constitutes a match to a request often differs from other fields. We require new definitions of similarity and techniques for evaluating them. For example, the ontology-matching software used by Alberto's collaborator must identify similar concepts in multiple ontologies using more than syntactic matching. For structured items, similarity can be based on high-dimensional feature representations of the components, aggregating differences on individual features into an overall similarity measure. When Alicia is seeking analogs to neuroblastoma in other species, diseases could be represented by their characteristics, and similarity measured on that basis. Defining similarity for nonnumeric information is challenging; systems should include subjective relevance and similarity, providing a vehicle for personalized behavior as well. Due to its growing importance, XML document matching deserves special mention.
- **Request evaluation**: New forms of similarity require new evaluation techniques, including approximate evaluation techniques. For example, being distance-based, Alicia's search for neuroblastoma analogs will be optimized and executed differently than if exact match of features is required. Multimedia and continuous-media searches with similarity-based comparisons on the content of the media itself (not just on the metadata) need special attention. An implication of similarity-based evaluation is that there is no unique, provably correct answer to a user's request. Combining distinct search mechanisms could give increased effectiveness and quality for the final information provided. Such diversity may lie in algorithmic aspects of search, such as the similarity metrics used, but may involve more fundamental aspects such as the user-interaction style itself (database-style queries, hypertextlike browsing, annotation-based correlation) or the form of information (e.g., correlating geospatial and semistructured data). DLs will often reside in a distributed environment, possibly served in part by external, autonomous information providers under service agreements, giving rise to new forms of search that are far from understood. Examples include cooperative search, where the participating nodes cooperate to identify the optimal answer to a request; competitive search, where they work to increase their own benefit as well; and self-organizing search, where the participants may be rated, weighted, etc. A user request may have multiple interpretations that depend on any part of its context. Thus, context-based search is another area for further investigation, as is ontology-based interpretation of the semantic context of a request. (Just capturing the information and process context of a request is a major challenge.)
- **Process management**: Besides passive information, active processes play a major role in DLs and require special forms of management. Processes appear in DLs at two different levels: First, processes may actually be stored in a DL, e.g., Alice may have several interdependent simulations for analyzing different behaviors of proteins; second, user interactions with a DL form processes, e.g., the actual steps that Alice takes to identify and annotate information. Keeping track of how information is derived through processes, monitoring processes as they unfold, process execution representation for reproducibility, and process optimization and tuning are all emerging as issues in DLMSs.
- **Information and process quality**: Given the expected approximate nature of information analysis within DLs, the quality of the resulting information and of the process to identify it needs measures of the success attained. One may imagine complex metrics to fill that role: completeness, accuracy, resolution and uncertainty, timeliness, and freshness for the resulting information; cost and response time for the process generating that information. Alicia wants the most up-to-date documents on neuroblastoma, while quickly finding the information on Etruscans so he can annotate it before class is Alberto's concern. Other metrics for specific situations must integrate into the general approach. Some aspects of quality cannot be readily quantified. For example, trust plays a major role in human interactions and is equally important in human-DL interactions. It is unclear how to evaluate alternative sources based on trust in their creators or curators. Another example is the existence of explanations (e.g., provenance). Measuring the quality and usefulness of explanations is critical but by no means straightforward.

C. Data storage

The content of DLs must be manipulated and prepared to best serve the other stages of the information life cycle. The diversity of content, and of operations upon it, generates a rich set of data-storage problems. We divide these problems into the three categories below.

- **Information space organization**: At the highest level, the issue is preparing information so that it is useful or interesting to the users. Injected information is analyzed to generate additional information that can serve people's needs. Knowledge discovery and data mining play roles here, for classification, clustering, ranking, and other forms of automatic or manual learning. The requirements are more challenging in a DL environment than in a business environment because of the plethora of data types and the different types of data mining needed. DL content is dynamic and needs support for versioning and archiving. Finally, recall that DLs are envisioned as meeting points for human discussions and collaborations, so content organization should facilitate such activity.
- **Feature space organization**: Indexing data items on individual features is key to efficient access. This old problem takes on a new life in a DL environment. The great variety of data in DLs demands new kinds of indexes, e.g., indexes for manuscripts (for their texture and visual layout as well as content), indexes on structure (such as for proteins), indexes for information in multiple media. Since much analysis and evaluation in DLs is approximate, indexes must support similarity queries and approximation of data features. Indexing of text remains a great challenge in all but its simplest forms. Stemming, indexing all meaningful parts of composite words, abbreviations, multilingual indexing, and indexing by text context are issues that need particular attention. Finally, the importance of some data formats for DLs, such as XML, makes indexing techniques for those formats especially significant.
- **Data formats**: Other issues around specific data formats require further study. A key problem is cost management for XML requests: The complexity of the format makes estimating the cost or result size of XML queries challenging. Metadata storage and retrieval is always an important problem and requires new approaches for the rich content of DLs.

