Details of the research proposal

Problem Identification

Digital libraries (a.k.a. electronic libraries, electronic information management systems, networked information systems, etc.) are new to most developing countries, yet we in developing countries have the most to benefit from an unfettered sharing of electronic resources. As collections begin to emerge from the research, library and archive communities, tools and guidelines need to be available in a timely fashion to guide future digital libraries to better meet the needs of local communities while still participating in and contributing to global efforts. This project aims to investigate techniques, models and tools for constructing flexible digital libraries to address the question of how we can effectively and efficiently build flexible digital libraries based on simple components arranged into a network of services. Prior work has already demonstrated the usefulness of developing parts of a digital library as Web-based service components. However, it remains to be shown that the reusability and simplicity inherent in that approach can be also applied to the design of complete systems. This project proposes to investigate the following issues that need to be addressed before a component approach can be adopted holistically by production-quality digital libraries:

a) how to create visual interfaces to compose components into complete systems, as well as the specification of such connections between components,
b) testing techniques and module registries to support the translation from individual components to systems of components,
c) packaging of systems for use at remote sites and how flexibility can be maintained while promoting rapid deployment, and
d) techniques for specifying and designing user interfaces to component systems based on the specification of workflows.

Rationale and Motivation

Digital libraries are software systems that manage electronic resources and make them accessible to users. Traditionally, this software has either been hastily hacked together by Internet–savvy programmers or developed through a regular software development cycle. While the former has resulted in faster turnaround times (albeit of less sophisticated software), the latter has resulted in sophisticated software that is delivered too late to meet the changing needs of its audience. As research progresses in the field, it is fast becoming recognised that current models in software engineering need to be integrated and applied to digital libraries. Most important among these models are the pivotal role of simplicity of design and the construction of larger systems from components (Gladney, et al., 1994; DELOS, 2001).

Until recently, the tools to enable such simple component-based models have not existed but this is gradually changing. The Open Archives Initiative (OAI) (OAI, 2003) led the way by developing a simple standard for sharing of data among disparate networked systems. Their Protocol for Metadata Harvesting (PMH) (Lagoze, et al., 2002) is arguably the only common interoperability mechanism supported by most academic digital libraries, e.g., the Physics Preprint Archive (arXiv, 2003), and a growing number of commercial library vendors. This OAI-PMH was designed as a "low-barrier to interoperability" (Lagoze and Van de Sompel, 2001), thus embodying the elusive goal of simplicity. Many communities have, in adopting this standard, created software tools that are essentially components that are added to their existing digital libraries. In addition, some new systems, notably the Networked Computer Science Technical Reference Library (NCSTRL) (NCSTRL, 2003), replaced their legacy system with one that was designed from scratch to use components that communicated using the OAI-PMH (Anan, et al., 2002). Thus the OAI-PMH embodies both the desirable properties of being simple and modular.

Part of the success of the OAI-PMH is due to the fact that many of the digital library toolkits created in recent years support the protocol by default. In parallel with the activities of the OAI, the University of Southampton created the EPrints software (OpCit, 2003) to manage a pre-print archive or simple refereed journal. DSpace (Smith, et al., 2003) is another package that was recently developed by Hewlett-Packard in collaboration with MIT in order to implement a distributed document management system for a university campus. From a different perspective, the Greenstone package (Witten, et al., 2000) from the University of Waikato uses advanced document analysis and full-text indexing mechanisms to organise collections of documents for efficient user access. All of these digital library toolkits support the OAI protocol in some way or the other and can therefore serve as coarse-grained components of a larger distributed system. This is the approach taken by NCSTRL, which encourages the use of any software as long as it is OAI-compliant.

While all of the systems discussed, and others, can provide the basic digital library infrastructure for a collection, each has its own workflow and suite of user services. There are mechanisms for configuration and extensibility in each (e.g., document analysis plugins in Greenstone) but these are specific to each software package and only support limited extensibility. There is no general model for extending such systems, replacing individual components or interconnecting the systems at a level other than gross data transfer interoperability. A typical digital librarian in search of software will need to assess each of the available packages, decide which comes closest to meeting her or his needs and then customise the software and adapt to its workflow. This is far from ideal – in an ideal situation, it should be possible to combine the best components from each system and design a solution that follows the workflow model known to users of the archive instead of the other way around (Borgman, 1998).

