Digital Libraries: Issues and Architectures

Peter J. Nürnberg, Richard Furuta, John J. Leggett, Catherine C. Marshall, Frank M. Shipman III
Center for the Study of Digital Libraries
Texas A&M University
College Station, TX 77843-3112 USA
{pnuern, furuta, leggett, marshall, shipman}@bush.cs.tamu.edu

ABSTRACT
The research field of digital libraries must be viewed as a union of subfields from a variety of domains combined with new research issues in order to realize its full potential. A clear exposition of the research issues involved has not yet been given. Most approaches to building digital library systems have thus far been limited to addressing specific digital library problems as variations of problems from other fields. This paper presents a taxonomy of digital library elements. Consideration of the elements in this taxonomy helps suggest a variety of issues. Example elements and some issues they suggest are used to populate the taxonomy. The paper continues by presenting a general digital library system architecture. Issues suggested by the taxonomy are shown to have implications at many levels of digital library system architectures for both design and implementation. This is illustrated by considering the implications of one issue (personalizing presentations) at several architectural levels and in the context of a set of current technologies.

Keywords: digital library issues, digital library architecture, databases, physical libraries, World Wide Web

INTRODUCTION

The emerging field of digital libraries brings together participants from many existing areas of research. Currently, the field lacks a clear agenda independent of these other areas. It is tempting for researchers to think that the field of digital libraries is a natural outgrowth of an already known field. From a database or information retrieval perspective, digital libraries may be seen as a form of federated databases. From a hypertext perspective the field of digital libraries could seem like a particular application of hypertext technology. From a wide-area information service perspective, digital libraries could appear to be one use of the World Wide Web. From a library science perspective, digital libraries might be seen as continuing a trend toward library automation. There is some truth to these perspectives (as well as others) but none address the field as a whole and its research agenda. The field of digital libraries will be limited if viewed only as a subfield of prior research interests. To realize its full potential, the field must be viewed as a union of subfields from a variety of domains combined with additional goals, and thus new research issues. Digital library research must both respect the existing tradition of our physical libraries and transcend current practice in developing a new, broader research agenda.

What are the research issues central to digital libraries? One issue might be how to digitize objects and put them on-line. A second might be how to include new forms of information that do not have temporal or tangible representation necessary for inclusion into physical libraries. Another could be how to locate materials in the new digital library. Yet another would be when to use and when to transcend the existing technologies and traditions of the physical library in its digital form. Still other issues stem from the problems of information overload created by new information technologies. This paper presents a framework for thinking about the field of digital libraries and the research issues that are part of it and demonstrates how these issues affect digital library systems.

The next section gives an analysis of the digital libraries field by positing that the digital library can be modeled to some degree after the physical library, and discussing the relationship between the two. In order to show the breadth of the research agenda in digital libraries, a taxonomy of the of elements of the digital library, and some issues raised by considering these elements is then presented. Following this, a general system architecture for digital library systems is presented. Issues suggested by considering the prior taxonomy are shown to affect many layers of these systems.

PHYSICAL AND DIGITAL LIBRARIES

Why is a digital library called a library at all? This question has been addressed by various members of this research community. Miksa and Doty [10] discussed the notions of collection, information sources, and place with respect to physical libraries and how these notions might carry over into the digital realm. Levy and Marshall [6] considered how work practices in physical libraries might be used in the design of digital libraries. The physical library can provide the starting point for discussing the elements and domains of digital libraries. An element of a library is a constituent part of the library. A domain of the library is the universe from which the library materials are drawn.

Elements

It is helpful to consider three broad classes of library elements: data, metadata, and processes. *Data* are library materials. *Metadata* are information about the library and its materials. *Processes* are active functions performed over library elements. For example, a book in a library may be thought of as being data of that library. An index over book titles (in a card catalog, for example) may be thought of as library metadata. The act of a librarian helping a patron find a book by suggesting the use of the card catalog may be thought of as a process.

This classification is vague, in the sense that it may be difficult or impossible to classify any given library element as distinctly belonging to a particular class. It may be possible to view a single element as belonging to all three classes. However, this classification is useful since it provides a framework for discussion about library elements. Physical library elements often fulfill some role for a given library user at a given moment. These roles often can be assigned in specific cases in a meaningful way.