D. Information integration and derivation

Locating and accessing information is only part of satisfying a user's request. Putting that information in the form most useful to the user may involve considerable postprocessing. Information from multiple sources may need to be combined into a single ranked list, have further quality metrics applied, or be summarized. When Alicia searches for material on neuroblastoma, the results might include published papers, data sets, and Web pages. She may want the combined results ranked based on the coverage of the protein products and pathways she believes are involved in the disease. Formatting and rendering may be required to make the retrieved information more easily understood. For example, Alicia may want information elements connected to expression of genes at particular anatomical locations associated with a 3D brain model. IT infrastructure for DLs should support the required result manipulation and presentation.

- **Result manipulation**: A particular DL request might consult a dozen sources, each with separate rankings. Ideally, these separate lists should be combined into a single ranked list, or otherwise organized, but the rankings are not necessarily comparable. Deriving correspondence between items coming from heterogeneous sources may require a semantic understanding of the underlying data. Thus, joint ranking may require retrieving the actual content in order to rerank it locally. Furthermore, joint rankings could depend on multiple criteria, such as match to keywords, freshness, and novelty relative to information already retrieved. Users will need simple ways to express preferences over multiple criteria. Most information-searching techniques are based on relevance, but quality assessment will become increasingly important. Quality assessment will also rely on explicit annotations, document context (Is the source curated? Is the document linked from a site previously judged reliable?), and agreement with retrieved information already judged accurate. Sometimes only small portions of retrieved documents or data are of interest. While there has been significant work on document summarization, *task-specific* summarization needs further work. When Alberto searches for museum reports and auction catalogs, he may be interested in certain categories of artifacts, or ones that are changing hands between the public and private sector.
- **Rendering and presentation**: The Web-search-engine approach of outputting results as a listing of hundreds or thousands of retrieved items is not conducive to complicated information tasks. Visual or alternative textual displays of large retrieval sets could enhance user performance. We mentioned Alicia visualizing information relative to a graphical model of the brain. Alternatively, she might want to see documents presented according to her local hierarchy of gene function. This latter instance exemplifies "personalized presentation": presenting data based on personal factors, such as background of the user, relation to local information, or interaction history. A user might not want to see all results at the same level; it may be more useful to organize the

result set into a navigable presentation, with its own internal page and link structure.

- **Information-space visualization**: The initial goal of a user researching a domain might not be finding particular documents, but rather getting the "lay of the land." He or she may want a view of the information space as a whole. Some DLs already aid the user in this regard, with geographic or temporal maps of holdings. Visualizing the information space covered by interoperating DLs is more challenging. When the user starts seeking particular documents, it may be important to situate them in the context of the larger information space. A simple example: Web search engines return individual pages, but correct interpretation of page content may require additional pages from the site that lead to the page of interest.

E. Information enrichment

Once information has been located, extracted and rendered from a DL, the story is not over. Interaction with DLs is generally in support of some task. That task can require further analysis, arrangement, and enhancement of the information obtained. Recall how Alberto annotated and linked the information items he found to make them more coherent and call out connections for his discussion section. We believe most tasks that employ a DL will involve enrichment of the information obtained. The IT infrastructure of DLs should provide capabilities to meet these needs, such as maintenance and propagation of pedigrees; annotation creation, management, and sharing; and restructuring and combining information from multiple queries. We describe these capabilities below.