To get closer to this ideal, some component frameworks have emerged in the last year to attempt to model systems as networks of loosely connected components instead of the traditional monolithic model. The Open Digital Library project (ODL) (Suleman and Fox, 2001; Suleman and Fox, 2002) has generalised the well-understood syntax and semantics of the OAI-PMH to support general inter-component communication. This generalisation was then used as the basis for designing a suite of simple protocols to support search engines, category-based browsing, recommendation systems, annotation engines and other typical services expected by users of a digital library. Components, corresponding to each of these protocols, were created and connected together (see Figure 1 in the attached file for the general system model) to test the performance of such systems and the ability of the model to elaborate various different types of digital library systems. The results of such tests (Suleman, 2002) showed that the model has much promise. At the same time, feedback from users and developers has indicated that while simplicity of the individual components is useful, much work still needs to be done in order to simplify the process of going from a set of components to a fully-fledged and seamless digital library.

Concurrent with the development of the ODL model, similar efforts were underway on the OpenDLib project (Castelli and Pagano, 2002). The aims of both projects are similar, but the approach differs in that OpenDLib uses a transport layer that is composed of custom protocols layered over SOAP (Box, et al., 2000). Lessons learnt from both projects can ultimately lead to the creation of a standardised component model.

More recently, the ODL project has led to the DL-in-a-Box project (Luo, et al., 2003), that provides ongoing support for components and is looking into the application of components to practical use cases. Also, the Open Component-Based Knowledge Hypermedia Applications Management project (OCKHAM) (OCKHAM, 2003) has started to look into the requirements for open reference models, catalysed by the ODL framework and related work from a theoretical perspective. Specifically, OCKHAM has suggested applying the REST model (Fielding and Taylor, 2002), an emerging theoretical framework that models and explains the success of the World Wide Web. The REST model formalises the simplicity inherent in the design of HTTP (Fielding, et al., 1999), explaining its widespread success and providing a basis for the design of future protocols. This directly vindicates and informs the approaches taken by OAI, ODL and, to a degree, OpenDLib.

Thus, the stage has been set. Models have been proposed to support flexible digital libraries, and simplicity of components has proven to be popular, especially in the context of the OAI-PMH. The natural next step is to investigate methods of creating complete digital libraries from components in a simple and flexible manner. Digital libraries are no longer filled with mystique – instead, each provides a subset of well-known services to users. These services should be part of a palette of tools available to system designers with widely differing needs and in various contexts. While such simplicity in system design is universally useful, it is vital to support digital library efforts in countries where resources are scarce and solutions need to be focused and efficient.

Research into the composition of components to build digital libraries is of practical importance to local digital library efforts while simultaneously satisfying multiple themes of the research agenda defined by NRF in its ICT programme. Building distributed systems from components is directly related to the "software customisation and integration" theme and since components are distributed over the Internet, this work is also related to the "Internet and mobile application" theme. Componentised systems also benefit greatly from "standard and quality deployment" as became evident during the OAI-PMH design, where tools for standardisation where among the most important factors contributing to the protocol's success. Specifically, work done on the Repository Explorer to test for protocol-level compliance of OAI components (Suleman, 2001; Suleman, 2003) has already been generalised but future work on specification-based testing will potentially benefit not just digital libraries but all distributed component models, including those based on Web Services and SOAP (Box, et al., 2000).

In terms of telecommunications and networking, componentised digital libraries contribute to most of the identified research themes, if not all. Protocol design was a key part of developing components distributed over the Internet and such protocols must be evaluated further for their ability to support overall system design and configuration in addition to inter-component communication. Local conditions of poor network connectivity impact on such experimental protocol design to encourage minimalist approaches and robustness of algorithms. The OAI-PMH, while achieving much widespread acceptance in the developed world, has had almost no support from developing nations and there has been no known history of open componentised digital libraries. As a consequence, much of the design and algorithms that are related to such protocols have assumed the existence of broadband network connections, which is far from the reality in developing countries. This work will provide a much-needed mechanism to test the robustness of common practices in networked digital libraries and will establish new practices for communicating among and within systems with realistic expectations. Lastly, this work will contribute to the areas of distributed systems and services since the component-based digital libraries under examination are network-distributed services. Investigations into aspects such as scalability and robustness of components generalise to other distributed systems, while services within the context of digital libraries are driven by user needs in a networked environment so include the full gamut of possibilities.

Finally, digital libraries as networked information systems fall within the research theme of "human-information interaction", especially the "use, storage, retrieval and sharing of information". The services provided by a digital library focus largely around the retrieval aspect while interoperability among systems is a direct realisation of the "sharing of information" research area. This sharing occurs at multiple levels within a component-based digital library: at the extremities where information is shared with external systems, and within the system where information is exchanged among independent components as specified by the component model.

