A Proposal and Prototype for an Information Systems Digital Library

John R. Venable

School of Information Systems
Curtin University of Technology
Perth, Western Australia
VenableJ@cbs.curtin.edu.au

Abstract

This paper describes an initial prototype for an Information Systems Digital Library (ISDL) for the free or low cost input, storage, full-text search, and retrieval of all kinds of publications relevant to the field of IS. The prototype is intended serve as a discussion point for the worldwide IS community, with the aim that an enhanced ISDL eventually be provided by and for the IS community. The paper proposes possible objectives for an ISDL, describes the features and interface of the prototype, and outlines current and planned research in providing such a system to a worldwide virtual community.

Keywords

Digital library, requirements, prototype

INTRODUCTION

The Information Systems Digital Library (ISDL) project proposes to provide a digital library system that would support IS researchers, students, and practitioners around the world. Among other things, an ISDL is intended to provide free or low cost browsing, full-text search, and retrieval of all kinds of literature relevant to the information systems field. We envision that such a system could provide its services through a community-based approach, e.g. as part of or an adjunct to ISWorldNet.

This paper describes an initial prototype for such a system, including its requirements, architecture and rationale. The following sections of the paper describe the motivation and a proposal for an ISDL, the requirements for a demonstration ISDL prototype, and the features and interface of the current version of the ISDL prototype. Finally, we present conclusions and future research directions.

MOTIVATION

It is currently somewhat difficult to locate and obtain recent, relevant publications in the field of information systems. Figure 1 shows a greatly simplified overview of the situation. Information systems publications come from many sources and via many distribution means. A researcher or student seeking publications is faced with a huge number of choices, which vary widely in their support for searching. Computer-based search indexes often provide only searches of keywords and/or abstracts. A major problem is that the various distribution means typically have low coverage of the available IS publications, for example including only a subset of the journals and not including important conferences' papers, workshop papers or working papers. A researcher is then forced to consult multiple sources (with the consequent time/effort) and yet still being unsure of finding relevant publications.

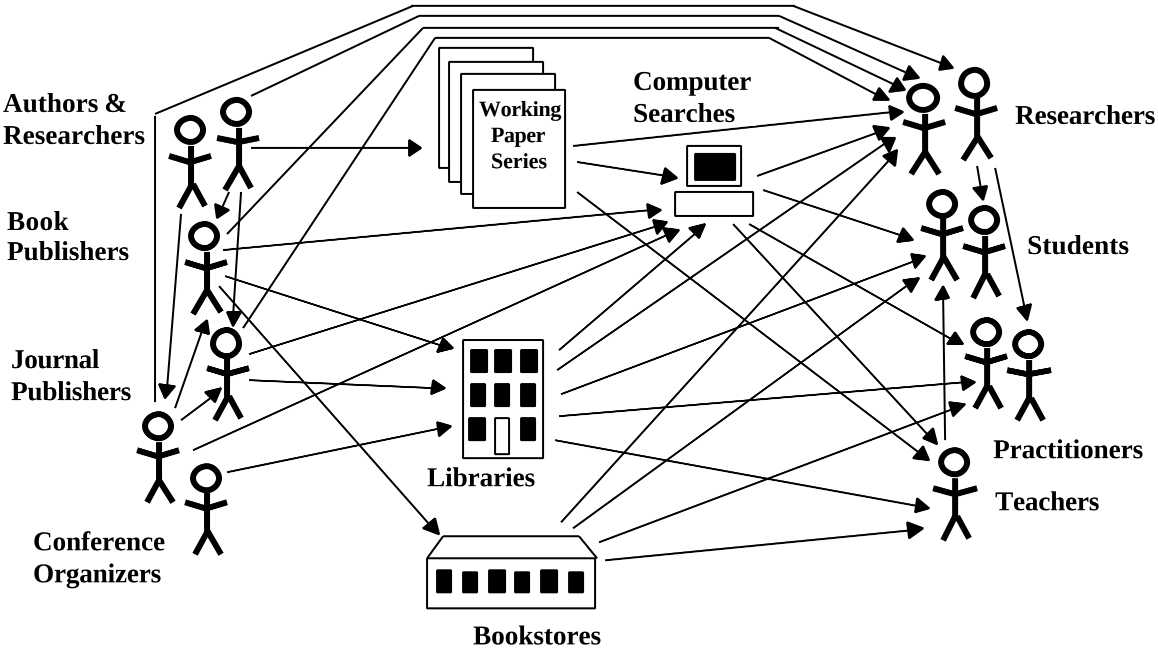


Figure 1: The Current Situation (from Venable *et al*, 1996)

Once a relevant IS publication is identified, the effort, cost, and/or time necessary to obtain a copy of the publication varies widely, and may even be significant enough to be prohibitive.