- **Provenance, pedigree, and citation**: We distinguish *provenance* as the ultimate source of an information item plus the intermediate points by which it reached its current location and *pedigree* as how an information item was derived: queries, processing steps, and inputs used to determine it. Both are important in judging the quality and applicability of information for a given use and for determining when changes at sources require revising derived information. Provenance, pedigree, and accurate citation depend on stable, fine-granularity addressing in DLs, which few DLs provide currently. While Digital Object Identifiers (DOIs) [18] and other proposals move in this direction, they generally lack citation at a subdocument level. Even less support exists for uniformly citing subportions of nontextual media, such as images, video, and data sets.
- **Annotation**: Annotation of content was an aspect of intellectual activity well before the digital age. There have been commentaries and interposed notes almost as long as there have been written languages. Annotations serve to select, explain, question, augment, and connect information elements. Fine-granularity addressing supports "stand-off" or "superimposed" annotations when it is inappropriate to modify the underlying content. DLMSs should provide for creating annotations, by manual means and intelligent processing of content. They should support storage, selective sharing, and configurable presentation of annotations. A challenge here is annotations that span DLs (such as point-to-point links): How can one DL become aware of annotations connecting to it that are managed by a different DL? Finally, other DLMS services, such as querying, should maintain annotations with an information element when it is extracted or reformatted.
- **"Recombinant information"**: The form in which information is delivered is seldom the most apt for a user's task, even with integration and presentation capabilities described previously. A user may obtain information through numerous separate requests from several DLs, so the information of interest is not processed all at once. DLMSs, or closely integrated tools, should help users interactively segment, combine, restructure, group, and organize the results of their queries. A particular application is producing a new document targeted to a particular task from excerpts of existing documents. For example, Alberto may want to construct "virtual auction catalogs" for particular classes of artifacts from entries in real catalogs.

3.3 Supporting environment

Progress on many issues in user-level functionality depends on advances in the infrastructure and general environment for DLMSs. In particular, metadata and ontologies are a critical underpinning for several of the capabilities described. Metadata is involved with query of nontextual information, context-based search, process management, feature-space organization, and information enrichment. Ontologies have a role in natural-language interaction, query formulation and interpretation, new similarity measures, classification, schema understanding, and information integration. DLMSs will need metadata and ontology management, such as construction, extension, storage, evolution, and versioning. While these functions have many uses in other domains, DLs provide challenging applications of them.

One resource of particular value for DLs is a corpus of information models, database schemas, and ontologies for online information sources. While harder to find and extract than simple document contents, such a collection has value in several contexts. It provides a measure of the diversity of models and organizations for online information. Does the diversity of modeling structures vary markedly across intellectual domains? Do resources related to art show more structural diversity than those for bioinformatics? Do such differences indicate which approaches are better suited for information integration in those domains? Second, it is a basis for investigating feature extraction for clustering or classifying models. IR-type techniques applied to the feature space could judge similarity of a newly found model to existing ones, helping identify the topic of a resource during information discovery. Finally, the corpus can support design aids for conceptual models for new DLs.

4 Recommendations

We have presented research issues in systems architecture, user-level functionality, and supporting environment. We list activities to target these issues below.

Principles of DLs: Much work to date is on the “art” of DLs – developing techniques for content capture and digitization, construction of DLs for particular purposes, learning good practices for operation and curation, and attempting integration of multiple DLs. Significant future activity should be devoted to moving the field of DLs towards a science by studying and articulating its basic principles. Such activity encompasses both formal modeling of DLs, their contents, their use, and their communities, as well as developing normative frameworks and protocols for DL architecture, infrastructure, and interoperation.

Community resources and repositories: The “community” here is DL researchers and developers. We see value in collecting information across a range of information services so that researchers have wide enough coverage of the domain to produce valid conclusions and developers can have a sense of what has worked or not previously. One resource is a compendium of case studies of DL projects, documenting requirements, development methods, construction and operation costs, features, holdings, economic models, and evaluation. A second resource is a corpus of information structures, metadata schemas, and data samples to support statistical approaches similar to those used in IR.

Technology development: Referring back to our scenarios, a wide range of capabilities useful in the context of DLs do not yet exist or exist only in a limited form. Much of the work on user-level functionality requires design and development of new technology. Activities in this area should span all stages of the information life cycle.