From a developmental angle, digital libraries fundamentally change the landscape of access to information and therefore affect the quality of life for all. This work, by promoting simpler models for digital libraries, will make them more accessible to archivists working in resource-poor conditions. This, in return, makes information more readily available to the ordinary student, teacher or researcher (within the constraints of basic Internet access), levelling the playing fields by removing access to information as a barrier to learning.

References

arXiv (2003), Physics Pre-Print Archive. Website http://www.arxiv.org

Anan, Hesham, Xiaoming Liu, Kurt Maly, Michael L. Nelson, Mohammad Zubair, James C French, Edward A. Fox and P. Shivakumar (2002), "Preservation and transition of NCSTRL using an OAI-based architecture", in Proceedings of the Second ACM-IEEE Joint Conference on Digital Libraries, Portland, OR, USA, pp. 181-182.

Borgman, Christine (1998), From Gutenberg to the Global Information Infrastructure, MIT Press.

Box, Don, David Ehnebuske, Gopal Kakivaya, Andrew Layman, Noah Mendelsohn, Henrik Frystyk Nielsen, Satish Thatte and Dave Winer (2000), Simple Object Access Protocol (SOAP) v1.1, W3C, 8 May 2000. Available http://www.w3.org/TR/SOAP/

Castelli, Donatella and Pasquale Pagano (2002) "OpenDLib: A Digital Library Service System", in *Proceedings of 6th European Conference on Research and Advanced Technology for Digital Libraries* (ECDL2002), Rome, Italy, 16-18 September 2002, pp. 292-308.

DELOS (2001) *Digital Libraries: Future Directions for a European Research Programme*, San Cassiano, Alta Badia, Italy, 13-15 June 2001. Available http://delos-noe.iei.pi.cnr.it/activities/researchforum/Brainstorming/brainstorming-report.pdf

Fielding, R., J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee (1999), *RFC2616: Hypertext Transfer Protocol – HTTP 1.1*, Network Working Group, June 1999. Available ftp://ftp.isi.edu/in-notes/rfc2616.txt.

Fielding, Roy T., and Richard N. Taylor (2002), "Principled Design of Web Architecture", in ACM Transactions on Internet Technology, Vol. 2, No. 2, May 2002, pp. 115-150. Available http://www.ics.uci.edu/~taylor/documents/2002-REST-TOIT.pdf

Gladney, H., Z. Ahmed, R. Ashany, N. J. Belkin, E. A. Fox and M. Zemankova (1994), "Digital Library: Gross Structure and Requirements", *Workshop on On-line Access to Digital Libraries*, June 1994.

Lagoze, Carl, and Herbert Van de Sompel (2001), "The Open Archives Initiative: Building a low-barrier interoperability framework", in Proceedings of the ACM-IEEE Joint Conference on Digital Libraries, Roanoke, VA, USA, 24-28 June 2001, pp. 54-62.

Lagoze, Carl, Herbert Van de Sompel, Michael Nelson and Simeon Warner (2002), *The Open Archives Initiative Protocol for Metadata Harvesting – Version 2.0*, Open Archives Initiative, June 2002. Available http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

Luo, Ming, Geoff Filippi and Edward A. Fox (2003), Digital Libraries in a Box. Website http://dlbox.nudl.org/index.html

NCSTRL (2003) Networked Computer Science Technical Reference Library. Website http://www.ncstrl.org

OAI (2002), *Open Archives Initiative*. Website http://www.openarchives.org

OCKHAM (2003), Open Communities for Digital Library Development. Website http://ockham.library.emory.edu/index.php

OpCit (2003), E-Prints. Website http://www.eprints.org/

Smith, MacKenzie, Mary Barton, Margret Branschofsky, Greg McClellan, Julie Harford Walker, Mick Bass, Dave Stuve and Robert Tansley (2003), "DSpace: An Open Source Dynamic Digital Repository" in D-Lib Magazine, Vol. 9, No. 1, January 2003. Available http://www.dlib.org/dlib/january03/smith/01smith.html

Suleman, Hussein (2001), "Enforcing Interoperability with the Open Archives Initiative Repository Explorer", in *Proceedings of the ACM-IEEE Joint Conference on Digital Libraries*, Roanoke, VA, USA, 24-28 June 2001, pp. 63-64.

Suleman, H. (2002), Open Digital Libraries, Ph.D. dissertation, Virginia Tech. Available http://scholar.lib.vt.edu/theses/available/etd-11222002-155624/

Suleman, Hussein (2003), *OAI Repository Explorer*. Website http://purl.org/net/oai_explorer

Suleman, Hussein, and Edward A. Fox (2001), "A Framework for Building Open Digital Libraries", in *D-Lib Magazine*, Vol. 7, No. 12, December 2001. Available http://www.dlib.org/dlib/december01/suleman/12suleman.html

Suleman, H., and E. A. Fox (2002), "Designing Protocols in Support of Digital Library Componentization", in *Proceedings of 6th European Conference on Research and Advanced Technology for Digital Libraries* (ECDL2002), Rome, Italy, 16-18 September 2002, pp. 568-582.