Because this classification concerns elements in the library, it ignores differences in roles played by people interacting with the library, the various ways in which these roles are being reassigned in the digital library, and the different high-level tasks performed by people fulfilling these roles. These are of course all important issues, but will not be considered here.

This classification of physical library elements can be applied to digital library elements as well, with the same understanding that a given element may be thought of differently by different users at different times.

Domains

A physical library deals primarily with physical data, whereas a digital library deals primarily with digital data. Of course most modern libraries deal with both, but it is useful for sake of discussion to consider hypothetical "all-physical" and "all-digital" libraries as foils.

If physical libraries primarily contain physical data and digital libraries primarily contain digital data, then how can digital libraries preserve and disseminate the vast amounts of existing physical data? Instead of containing the physical data itself, digital libraries will contain digital translations of this data. The term translation is used, because the process of generating these digital representations of physical data is not necessarily a completely meaning-preserving process. The product may not be perceived by users in the same way that the source is perceived since their media of presentation are necessarily different [9].

It might be tempting to think that if there are differences between analogous physical and digital objects, they have no practical consequence. This would imply, however, that all such differences are already known. Not only is this not the case, but it is not even clear that all such differences can ever be known, because one cannot know, a priori, all the important characteristics of an object in any situation [13]. Without knowing all of the differences between physical and digital objects, how could one claim that these differences are insignificant?

The magnitude of differences between physical and digital analogs may be related to the accuracy of the physical/digital translation. A spectrum of translation quality certainly exists. Without more research into the effects of translating material between physical and digital form, it is difficult to know the accuracy of such translation.

The difference between the physical and digital domains also has implications for translating the metadata and processes of physical libraries. Some of the metadata and processes of a physical library (e.g. card catalogs and shelving) are themselves physical elements, and thus, the discussion of translations as formulated above applies. However, even those elements of the physical library that have no direct physical reality (e.g. the Library of Congress classification scheme) are often inextricably tied to the physicality of data and the library itself. These abstractions also need to be translated into the digital realm.

In summary, though both physical and digital libraries may be thought of as sharing certain goals and consisting of elements that may be classified similarly, the domains of the two types of libraries differ. Digital libraries will deal with translated physical elements, conceptual elements of the physical library adapted to the digital realm, and completely

	Data	Metadata	Processes
Translations of Physical Library Entities	Book Journal Movie	Static index Classifications Spatial arrangement	Acquiring data Suggesting sources Helping locate sources
New Digital Library Entities	Hypernovel Scientific visualization Computer program	Dynamic index Personalized structure Annotations	Full-text searching Personalizing presentation Retrieving by agents

Figure 1: Taxonomy of Digital Library Elements.

new digital elements with no apparent physical library analog (e.g. hypernovels). Differences between physical library and digital library elements have created many open problems concerning how to adapt the tradition of the physical library into the digital realm.

TAXONOMY OF DIGITAL LIBRARY ISSUES

Given the above discussion, it is reasonable to classify the elements in digital libraries along two axes. Firstly, elements may be classified as data, metadata, or processes. Secondly, these elements may be translations of physical library elements or new digital library elements with no clear physical library analog. This results in the grid shown in Figure 1.

Each section of the grid is discussed below. Examples of elements that may be thought of as belonging to the section in question are given, followed by an issue particularly relevant to that section. These issues and their positions in the grid are shown in Figure 2. As stated earlier, a given element may be thought of as being classified in many different sections on the grid, but elements are placed so that some typical use of that element is highlighted. Also, problems raised in each section may (and often do) apply to other sections as well, but may be thought of as having special significance in their respective sections.

Translations of Physical Library Data

It is easy to find examples of physical library data that are translated into digital form routinely. For example, books, journals, and movies are all examples of physical library data that are scanned, digitized, or otherwise translated into electronic form [5].

A central problem in translating physical library data is deciding which aspects of the original merit consideration in the translation process. When translating a book into digital form, when does an ASCII representation of the text suffice? When must each page be scanned as a photograph would be? How are such decisions to be made? These questions involve many tradeoffs, and answers cannot be known in the general case [7].

It is not even clear which characteristics of an object are most meaningful. Many characteristics of physical data, such as size and shape of a book, may be meaningful only to some people or in only some circumstances. Consider how grease smudges on the sides of auto parts manuals aid people in finding desired pages [4]. It is impossible to include every characteristic of a physical data object that may ever be deemed meaningful to any person, but ignoring meaningful aspects of an object during translation has important implications for the preservation of function in a digital library.