The end result of these difficulties is that, often, choices for seeking IS publications are made based largely on convenience of searching and retrieving. Consider for example the preponderance of on-line papers found in the reference sections in student papers these days.

A PROPOSAL FOR A COMMUNITY-BASED ISDL SYSTEM

Our vision of an ISDL is one of a community-based service, which supports the IS community's goals and values. As such, there are a number of possible goals and objectives. In our view, the goals and objectives of an ISDL should be determined by the IS community as a whole. In this section, we will briefly introduce some of the possibilities that we consider to be desirable for an ISDL system.

The primary purpose that we see for an ISDL would be to provide a single, unified source for flexible, full-text searching and retrieval of *any* kind of IS publication, at little or no cost, via the internet (see figure 2). While direct, free retrieval is preferable, where it is prevented by copyright or other interests, support should be given for obtaining the publication indirectly through physical library systems or at a cost from the publishers. The main objectives here are to increase the probability of locating relevant publications and to reduce the costs of both searching for and retrieving relevant publications.

An ISDL could also provide other substantial capabilities to support searching and/or retrieval. For example, an ISDL could provide longer-term (multi-session) storage and refinement of search query formulations and results. An ISDL might also provide librarian (human) or automated assistance for using search and other facilities. Collaborative searching could also be facilitated, either with other researchers or with librarians. Simple, topical browsing (Jones & Paynter, 1999) could also be made available. It would also be very useful to be able to easily retrieve publications that are referenced from other publications.

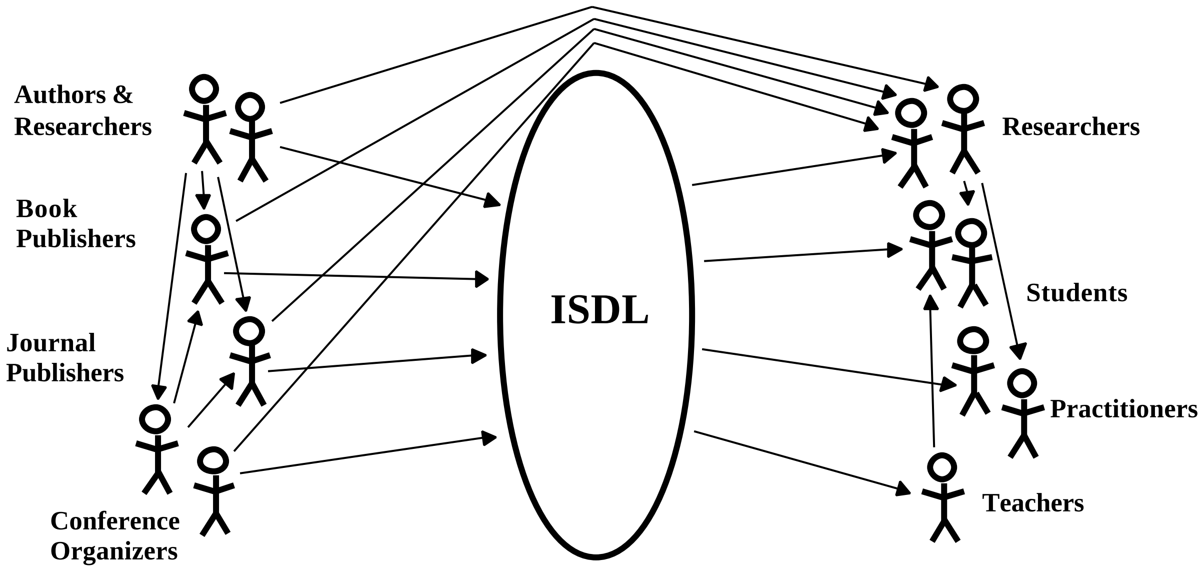


Figure 2: The Envisioned ISDL (from Venable *et al*, 1996)

An IDSL could also provide a host of other interesting features for users, such as provision of references and bibliographies in various paper or electronic formats (e.g. End Notes), incorporation of multi-media (including sounds, graphics, video, animations, annotations, etc.), virtual reality interfaces, and links to more traditional (physical) libraries.