Witten, I. H., R. J. McNab, S. J. Boddie and D. Bainbridge (2000), "Greenstone: A Comprehensive Open-Source Digital Library Software System", in *Proceedings of Fifth ACM Conference of Digital Libraries*, San Antonio, Texas, USA, 2-7 June 2000, pp. 113-121.

Research Aims

This proposal outlines a course of research into the issues faced by digital library designers who wish to adopt component models for their systems. Specifically, the following questions will be investigated:

1. Is it possible to compose (i.e., connect together) components using familiar visual interfaces such that the complexity of the process is understandable, manageable and ultimately decreased? Further to elaborating the issues involved and testing the usability and understanding of such visual interfaces by typical users, answering this question will also provide tools for building systems in experimental environments (and possibly for adaptation to production environments).
2. What support structures will be needed to encourage the adoption of an open component model? It is tractable to devise general-purpose specification-driven testing tools for components, and if so, what are the problems and limitations of such generalisations? How is registration of components handled for a system that is distributed over a network (e.g., using Web Services)? How are fault tolerance and load balancing included to address scalability and robustness of such systems? These questions aim to address the additional interfacing concerns introduced by splitting a system into multiple sub-systems.
3. How can digital library software be pre-packaged for different scenarios without compromising the flexibility of the model? This is of vital importance in developing countries such as South Africa where digital library technology is new and adoption is a hurdle because of the high start-up and maintenance cost for systems which are not "right-sized".
4. How is flexibility handled at the level of the user interface? Is it possible to describe the user interactions with the system using an interface language such that changes to the interface can be accomplished through visual tools? Is it possible to share descriptions of the functionality of digital libraries using declarative specifications? Can such declarative specification languages be used to model existing systems and thereby subsume existing methods (largely, hard-coded programs) for specifying how users interact with systems.

In summary, this research is aimed at addressing the problems introduced by dividing what is conceptually a single system into individual components. The expected outcome of this work is a set of proven guidelines and experimental tools to move the digital library community closer to an ideal of simple yet flexible digital library architecture.

A subsidiary aim of this project is to set up real-world testbeds of digital resources that will directly benefit researchers while indirectly providing an experimental platform for this and future digital library endeavours. Experiences from past and related initiatives at other institutions has shown that such a process in itself encourages researchers not directly related to the project to manage, archive and make their work accessible – while not being an aim of this research, this will be an expected and welcomed side-effect.

Research Design and Framework

The methodology for this work has two identifiable aspects: those elements applicable to all of the sub-questions that will be addressed; and elements that relate to each individual major issue. Thus, these will be treated as separate sections:

1. Overall research design

1.1. Data collection

As an enabling mechanism for the other aspects of this project, data collections need to be established in the form of local archives. These will be obtained by working with researchers to archive their publications in subject repositories recommended by their individual communities e.g., by setting up local nodes for the NCSTRL project, which already has more than 150 participants worldwide, notably excluding South African universities. In addition, institutional repositories for documents such as electronic theses and dissertations are becoming popular (South Africa has at least three members in the international Networked Digital Library of Theses and Dissertations (NDLTD) (http://www.ndltd.org) and it is proposed that these will be buttressed by additional components to make them interoperable with our and other efforts. In all of these cases, the work involves consultation with staff and the use of standard or emerging tools in the digital library community – no new software development will be necessary. The collections built over time will serve a two-fold purpose: the data can be used by experimental systems and the researchers – who would have had exposure to digital libraries – can provide a base group for evaluation of systems. Essentially, this nurtures a tradition of digital library use and understanding, which is lacking in South Africa and must be developed to support research and development efforts.

As a second front to data gathering, the OAI Protocol for Metadata Harvesting, which is pivotal to the component model, can be used to obtain data from remote sources. The OAI maintains a registry of archives currently sharing their collections of metadata and/or data freely with external entities. This includes notable organisations such as the US Library of Congress and collections such as OCLC's WorldCat – a collection of 4 million pointers to dissertations. Such data, or subsets thereof, can readily be obtained as and when necessary, over the Internet, in order to supplement and complement local collections and to expose local audiences to international resources and vice versa.