Technology application, testing, and dissemination: While new technologies must be useful for a wide range of DLs, they should nevertheless be deployed and evaluated on particular applications. Activities here should take place in conjunction with specific DLs, either extant or under development (rather than DLs created purely to test the technology). DL initiatives to target are the National Science, Technology, Engineering and Mathematics Education Digital Library (NSDL) [16] and EC-funded cultural projects, e.g., COLLATE [15], ECHO [2], CYCLADES [11]. While large DLs may appear to have the greatest technological challenges, delivering solutions for small systems with proportionately smaller costs also constitutes a significant problem.

5 Evolution of digital library technology and conclusions

Looking forward, the progress of the DL field can be evaluated along several dimensions.

Architectural dimension. Here we see increasing capabilities and dynamicity as more sophisticated system and network architectures develop. Some points along this dimension:

- Single systems: Early DLs were standalone systems, generally serving a single purpose.
- Homogeneous distributed systems: These replicate multiple instances of the same DL implementation, for distribution and scaling, but still tend to serve a single purpose. An example is Dienst [17].
- Heterogeneous distributed systems: These relax the requirement that all sites must be running identical software or providing the same capabilities. Instead, standardized service definitions allow interoperation among distinct DL implementations. Such systems tend to be directed at a given category of purposes. Individual sites generally participate in a single federated system. An example is the Networked Digital Library of Theses and Dissertations (NDLTD) [8].
- Dynamic virtual DLs: Ultimately, DLs will enter into multiple federations, which emerge dynamically in response to particular needs and give users and communities unified access to their combined resources. An example is Hyperdatabases [20].

Interoperation dimension. Along this dimension we see an increasing number of aspects in which DLs can interoperate:

- Search and retrieval: Search requests are forwarded to other DLs, and responses are combined and returned to the requester. Retrieval requests are routed to the appropriate storage server. The Dienst Index and Repository Services are examples.
- Metadata: DLs interact to exchange metadata, allowing local cataloging and indexing. Harvest [3] and the Open Archives Initiative [21] are examples.
- Security and authorization: DLs protect privacy and propagate privileges for client requests that span multiple sites; DLs also establish mutual trust relationships.
- Quality assessment: DLs exchange reviews, ratings, and evaluations of materials.
- Notification: DLs distribute user-interest profiles and provide efficient distributed notification.

Information dimension. This dimension delineates the sophistication with which individual DLs reason about the information they hold.

- Data: Holdings are regarded as simply text or bytes, possibly uninterpreted, or understood as sets of low-level (syntactic) features.
- Metadata: Particular, predefined facets of documents or data are recorded and available for searching or

exchange. An example is the Dublin Core metadata standard [6]. Generally, such facets support searching by people. To the extent that the domain of a facet is specified and formalized, such as spatial extent in Federal Geographic Data Committee metadata [7], it may support machine reasoning about content.

- Extensibly structured information: Extensible markup systems such as XML allow for the semantic tagging of content that may be used to support fine-grained filtering and retrieval of content.
- Knowledge representation: Use of ontologies and other domain models allows deeper semantic analysis of content and supports automated information transformation and integration across sites.

Service dimension. This dimension characterizes the complexity of processing that DLs and federations of DLs can manage on behalf of clients.

- Web service: A DL offers one or more capabilities as Web-callable services directly to a client. The service definitions might be proprietary or conform to a protocol (e.g., OAI (http://www.openarchives.org) or framework [e.g., ODL (http://oai.dlib.vt.edu/odl/)].
- Workflow management: DLs provide for definition, execution, and monitoring of compositions of services over distributed sites on behalf of a client [19].
- Agent hosting: DLs support multiple, communicating mobile agents operating on behalf of clients.

In each dimension, deployed systems tend to be at the first or second point, while research systems are further along on particular dimensions. The research agenda presented here will drive advances in all these dimensions, both in deployed and experimental systems. Developing generic infrastructure to build effective DLMSs is clearly of fundamental importance. It requires solutions to several "horizontal" problems that touch upon many aspects of a DLMS and its environment. Although prior results in related areas are valuable, many difficult problems are largely open. We have sought to identify the critical problems and suggest promising directions that we hope will push DLs into the mainstream.

Acknowledgements. This work was supported by the Information Society Technologies (IST) Program of the European Commission as part of the DELOS Network of Excellence (Contract IST-1999-12262) and by the National Science Foundation of the United States (Grant IIS 0226501).