Another area where an ISDL could be helpful is in provision and support of material for teaching purposes. One could make links from course materials to publications provided by the ISDL. One could even build introductory or advanced collections of materials related to particular a particular topic area within the auspices of an ISDL, similar to some of the facilities provided now in ISWorld (e.g. the publication references in (Patnayakuni, 1999))

A particularly important area in or view is that an ISDL library could greatly expand on the role of traditional libraries by providing a tighter connection with and support for scholarly discourse. For example, one could have discussions of particular publications within an ISDL (e.g. see the Reactions section of *Earth Interactions* (1999) and White (1999)) or have general discussions with references to publications available through the ISDL. The ability to reference publications from other publications and have tracing of their links (i.e. hyperlinks) supported by an ISDL would be particularly useful way of supporting scholarly discourse. These concepts follow directly from a view of the IS scholarly community as an inquiring system (Churchman, 1971) or as a (virtual) inquiring organisation (Courney *et al*, 1998). We further envision integrated, full-text searching of these ancillary annotations and discussions (in addition to the main publications). We view such features in an ISDL as natural extensions of Watson's (1994) paper dealing with earlier technologies' support for a world-wide scholarly community. An ISDL could also support the paper reviewing and publication cycle by providing access to publications in progress (see e.g., Sumner & Shum (1996), Roberts (1999), and the recent discussion on the ISWorld mailing list). Even ACM, a mainstay in the

IS are is considering giving access to its pre-print database via its digital library (ACM Digital Library, 1999).

On the more technical side, we believe that an ISDL should be an open system, with interfaces for services provided to other automated systems. For example, it should be possible to send formatted queries to the ISDL system for execution and have the ISDL return an electronic version of the query results for external processing. Similarly, remote systems should be able to retrieve documents if the ISDL internal reference (or some common external reference, e.g. ISBN, is known). Finally, trusted remote systems should be able to submit IS publications and documents for indexing, storage, and later retrieval.

In this section, we have identified a few possible goals, objectives, or features for an ISDL. We envision delivering an ISDL service as a part of ISWorldNet, thus making it freely available worldwide. However, we also believe that the actual choices for the requirements, design, implementation, and even operation and maintenance of an ISDL should be made by the IS community as a whole. It is also very important that an ISDL system provide suitable features that will be accepted by enough of both publication consumers and publication providers/publishers to ensure a critical mass of users (Venable *et al* 1996).

To this end we are conducting research on appropriate methods for developing a consensus for action by the IS community in building and ISDL. It is our belief that current technologies make an ISDL as described above feasible; economic and socio-political issues are the primary obstacles. To this end, we have been researching a way to use a web-based group support system (GSS) to support Soft Systems Methodology (SSM). Work in this area has proceeded from Venable *et al* (1996) with refinement and pilot studies (publications forthcoming).

Part of this method involves delivering an initial prototype ISDL to provide a discussion point and to make potential stakeholders aware of the possibilities an ISDL might offer. We also envision enhancing this prototype significantly in accordance with requirements to be determined by the IS community at large. In the following sections, we report on an initial prototype ISDL. Further work is needed before the prototype can be released for experimentation by the IS community.

REQUIREMENTS FOR AN ISDL PROTOTYPE

In this section, we describe the desired features and document the system requirements for a prototype ISDL. Note that these requirements address only the primary goal and objectives described in the section above. Other requirements/features would come out in discussion with the IS community.

Desired Features

In serving the needs above, the ISDL that we wish to create should include the following features (adapted and enhanced from Venable *et al*, 1996).

1. **Coverage of the IS field:** Publications about all topics considered by members of the IS community to be relevant to the field of IS would be included. The search capabilities should be used to identify relevant publications, not preconceived ideas of what is or is not part of IS.
2. **Coverage of all types of IS publications:** Journal and magazine articles, books, conference and workshop proceedings, working papers series papers, web pages, and anything else identified as useful would be included.