1.2. Specialized equipment and infrastructure

As one of the aims of this project, the techniques evolved must be applicable to a wide range of systems, from large institutional, national and international repositories to small collections of important resources. However, recognising that very large projects have the personnel and finances to develop custom solutions, this project is aimed primarily at defining the architecture of digital libraries that operate with tight constraints on resources. Thus, there is no unconventional equipment required. Instead, medium- to low-end PC servers using open source software are becoming the norm in the digital library community and these are all that is needed. Multiple servers are needed to test inter-component interaction over a network and to separate highly experimental pieces of software from stable archives that serve the purpose of evaluation in a production environment. As much of the work in this project is designed to be conducted by postgraduate students working independently, commodity workstations are required for each of those researchers.

Some funding has already been obtained from UCT for at least the first server to host experiments with components and other digital library technology in the immediate future (prior to any possible funding from NRF or other agencies). Also needed for all servers and workstations is a local network infrastructure and a realistically fast external Internet connection. The local computer science department provides 10Base-T connections for local machines, with a gradual upgrade to 100Base-T. Past experience has indicated this more than suffices for digital library experimentation. The external Internet connection is not fast and neither is it reliable – this thus provides a perfect experimental environment to test the robustness of algorithms and software.

1.3. Component Framework Design and Systems Integration

Primary responsibility for further developing the component framework to adhere to current Internet and digital library standards will rest with the principal investigator. This includes the

dynamic process of re-design, ongoing maintenance and testing of the ODL framework, in collaboration with local and international collaborators, in response to emerging requests for changes. A component suite will be maintained to support all experiments by students and collaborators associated with the project. Interest in the ODL framework from external parties has spurred discussion that may lead to the standardisation of inter-component interaction in the long term. The principal investigator will participate in and initiate such activity as and when an appropriate body of research is available to substantiate a need for standardisation.
In addition, while the research question has been divided into sub-questions aimed to define and encourage student participation in the larger project, integration among all of these parts is essential to prove the viability of the overall approach. This integration will be led by the principal investigator, to oversee the various parts and maintain a consistent and workable model as time progresses.
Dissemination is usually considered to be a separate activity post-experimentation, but in the context of networked digital libraries, it is a primary need since collaborators frequently provide support for experiments with remote systems. The principal investigator will encourage dissemination and actively seek collaborators to help with experimental validation of the approach taken by this project.

2. Specific issues

2.1. Visual Component Composition

Based on current practices in visual environments, a toolkit for connecting modules will be developed and applied to the problem of component composition in digital libraries. This will be based on a client-server architecture, where a client is used to design a typical digital library that is then instantiated on a server. The system will use existing components from the ODL project as its base, in order to ratify the existing body of work and suggest changes where necessary to support remote configuration in addition to remote use of components. As part of this work, the following aspects need to be defined and/or developed:

- a client digital library design tool
- a server daemon to communicate component descriptions to the client
- a language to describe the configuration of individual components
- a language to describe a particular network of components that form the basis for a specific digital library system

To test for complexity and understanding of the approach, user tests will be conducted to evaluate the ease of use of component composition. In addition the expressive ability of such a system will be tested by attempting to model well-known digital library toolkits using tool developed in this project.
This "visual component composition" element of the larger project is already being addressed in part by 3 students: 2 Honours students are building a basic GUI for component-based systems, which is being applied to the digital library design problem by a Master's student, all of whom have begun work on the project in 2003. With additional funding, the Master's student will continue this work into 2004, and chart out future work to be addressed beyond that.

2.2. Remote Component Management

With an increase in independence of components, there is a corresponding increase in management responsibilities. Building on experiences in developing validation tools for the W3C (Web Characterization Repository – http://repository.cs.vt.edu) and the OAI (Repository Explorer – http://purl.org/net/oai_explorer), tools will be designed and built to support general testing of component interfaces. Some initial work was done on extending/generalizing the Repository Explorer (which tests for adherence to the OAI-PMH protocol) but further work is envisaged to make testing a specification-driven process, based on formal descriptions of high-level protocols such as that used by OAI and ODL. This work will include the development or adaptation of a language for protocol specification and the specification of sequences of protocol interactions,

followed by the creation of reference implementations of tools that can understand such specifications and test one or more components based on it. Tests will be conducted with installations of components, either as individual pieces or as parts of larger systems. In addition, the generality of such an approach will be evaluated for its applicability to and implications for the testing of Web Services and Web-based services in general.
This work will be carried out in conjunction with a single Master's student in 2004-2005, possibly in conjunction with a group of Honours students. Substantial responsibility for the design and evaluation will be delegated to the Master's student, under supervision of the principal investigator, with Honours students focusing mainly on the implementation of tools.
Additional concerns related to the management of remote components include registration mechanisms for components to support remote configuration, load balancing and possibly a form of process migration. These issues are of sufficient size and adequate complexity to serve as Honours projects, targeted at 2004 and beyond, as registration and robustness concerns can be best addressed once a critical mass has been established in terms of data and experimental systems.