References

1. Alonso G, Casati F, Kuno H, Machiraju V (2003) Web services. Springer, Berlin Heidelberg New York
2. Amato G, Gennaro C, Savino P (2001) Searching documentary films on line: the ECHO Digital Library. In: Proceedings of the International Cultural Heritage Informatics meeting (ICHIM '01), Milan, Italy, September 2001, pp 147–155
3. Bowman CM, Danzig PB, Hardy DR, Manber U, Schwartz MF (1995) The Harvest Information Discovery and Access System. Comput Netw ISDN Syst 28(1–2):119–125
4. Berners-Lee T, Hendler J, Lassila O (2001) The Semantic Web. Sci Am 284(5):34–43
5. Crane G, Wulfman CE, Cerrato LM, Mahoney A, Milbank TL, Mimno D, Rydberg-Cox JA, Smith DA, York C (2003) Towards a cultural heritage digital library. In: Proceedings of the 3rd ACM/IEEE-CS joint conference on digital libraries, JCDL 2003, Houston, TX, June 2003, pp 75–86
6. Dekkers M, Weibel S (2003) State of the Dublin Core Metadata Initiative. D-Lib Mag 9(4)
7. Federal Geographic Data Committee (1998) Content standard for digital geospatial metadata. FGDC-STD-001-1998, Washington, DC, June 1998
8. Fox EA, Gonçalves MA, McMillan G, Eaton J, Atkins A, Kipp N (2002) The Networked Digital Library of Theses and Dissertations: changes in the university community. J Comput Higher Educat 13(2):3–24
9. Foster I (2002) The Grid: a new infrastructure for 21st century science. Phys Today 55(2):42–47
10. Fensel D, Wahlster W, Lieberman H, Hendler J (eds) (2002) Spinning the Semantic Web: bringing the World Wide Web to its full potential. MIT Press, Cambrdige, MA
11. Gross T (2003) CYCLADES: A distributed system for virtual community support based on open archives. In: Proceedings of the 11nth Euromicro conference on parallel, distributed, and network-based processing (PDP 2003), Genoa, Italy, February, pp 484–491
12. Hull R, Benedikt M, Christophides V, Su J (2003) E-services: a look behind the curtain. In: Proceedings of the 22nd symposium on principles of database systems, San Diego, June, pp 1–14
13. Suleman H, Fox EA (2001) A framework for building open digital libraries. D-Lib Mag 7(12)
14. (2003) Digital Library Information-Technology Infrastructures. Report of the DELOS-NSF Working Group on Information-Technology Infrastructure, November 2003. Available at the following URL address: `http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/joint-wgs.html`
15. Keiper J, Brocks H, Dirsch-Weigand A, Stein A, Thiel U (2001) COLLATE – A web-based collaboratory for content-based access to and work with digitized cultural material. In: Proceedings of the International Cultural Heritage Informatics meeting (ICHIM '01), Milan, Italy, September 2001, pp 495–511
16. Lagoze C, Arms WY, Gan S, Hillmann D, Ingram C, Krafft DB, Marisa RJ, Phipps J, Saylor J, Terrizzi C, Hoehn W, Millman D, Allan J, Guzman-Lara S, Kalt T (2002) Core services in the architecture of the national science digital library (NSDL). In: Proceedings of the ACM/IEEE joint conference on digital libraries (JCDL 2002), Portland, OR, June 2002, pp 201–209
17. Lagoze C, Davis JR (1995) Dienst – An architecture for distributed document libraries. Commun ACM 38(4):47
18. Paskin N (2003) DOI: a 2003 progress report. D-Lib Mag 9(6)
19. Schuler C, Schuldt H, Tuerker C, Weber R, Schek H-J (2004) Peer-to-peer execution of transactional processes. J Cooper Inf Syst (to appear)
20. Schek H-J, Schuldt H, Weber R (2002) Hyperdatabases – infrastructure for the information space. In: Proceedings of the 6th conference on visual database systems (VDB'02), Brisbane, Australia, May 2002, pp 1–15
21. Van de Sompel H, Lagoze C (2002) Notes from the interoperability front: a progress report on the Open Archives Initiative. In: Proceedings of the 6th European conference on research and advanced technology for digital libraries, Rome, September 2002. Lecture notes in computer science, vol 2458. Springer, Berlin Heidelberg New York, pp 144–157