3. **Combined full-text and metadata searching of publications:** Full-text searching is considered to be important as a means of overcoming difficulties with keyword based searches. For example, it allows searching of references or for citations. However, support for searching by metadata (author, publication type, date, language, etc.) about the publication, either alone or in combination with full-text searching, is also necessary. For example you might wish to search only journal publications or for recent publications.
4. **Simplified retrieval of publications:** Direct retrieval of publications wherever possible is the goal, possibly constrained by the need to provide compensation for the authors and/or publishers (see point 5).
5. **Protection of copyright and authors'/publishers' interests:** It is important that publishers and/or authors receive just compensation from those who consume the publications. This can be done by limiting the direct retrieval of documents, either by providing only citations and/or abstracts directly, or by providing indirect means whereby the publication consumers must pay for the publication, as chosen by the publisher.
6. **Facilities to input, index, catalogue, compress, and store publications and information about them:** An important objective is to consider how to automate as much of the input process as possible. Our intention is also that there is as little administration as possible.

Requirements for the Prototype

The above features are addressed to varying degrees in the current prototype. Figure 3 shows a context data flow diagram of the intended system. Note that a library patron could retrieve a publication directly from the ISDL or indirectly from the publisher, depending on the retrieval permission for the document set by the publisher or author.

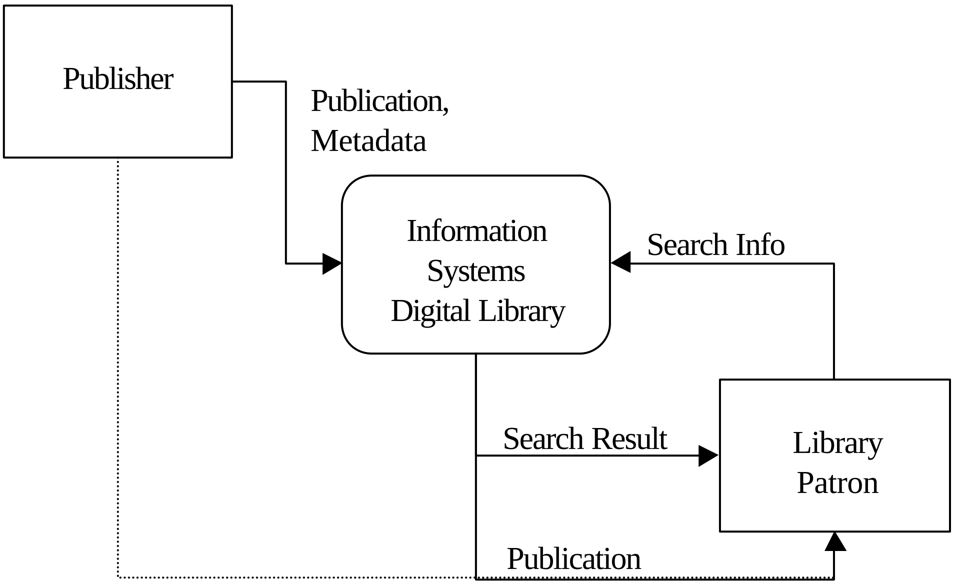


Figure 3: Context Diagram for ISDL System

The ISDL can be broken down further into three main processes, receiving publications, searching for publications, and retrieving publications, as shown in the top level data flow diagram in figure 4.

The metadata input and stored by the system (see figure 4) should include:

1. **Document metadata:** This is information directly about the publication, including citation information (e.g. authors, title, date, publisher, journal issue, pages, language, and publication type – journal article, conference paper, working paper, etc.), descriptive information (e.g. keywords, length, figures, format, etc.), and retrieval information (e.g. URL of original document, permission for retrieval – full-text, abstract-only, citation-only, etc.).
2. **Author metadata:** Information for validating and contacting the author(s), e.g. name, DOB, current address.
3. **Publisher metadata:** Information for validating and contacting the publisher, e.g. name, address, contact person, authorised username (the publisher's agent for using the ISDL).
4. **Serial publication metadata:** Information about a publication series, e.g. name, short name, type (journal, magazine, working papers, conference, workshop, etc.).
5. **Conference or workshop metadata:** Information about a conference, e.g. name, location, date(s), short name. This information is used in citations.