2.3. User Interface Descriptions

The user interfaces of most digital libraries are fixed, or configurable from the perspective of appearance but not workflow. With the introduction of a flexible model for the underlying or back-end components of a digital library, the user interface can change in terms of both aspects. This part of the project will look into creating a representation of user interaction as a specification that can easily be changed with the changing architecture of the back-end system. This specification will be modelled upon a finite state machine – some work in the existing literature on user interfaces indicates the feasibility of this approach. Thus, changes to the state of the system will be modelled as the result of user interactions (or system-level interactions), resulting in specified component interactions and transitions in the user interface. Various emerging XML standards promise to be enabling technologies and these will be investigated. To test the specification in practice, an "engine" will be constructed to provide a digital library interface between users and component-based services.
This approach will be compared to approaches taken in existing systems, as well as emerging technology for Web interfaces e.g., portals. Such interface descriptions will be applied to model existing and new digital library projects, thus evaluating the generality of the specification and driver combination. In addition, qualitative feedback will be obtained on the design approach and implementation(s).
Ultimately, this aspect of the project will be combined with the components composed in the "Visual Component Composition" aspect to create complete digital library systems encompassing both user and system elements.
This specification language and user interface generalisation is intended as a Master's student project, to begin in 2004 with a prototype implementation for proof-of-concept, followed by a comprehensive and complete language specification, modelling and testing in 2005.

2.4. Packaging and Transparency

In order for component-based digital libraries to prove their practical benefit to society at large, it must be possible for them to be deployed as installable software packages. This aspect of the larger project will look into how the components, after being composed into systems, can be integrated into single packages based on succinct descriptions. It will involve a study of:

- how packages can be created and shared,
- how individual pieces can be upgraded in future,
- how the architecture can be modified as needed, post-deployment,
- how the individual configurations of components can be coalesced to drive a global configuration mechanism, and
- how existing systems can be modelled as packages involving components.

A packaging system will be created to bundle components together based on digital library descriptions, and to create installable versions thereof. This will be informed by and contribute to best practices on issues such as resolving of dependencies, as exemplified by the myriad package managers currently available for open source operating systems.
These packages will be tested by users who will be surveyed to measure ease of use, modifiability and levels of maintenance.
Bootstrapping techniques will also be looked into. Experiments will be conducted to assess if distributing minimal systems, which then download and install components as needed, can be applied in this scenario.
This work is aimed at a level suitable for a single Master's student, with elements of design, implementation and evaluation to be conducted independently, with appropriate supervision from the principal investigator. Initial work in the early part of 2005 will include a study of existing systems, followed by implementation of packaging tools in late 2005 and early 2006. This is expected to be followed by an evaluation by peers both locally and in the digital library community at large.

Progress to Date

The principal investigator has worked extensively on the initial development of the existing ODL component framework as part of his Ph.D. research. This has included the design of numerous protocols, implementation of reference components, development of test systems (albeit hard-coded), and initial evaluation of the usefulness/simplicity of components and the feasibility of inter-component interaction from a performance perspective. The developers of various digital library systems have consulted with him in order to build ODL components into their systems, for prototyping and/or production use. This proposal is motivated largely due to the difficulty experienced by collaborators in integrating such components to form complete systems.
As a member of the OAI Technical Committee, he has contributed to the design and testing of the OAI Protocol for Metadata Harvesting (OAI-PMH) (versions 1.0, 1.0 and 2.0). He developed and actively maintains the Repository Explorer, a Web-based tool to test for compliance with the OAI-PMH. This is a precursor to some aspects of the proposed work and the experiences gleaned in past efforts will inform future endeavours. He also has been instrumental in popularising the OAI-PMH by developing and distributing templates and toolkits for implementation of the OAI protocol on clients and servers.
In a largely consultative role, he has assisted with the design of networked digital libraries related to the following projects:
-Networked Computer Science Technical Reference Library (NCSTRL, www.ncstrl.org),
-Networked Digital Library of Theses and Dissertations (NDLTD, www.ndltd.org),
-Computing and Information Technology Interactive Digital Educational Library (CITIDEL, www.citidel.org),
-AmericanSouth.org (americansouth.org)
-iLumina (www.ilumina-dlib.org)
In addition, he has been a lead developer/maintainer of the Computer Science Teaching Centre (CSTC, www.cstc.org) and the Web Characterisation Repository (repository.cs.vt.edu).
Various publications have emerged from the work on networked digital libraries and component-based models for digital libraries, as listed in this document.