Translations of Physical Library Metadata

Examples of physical library metadata are plentiful. Long-lived indexes (such as those in card catalogs), classification schemes (such as the Library of Congress classification scheme) and spatial arrangement of library materials are three examples.

A problem with translating such physical library metadata is that often either the metadata itself or its application is influenced by the physicality of the data. For example, the spatial arrangement of data objects in a physical library conveys meaning and is a form of metadata. Spatial arrangement of objects is meaningful because the objects have some physical presence. How can this be translated into the digital realm? Is a virtual reality approach, in which digital objects are associated with some virtual physical presence in a virtual physical place, the correct way to translate this metadata? Or, is the correct approach one that spatially arranges abstract images in an abstract space?

While spatial arrangement of library materials is a physical library metadata element with physical presence, other metadata with no direct physical reality must also be translated, or adapted in its application, if it is to be used in a digital library. For example, the Library of Congress classification scheme may not have any physical reality itself, but its application is sometimes constrained by the physicality of the objects it classifies. For example, such a classification scheme is often used to guide the physical location of data in a library, because placing like-classified objects in physical proximity can aid patrons in locating data. If a library has one copy of a book, but the book could be classified in more than one category, how is the book to be located? It can effectively only be co-located with

	Data	Metadata	Processes
Translations of Physical Library Entities	What to translate?	How to translate metadata that is dependent on data physicality?	How to provide tools for human involvement in these processes?
New Digital Library Entities	How to account for the continual rapid evolution of new data types?	How to insure consistency of separately maintained metadata?	How to distribute computation?

Figure 2: Issues Raised by Considering the Taxonomy of Digital Library Elements.

sources of one classification. This same limitation does not hold for digital objects located in a virtual space.

Translations of Physical Library Processes

Many kinds of physical library processes exist. Three examples of such processes are acquiring data, suggesting the usefulness of elements, and aiding in the location of elements. An example of acquiring data is choosing new books to add to a library. Suggesting the usefulness of elements might take the form of a patron identifying potentially helpful data and metadata sources to a colleague who might otherwise not have known about nor used these sources. An example of aiding in the location of elements is a library worker helping a patron locate an object given incomplete information.

One characteristic shared by many physical library processes is that they are performed by human beings. A key problem in translating such physical library processes into the digital library realm is how to provide human beings with tools to assist them in performing these often informal processes, especially since digital library patrons and librarians cannot rely on co-location with people likely to be helpful. This problem is particularly important given the inherently collaborative nature of many tasks performed in the library [3, 8, 12].

New Digital Data

Hypernovels, scientific visualizations, and active computer programs are all examples of new digital library data that do not have clear physical library data analogs. It could be claimed that novels on paper are clear predecessors to hypernovels, but hypernovels have many characteristics that qualitatively differentiate them from their paper counterparts [11]. It is certainly conceivable to build a library of active computational objects. Also, many physical objects not currently included in the physical library due to space or other restrictions (e.g. transcripts of radio programs or videos of television shows) may have digital analogs in the digital library.

One problem faced by digital library designers and implementers when considering new digital library data is that new types of this data are constantly and rapidly evolving. While it is true that new physical types of data are constantly evolving, the pace of change in the digital realm is currently greater, because of immaturity of new digital data types. New potentials are constantly being recognized and used. It is particularly difficult to design or implement a digital library if the types of data to be included in the library are not yet known.

New Digital Metadata

Many new kinds of metadata are possible in a digital library. Three examples are dynamically generated indexes, personalized structures over library elements, and annotations. Dynamically generated indexes may have relatively short life-spans compared to the long-lived indexes of the physical library. One example of personalized structures are user- or group-specific sets of hypertext links over some set of library elements. Annotations are virtual modifications of data objects by patrons – these modifications exist separately from the data but may be always displayed with the data for a particular user or group, thereby effecting a “virtual” modification [7].

A problem with new digital library metadata is that much of it is personal, and thus may be stored separately from the data over which it applies, leading to possible consistency errors. If many users build structure over certain data in a library, and that data changes, what should be done with all of the metadata that is in some way invalidated by this change? This is certainly a problem in the physical library. Because most physical library metadata resides in the library itself, however, it may be easier to modify the metadata to reflect any changes in data. With personal digital library metadata, all such copies of metadata may not be known. To what degree is the digital library system responsible for propagating changes to patrons with metadata that relies on the changed material? How can this propagation be effected?