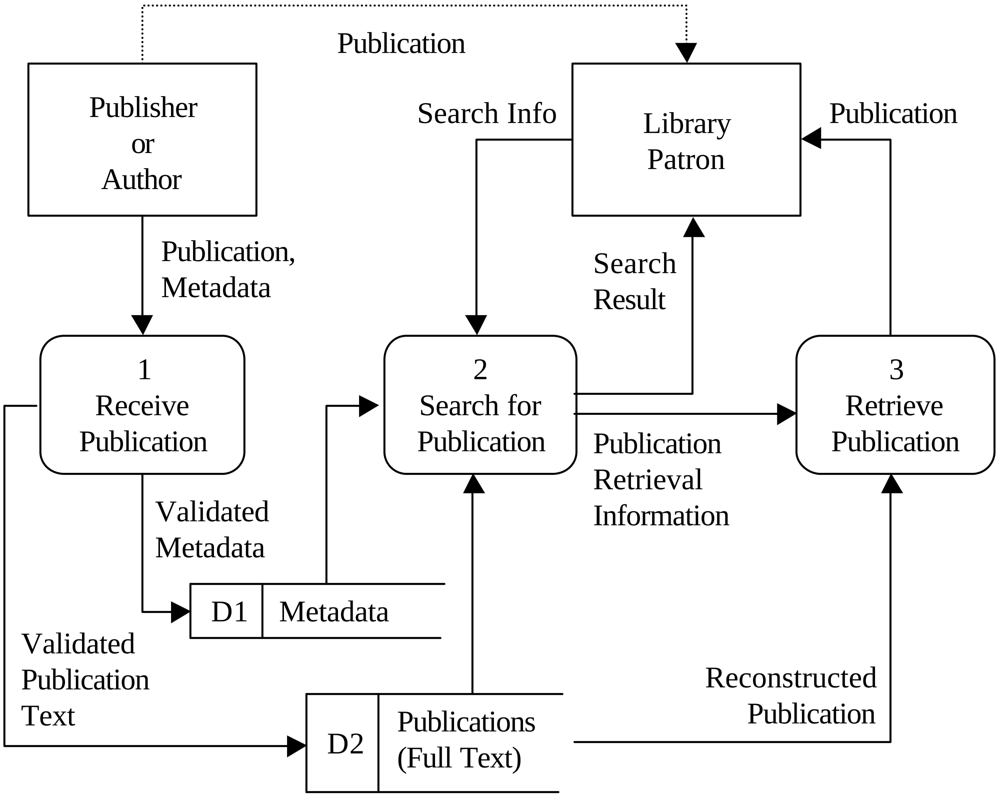


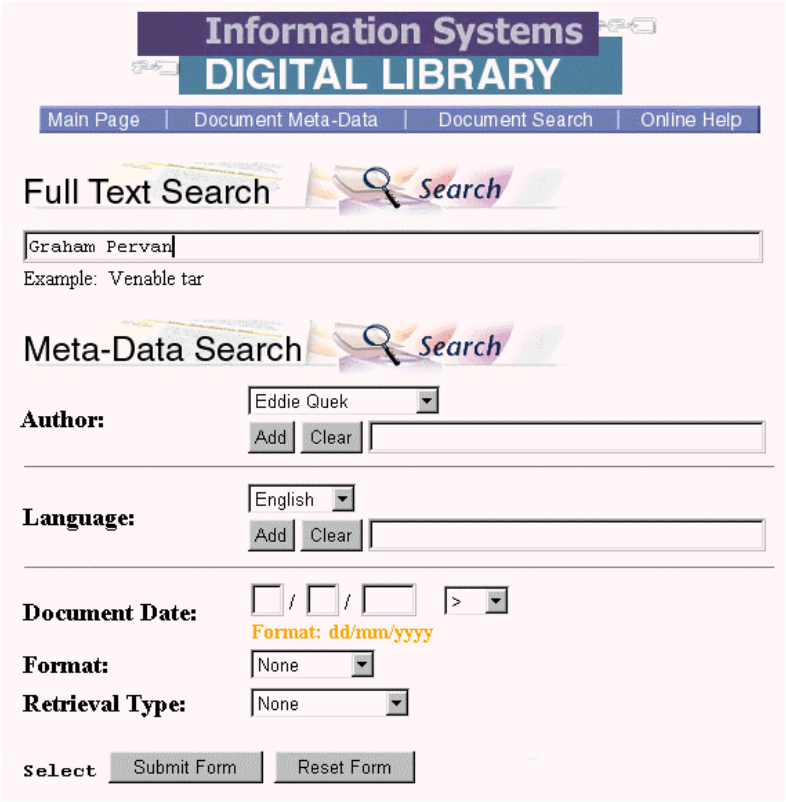
Figure 4: Top Level DFD of ISDL

DESIGN AND IMPLEMENTATION OF THE PROTOTYPE

An initial prototype of an ISDL has been built (ISDL, 1998). The ISDL extends technology developed by researchers working on the New Zealand Digital Library (NZDL, 1999, Witten *et al*, 1998). The current prototype was built using CGI scripts and runs on a UNIX web server. It incorporates many, but not all of the requirements described above. In the next subsection we describe the features of the prototype. In the subsequent subsection, we discuss the limitations of the current version of the prototype.