Potential Impact on HR Development

This proposal is specifically designed to encourage the training of students in the principles and techniques involved in building modern networked digital libraries. This is a critical skill needed in the IT community to develop and enhance future infrastructure for information sharing. While postgraduate students will benefit directly, their experiments will involve scores of researchers and

fellow students who will gain a better understanding and appreciation for digital libraries by exposure to the technology.
This project also aims at promoting, directly and indirectly, collaboration among researchers through the medium of digital libraries. Such collaborations enhance the quality of individual research works, while contributing to a sense of community among researchers.

Potential Impact on Redress and Equity

Digital libraries, by their very nature, address imbalances in the information arena by making information more accessible to regular users. This is especially relevant in the area of journal publications. These are traditionally expensive to obtain and therefore only easily available to universities that subscribe at high costs to themselves, the state and their students/academics. One of the main aims of digital library research is to make it possible for academics to archive, review and publish work without the need for expensive intermediaries such as publishing houses. Even if such works are already published, simple digital libraries can be used to make the publications available locally, as is allowed by many publishers already (e.g., ACM, Springer). This project aims to make the technology accessible to researchers so that they, in turn, may make the products of their research accessible to the larger research community.
Component-based systems are important to support research into online information systems because they provide the mechanism for highly specialised projects to be incorporated into larger systems. This allows researchers in information systems at smaller universities to easily contribute to digital library (and related) research without the need to first develop their own basic tools and frameworks.
Eventually, digital libraries are the mechanism to get high quality information about Africa out to the rest of the world, and in the reverse direction, for us to obtain high quality electronic resources. Further, this will be possible without having to develop massive software systems if our models, and the models adopted by the rest of the world, are based on simplicity.

Potential Outcomes

This project has many potential outcomes for the research community in digital libraries, the research community in general and the archiving and library community.
The research community in digital libraries is made up of widely differing projects with similar or related aims. By providing a simple and workable framework to integrate digital library services and data collections, blurring the boundaries between single digital libraries and highly distributed systems, this project will hopefully facilitate greater interaction among researchers. Experimental systems can then be built to combine unique and highly focused services from different researchers. In addition, such systems can trivially contain the basic services so much time will be saved by avoiding reinventing the wheel, as is currently done by most research groups.
By providing a model for building digital libraries that are more flexible and simpler to understand and maintain, this project will potentially affect the acceptance rate of such technology. Thus, more institutions and archivists will be able to make organised collections available, in general bolstering the research community by making access to information easier. This can go beyond the research community to regular man-on-the-street Internet users who seek information on particular topics. If such information is available, simple and flexible digital libraries make it possible for those who possess the information to easily organise their collections and make them accessible.
This project also has implications for the library and archiving community in terms of cost savings. Modular programming embodies a spirit of reuse and this is currently lacking in the digital library community. Rather than develop software from scratch for every changing need, a standard suite of components and standard protocols for use of those components will greatly assist libraries in making collections available online – this work potentially removes the remaining barrier of how to

compose those components into complete systems with user interfaces and associated tools to maintain the systems.
In terms of artefacts, this project will contribute a suite of tools that may be used to design digital libraries or that may be adapted to similar projects in distributed systems and/or Web Services. This includes tools to design back-end distributed systems, tools to test component interfaces and tools to design and modify user interfaces. Software packages produced as part of the work also may be used in research environments to make local resources accessible to local and/or international audiences. Finally, the collections developed during this project will serve as testbeds for future digital library experimentation locally and abroad, where they will encourage future international efforts to acknowledge and cater for the needs of countries such as South Africa.
Ultimately, by advancing the science of building digital libraries, this project hopes to make information more accessible to users. In South African society, where the cost of information is abnormally high, any move towards digital libraries makes it easier for people to access information that would normally be beyond their reach (philosophically and physically). A typical example of this is the Networked Digital Library of Theses and Dissertations (NDLTD), which is making it possible for students and researchers to obtain electronic copies of theses and dissertations online. This was previously only possible through inter-library loan. Now, anybody anywhere in the world can find and get access to a thesis if it is part of a member digital library at a participating institution. NDLTD achieves this by using some components developed as part of the initial ODL development, and is therefore a suitable representative for the breadth of opportunities made possible by the adoption of a simple and flexible digital library model.