New Digital Processes

Finally, the digital library allows new processes not found in the physical library. Specifically, processes such as full-text searching, personalizing presentations, and retrieving by agents are new digital library processes. Full-text searching refers to querying a full-text index. Personalizing presentations involves access control issues as well as tailored screen layouts. Retrieving by agents involves programs that search data autonomously and report findings to users.

One problematic aspect of these new processes is that they involve computation that may access large amounts of library data or metadata. A central problem is how to distribute the computation needed to maintain these processes. For example, how much of the computation involved in personalizing presentation of information should be done by the server and how much should be done by the client? If such processes are computationally expensive, how can this load be fairly distributed? What is the optimal mix of client / server communication, server-side computation, and client-side computation for effecting these processes?

DIGITAL LIBRARY SYSTEMS

The taxonomy of issues presented in the previous section illustrates the wide range of problems to be considered when designing and implementing a digital library. This section presents a conceptual template of a general digital library system architecture and illustrates by example how issues identified in the previous taxonomy can have implications in several areas of this architecture. The section closes by considering the role played by some of today’s current technologies when constructing a digital library system.

Digital Library System Architecture

Conceptually, a digital library system may be thought of as mediating certain kinds of interactions among people and computing systems. Figure 3 shows some relationships and interactions among several parts of the digital library and several people and systems external to the library. To help clarify the interactions occurring in these relationships, the computing resources in this figure have been partitioned into server resources and client resources. This allows the classification of computer-supported relationships into human/human, human/client, human/server, and client/server classes.

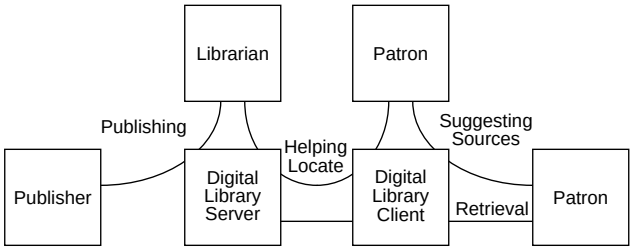

Figure 3. Conceptual Role of a Digital Library System with Example Relationships.

The real relationships are often more complicated than shown. For example, publishing in the digital library is not strictly a relationship between publisher, librarian, and the digital library server. Patron needs, budgetary constraints, limitations of library computing resources, and a number of other factors may be involved. Any robust digital library system should provide support for these complex relationships.

The client and server computing systems may each be further subdivided. Each may be thought of as consisting of three parts: the back-end, the "middle-end", and the front-end. Both the back-end and the front-end of a system define interfaces between the system itself and some external entity. A system front-end normally provides services to external clients, while the back-end is provided with services from external servers. The middle-end provides some intermediate mapping between the front- and back-ends. Figure 4 illustrates the same entities as shown in Figure 3, but with the divisions of the client and server into their respective back-, middle-, and front-ends.

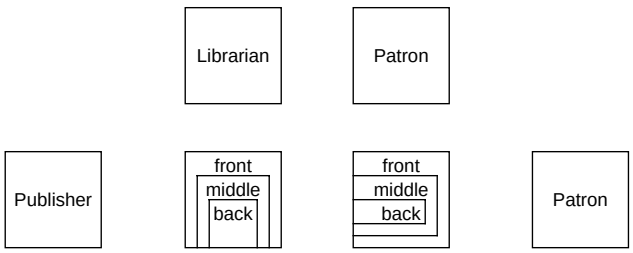

Figure 4. Digital Library System Architecture.

Mapping Issues to Solutions

The issues identified in the taxonomy presented in the previous section may have implications in several areas of the digital library system. This section illustrates this point by taking one issue raised previously and identifying the areas of the digital library system that are affected.

Consider the issue raised in the discussion of new digital processes – how can the computational and storage load be equitably divided between client and server for these new processes. Specifically, consider the new digital process of personalizing the presentation of material.

Addressing this issue cannot be confined to any one part of the digital library system. The publishers of digital library data must consider *how to format* their data stored in the server back-end so that it may be presented in a personalized way on the client side. The server middle-end must address *how much preprocessing* should be done, which involves a tradeoff between possibly sending too much unprocessed data versus spending too much computing time on the server side. The server front-end and the client back-end must agree on *which protocol* to use to send the semi-processed data. The client middle-end must address *how to distribute* data retrieved from the server among many displays on the client front-end processes. Finally, the client front-end must address *how to make personalization of presentation a usable feature* for library patrons. These points are just some examples of what must be considered at different levels of a digital library system to address one element or issue raised in the above discussion on the taxonomy of elements.