Prototype Description

The prototype as built explores the need to capture metadata from the authors and/or publishers with as little intervention by a system administrator as possible. It also explores an interface for combining full-text search with metadata based searching for publications. Figure 5 shows the search screen.



Graham Pervan

Example: Venable tar

Author: Eddie Quek

Language: English

Document Date: / / >

Format: dd/mm/yyyy

Format: None

Retrieval Type: None

Figure 5: ISDL Prototype Search Screen

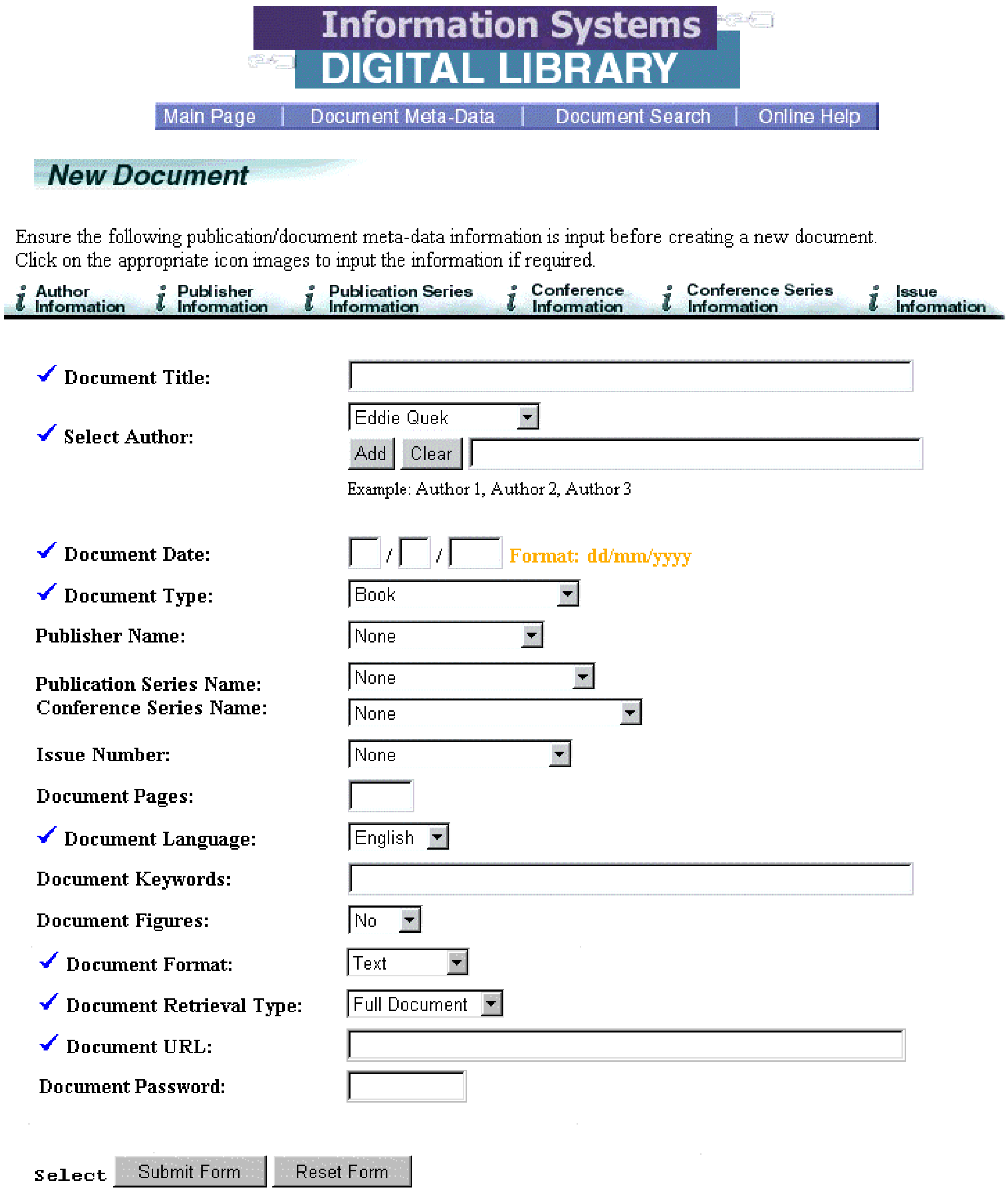
Users can enter text for full-text search (including authors in reference sections as, as in figure 5) and/or other information to narrow the search, as shown. If fields are left blank, they are not used to constrain the search. If both text and metadata are entered, a publication must meet both text and metadata constraints to return a hit.

Figure 6 shows the main document entry screen. The tick marks denote mandatory fields. Pull down (combo) boxes ensure that authors, publishers, etc., are selected only from previously entered data, thereby improving data validation. Additional screens (reached from the horizontal menu just above “Document Title”) provide for entering and maintaining this additional data.

Limitations of the Prototype

The current prototype has a number of significant limitations that need to be rectified before the system is put into actual operation.

First, the current version does not use the *mg* (managing gigabytes) software (Witten *et al*, 1999), which drives the NZDL system. *Mg* provides very fast searching capabilities, as well as significant data compression. The current version simulates this using UNIX text search tools, such as *grep*, which will not scale up for a large number of publications.



Information Systems
DIGITAL LIBRARY

Main Page | Document Meta-Data | Document Search | Online Help

New Document

Ensure the following publication/document meta-data information is input before creating a new document.
Click on the appropriate icon images to input the information if required.

i Author Information *i* Publisher Information *i* Publication Series Information *i* Conference Information *i* Conference Series Information *i* Issue Information

✓ Document Title:	
✓ Select Author:	Eddie Quek Add Clear Example: Author 1, Author 2, Author 3
✓ Document Date:	/ / Format: dd/mm/yyyy
✓ Document Type:	Book
Publisher Name:	None
Publication Series Name:	None
Conference Series Name:	None
Issue Number:	None
Document Pages:	
✓ Document Language:	English
Document Keywords:	
Document Figures:	No
✓ Document Format:	Text
✓ Document Retrieval Type:	Full Document
✓ Document URL:	
Document Password:	

Select Submit Form Reset Form

Figure 6: ISDL Prototype New Document Input Screen

Second, the current version still involves significant system administrator overhead, much of which could be automated. For example, when a new document is entered, the system administrator must manually invoke *ftp* tools to fetch the original document. Similarly, once the document is fetched, the system administrator must give a command (provided by a menu selection on a system administrator screen) to convert the document from either postscript or html into plain text.

Third, MS Word and other word-processor formats are not yet supported. However, authors and/or publishers could simply provide plain-text versions of the documents.

Fourth, security measures have not been implemented. For the system to be useful, the people entering information about publishers, authors, conferences, and so forth must have assurance that no one else can change the data that they have entered. It is intended that publishers' and authors' ability to enter, edit, and delete information should at least be password protected. Fifth, no batch processes for entry of a large number of documents are provided. It must be convenient for conference organisers, journal publishers, etc. to easily send information about a large number of documents, perhaps drawn from their own database systems.

Finally, there are various other small omissions and errors. E.g., a metadata search cannot be constrained by publication type, such as journal publications only.

The principal extensions of the ISDL prototype to the NZDL are (1) the provision of forms/screens for publishers, authors, series editors, conference organisers, etc. (the copyright owner) to input metadata about new publications, which triggers the ISDL to collect the publication, (2) the ability for the copyright owner to specify limited retrieval of or access to the publication, (3) the ability to search on the metadata (possibly in combination with a full-text search), and (4) the ability to handle documents of various kinds. The main NZDL collection of computer science technical reports collects publications directly (with little human intervention) from known web sites and does not collect metadata other than file size, locations, etc. The ISDL and NZDL differ in philosophy in that the NZDL avoids copyrighted material and supports a homogeneous collection of documents while the ISDL prototype is a first attempt to cover copyrighted material of diverse kinds. NZDL research has also focussed on metadata extraction (e.g. author, title) rather than collection of metadata from the document originator. However, it should be stressed that the current ISDL prototype is only an untested concept exploration prototype while the NZDL has become a useful, industrial strength tool at the same time that it is a research platform.