Current Technologies

This section closes by considering how one set of current technology maps to the general digital library system architecture, and how the example of personalized presentations is addressed by this current technology. The technology considered is a set of WWW clients communicating with httpd servers that use Common Gateway Interface (CGI) scripts and/or binaries to access a database [2]. This system and its mapping to the terminology presented above is shown in Figure 5.

Consider how this technology answers just the questions raised in the above section. There are many ways for publishers to answer the question of *how to format* their data. Several popular formats exist for digital data translated from the physical realm, such as Graphics Interchange Format (gif) for still video images or ASCII for plain text. Publishers of database data may choose any of these popular formats appropriate for their needs, since many of the more popular formats can be handled on the client front-end. Formats for new digital data types are still forming, such as the evolving HyperText Markup Language (HTML) for hypertextual documents [1]. There are no generally agreed upon formats for more exotic digital elements such as process-based dynamic hypertexts.

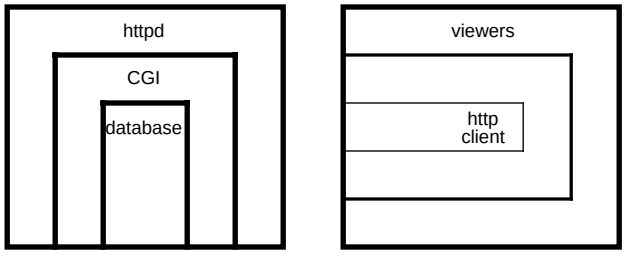

Figure 5. Current Technology Mapping to Digital Library System Architecture.

Distinct processes are separated by heavy lines. Divisions that may or may not imply separate processes are marked by medium lines. Hypothetical intra-process divisions are marked with light lines.

The question of *how much server-side preprocessing* of the data can be done by CGI scripts is difficult to answer. On the one hand, these scripts are capable of arbitrary computation, and can be passed meaningful strings appended to URL's. However, the scripts themselves are static. In current practice, because presentations are rarely personalized at the client front-end, CGI scripts rarely do much preprocessing of the retrieved data before passing it to the server front-end.

The question of *what protocol* is to be used between the server front-end and the client back-end seems to be temporarily resolved in favor of a mix of http, ftp, gopher, and a handful of other protocols. New protocols can clearly be and will need to be added to support new data types by adding new URL access methods. However, the fact that the same object referenced by two URL's with different access methods may have different (non-access method) identifiers does not allow easy dynamic negotiation of protocols between server and client. One research issue to consider is the effects this dependence of the identifier has on the access method.

Currently, most Web clients do not support multiple front-ends in any meaningful way. This means that multiple front-ends require the back-end to replicate server calls even if they are displaying the same data. Thus, the current technology does not address *how to distribute* client-retrieved information to multiple client front-ends.

Finally, current Web clients only allow a small degree of personalization of presentation. This is essentially limited to specifying viewers for non-inlined data, specifying some parameters for how to display in-lined data, and possibly providing information to the server via an HTML forms interface about what kind of data should be retrieved. Thus the only personalization of data in the client front-end concerns display of data and not access to data. Web clients need to provide more tools to patrons of digital libraries to *allow easy personalization* of data with respect to both presentation and access.

In summary, Web clients communicating with httpd servers using CGI scripts to access databases has technology in several of the areas of the general digital library system architecture outlined above, with the exception of an identifiable client middle-end to handle multiple front-ends corresponding to one client back-end. Some issues, such as how to format new data types and what protocols to use to communicate this data, can be addressed somewhat independently and solutions can be integrated at a later time. Other issues, such as client-side filtering of information that allows personalization with respect to access, are not currently addressed.

CONCLUSIONS

Physical libraries provide a good starting point for discussion of digital libraries. Elements of both the physical and digital libraries may be categorized as data, metadata, or processes; these categories are determined in specific instances by the intended use of elements by librarians, patrons, or others. Data, metadata, and processes of the physical library must be translated into the digital domain if they are to be used in the digital library. Additionally, there are types of library elements with no clear physical library analog – wholly new digital library elements. These observations led to the development of a taxonomy of digital library elements.

Issues raised by the taxonomy of digital library elements have implications at several levels of digital library systems. Examining the problem of personalizing presentations identifies sample issues at all levels of the architecture. Specifically, considering personalizing presentations led to identifying issues of data format (server back-end), server-side preprocessing (server middle-end), protocols (server front-end to client back-end), client-side distribution (client middle-end), and user tools (client front-end). By first identifying a digital library issue, and then considering the implications for system design and implementation, the myopia of considering issues at one architectural level isolated from issues at other levels is avoided. Also, by applying this approach from *digital library issue* to *digital library system solutions*, system designers and implementers can better understand that decisions made at one architectural layer about seemingly low-level issues (e.g. how to format data) can affect high-level capabilities (e.g. personalizing presentations) provided to the end-user.

The field of digital libraries presents a set of complex issues, and solutions to these problems will require a blending of approaches from a variety of fields. Claims that any one technology has solved all of the issues posed in the design and implementation of digital libraries fail to address the entire problem. For example, proponents of the view that federated databases solve the technical issues of digital libraries have only considered technology at the server back-end to handle already made translations of physical library data and metadata. Even augmenting such databases with other current technologies such as Web clients, httpd's and CGI scripts does not provide a fully functional digital

library system. Instead, any successful attempt at constructing a digital library system will need to address issues raised by considering the many different kinds of digital library elements throughout the various levels of the general digital library system architecture.

ACKNOWLEDGEMENTS

Part of the research described in this paper has been supported by the Texas Advanced Research Program under Grant No. 999903-155.

REFERENCES

[1] Berners-Lee, T. J. and Connolly, D. W. 1995. HyperText Markup Language Specification - 2.0 (IETF Draft).

[2] Berners-Lee, T. J., Cailliau R., Groff, J. F., Pollermann B. 1992. World-Wide Web: The information universe. *Electronic Networking: Research, Applications and Policy 2* (1) (Spring), pp. 52-58.

[3] Ehrlich, K., and Cash, D. 1994. Turning information into knowledge: Information finding as a collaborative activity. *Proceedings of the Digital Libraries '94 Conference*, (College Station, TX, Jun 19-21), pp. 119-125.

[4] Hill, W. C., and Hollan, J. D. 1992. Edit wear and read wear. *Proceedings of the Human Factors in Computing Systems '92 Conference*, (Monterey, CA, May 3-7), pp. 3-10.

[5] Lesk, M. 1991. The CORE electronic chemistry library. *Proceedings of the Fourteenth Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, (Chicago, IL).

[6] Levy, D. M., and Marshall, C. C. 1994. Going digital: a look at assumptions underlying digital libraries. *Communications of the ACM 38* (4).

[7] Løkken, S. 1993. Text Representations In Digital Hypermedia Library Systems. M.S. Thesis. Department of Computer Science, Texas A&M University. College Station, TX (Dec).

[8] Marshall, C. C., Shipman, F. M., and McCall, R. J. 1994. Putting digital libraries to work: Issues from experience with community memories. *Proceedings of the Digital Libraries '94 Conference*, (College Station, TX, Jun 19-21), pp. 126-133.

[9] McLuhan, M. 1964. Understanding media; the extensions of man. McGraw-Hill. New York.

[10] Miksa, F., and Doty, P. 1994. Intellectual realities and the digital library. *Proceedings of the Digital Libraries '94 Conference*, (College Station, TX, Jun 19-21), pp. 1-5.

[11] Moulthrop, S. 1991. Beyond the electronic book: A critique of hypertext rhetoric. *Proceedings of the Third ACM Conference on Hypertext (Hypertext '91)*, (San Antonio, TX, Dec), pp. 291-298.

[12] Schnase, J. L., Leggett, J. J., Metcalfe, E. S., Morin, N. R., Cunnius, E. L., Turner, J. S., Furuta, R. K., Ellis, L., Pilant, M., Ewing, R. E., Hassan, S. W., and Frisse, M. 1994. The CoLib project–Enabling digital botany for the 21st century. *Proceedings of the Digital Libraries '94 Conference*, (College Station, TX, Jun 19-21), pp. 108-118.

[13] Suchman, L. A. 1987. Plans and Situated Actions: The Problem of Human-Machine Communication. Cambridge University Press. New York.