CONCLUSIONS AND FUTURE RESEARCH

As built, the prototype meets many of our requirements for an ISDL, but falls short in a number of areas. Developing the prototype highlighted a number of issues that were overlooked in the requirements. Some of these were small, such as exactly what publication types were required. Others were larger, such as whether both authors and publishers could enter papers and what security would be required. We plan to develop an enhanced prototype that reaches an acceptable state to be made available to the Information Systems community, with the goal of spurring further development with the help of the IS community as a whole.

ACKNOWLEDGMENTS

The initial prototype software described in this paper was built according to the author's specifications by Chau Leong (Gerard) Chin, Yee Khiong (Jacky) Chang, Chu Yang (Eddie) Quek, Malcolm Alan Kite, and Heng Jim (Jim) Wong, under the project supervision of Dr Du Huynh of the School of Information Technology at Murdoch University. This paper expands on an earlier paper published in WAWISR'98 and has benefited from the comments of the anonymous reviewers of both WAWISR and ACIS and from my colleagues at Murdoch University.

REFERENCES

ACM Digital Library (1999) URL http://www.acm.org/dl/ (accessed May 1999).

ARL (1999) Association of Research Libraries, Conference on New Challenges for Scholarly Communication in the Digital Era: Changing Roles and Expectations in the Academic Community, URL http://www.arl.org/scomm/ncsc/conf.html#P3 (accessed June 1999).

Churchman, C. West (1971) *The Design of Inquiring Systems: Basic concepts of systems and organizations*, Basic Books, Inc., New York, NY.

Courtney, James, David Croasdell & David Paradice (1998) Inquiring Organizations, URL http://iops.tamu.edu/faculty/j-courtney/inqorg/inqorg.htm (accessed May 1999), also published in *Australian Journal of Information Systems.*

Earth Interactions (1999) On-line journal, Reactions section, URL http://EarthInteractions.org/E-JOURNAL/react/index.html (accessed May 1999).

Jones, Steve and Gordon Paynter (1999) Topic-Based Browsing Within a Digital Library Using Keyphrases, *Proceedings of Digital Libraries'99 The Fourth ACM Conference on Digital Libraries* (forthcoming), Berkeley, California, 11-14 August 1999.

ISDL (1998) Information Systems Digital Library (prototype). For current status and a link to the current installation, see the author's homepage at: http://www.cbs.curtin.edu.au/units/is/venable/Homepage/

NZDL (1998) New Zealand Digital Library, URL http://www.nzdl.org/ (accessed May 1999).

Patnayakuni, Ravi, ed. (1999) Information Systems Development: An Undergraduate Course Page (ISWorldNet), URL http://www.dis.unimelb.edu.au/staff/ravi/isworld/index.html (accessed May 1999).

Roberts, Peter (1999) Scholarly Publishing, Peer Review, and the Internet, *First Monday: Peer Reviewed Journal on the Internet*. 4, 4, April 5th 1999, URL http://131.193.153.231/issues/issue4_4/proberts/index.html (accessed May 1999).

Sumner, Tamara and Simon Buckingham Shum (1996) Open Peer Review & Argumentation: Loosening the Paper Chains on Journals, *Ariadne (The Web Version)*, Issue 5, September, 1996, URL http://www.ariadne.ac.uk/issue5/jime/ (accessed May 1999).

Venable, John R., Julie Travis, and Marc D. Sanson (1996) Requirements Determination for an Information Systems Digital Library, *Proceedings of the 7th Conference of the International Information Management Association* (IIMA'96), 4-6 December 1996, Estes Park, Colorado, pp. 35-46.

Watson, Richard (1994) Creating and Sustaining a Global Community of Scholars, *Management Information Systems Quarterly,* 18, 3 (September, 1994), URL http://www.misq.org/archivist/vol/no18/issue3/vol18n3art1watson.html (accessed May 1999).

White, John (1999) ACM Digital Library Enhancements, URL http://www.acm.org/dl/dl_enhance.html (accessed May 1999).

Witten, Ian H., Craig Nevill-Manning, Rodger McNab, and Sally Jo Cunningham (1998) A Public Library Based on Full-text Retrieval, *Communications of the ACM,* Vol 41, No 4 (April 1998), pp. 71-75.

Witten, Ian H., Alistair Moffat, and Timothy C. Bell (1999) *Managing Gigabytes: Compressing and Indexing Documents and Images* (Second Edition), Morgan Kaufmann Publishing, San Francisco.

COPYRIGHT