Research Outputs

Books

Suleman, H. (2002), *Open Digital Libraries,* Ph.D. dissertation, Virginia Tech. Available http://scholar.lib.vt.edu/theses/available/etd-11222002-155624/

(See Journal Articles for an article co-published as a book chapter)

Journal Articles

… Suleman, H. and E. A. Fox (2003), "Leveraging OAI Harvesting to Build a Union Catalog", to appear in *Library Hi-Tech*, Emerald Publishing.

Suleman, H. and E. A. Fox (2001) "The Open Archives Initiative: Realizing Simple and Effective Digital Library Interoperability", *Journal of Library Administration*, 35(1/2), pp. 125-145, November 2001, Haworth Press Inc, New York. Co-published in *Libraries and Electronic Resources*, edited by Pamela Higgins, Haworth Press Inc, 2001.

Conference Papers

… DRTC …

Fox, Edward A., Hussein Suleman, and Ming Luo (2002) "Building Digital Libraries Made Easy: Toward Open Digital Libraries", in *Proceedings of ICADL 2002*, pp. 14-24.

Suleman, H. and E. A. Fox (2002) "Designing Protocols in Support of Digital Library Componentization", *6th European Conference on Research and Advanced Technology for Digital Libraries* (ECDL2002), Rome, Italy, 16-18 September 2002.

Suleman, H. and E. A. Fox (2002) "Towards Universal Accessibility of ETDs: Building the NDLTD Union Archive", *Fifth International Symposium on Electronic Theses and Dissertations* (ETD2002), Provo, Utah, USA, 30 May-1 June 2002.

Suleman, H. (2001) "Enforcing Interoperability with the Open Archives Initiative Repository Explorer", *Proceedings of the First ACM-IEEE Joint Conference on Digital Libraries*, Roanoke, Virginia, USA, June 2001, pp. 63-64.

Suleman, H., E. A. Fox, and M. Abrams (2000) "Building Quality into a Digital Library", *Proceedings of the Fifth ACM Conference on Digital Libraries*, San Antonio, Texas, USA, June 2000, pp. 228-229.

Reviewed Conference Demonstrations

Fox, E. A., R. K. France, M. A. Gonçalves, and H. Suleman (2001) "Building Interoperable Digital Library Services: MARIAN, Open Archives and NDLTD", in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, Louisiana, USA, September 2001, p. 451.

Suleman, H. (2001) "Using the Repository Explorer to Achieve OAI Protocol Compliance", in *Proceedings of the First ACM-IEEE Joint Conference on Digital Libraries*, Roanoke, Virginia, USA, June 2001, p. 459.

Fox, E. A., R. K. France, M. A. Gonçalves, H. Suleman, and M. H. Lee (2000) "Building a Unified Digital Library for Theses and Dissertations", *3rd International Conference of Asian Digital Library*, Seoul, South Korea, December 2000.

Suleman, H., E. A. Fox, and M. Abrams (2000) "Building Quality into a Digital Library", *Fifth ACM Conference on Digital Libraries*, San Antonio, Texas, USA, June 2000.

Technical Reports

Suleman, Hussein and Fox, Edward A. (2002) *Beyond Harvesting: Digital Library Components as OAI Extensions*, Technical Report TR-02-25, Department of Computer Science, Virginia Tech. Available http://eprints.cs.vt.edu:8000/archive/00000625/

Magazines Articles

Suleman, H. and E. A. Fox (2001) "A Framework for Building Open Digital Libraries", in *D-Lib Magazine* 7(12), December 2001. Available http://www.dlib.org/dlib/december01/suleman/12suleman.html

Suleman, H., A. Atkins, M. A. Gonçalves, R. K. France, E. A. Fox, V. Chachra, M. Crowder, and J. Young (2001) "Networked Digital Library of Theses and Dissertations: Bridging the Gaps for Global Access - Part 1: Mission and Progress", in *D-Lib Magazine* 7(9), September 2001. Available http://www.dlib.org/dlib/september01/suleman/09suleman-pt1.html

Suleman, H., A. Atkins, M. A. Gonçalves, R. K. France, E. A. Fox, V. Chachra, M. Crowder, and J. Young (2001) "Networked Digital Library of Theses and Dissertations: Bridging the Gaps for Global Access - Part 2: Services and Research", in *D-Lib Magazine* 7(9), September 2001. Available http://www.dlib.org/dlib/september01/suleman/09suleman-pt2.html

Tutorials

Suleman, H. (2002) "Building Interoperable Digital Libraries: A Practical Guide to Creating Open Archives", half-day tutorial at the *Second ACM/IEEE Joint Conference on Digital Libraries*, Portland, Oregon, USA, 14-18 July 2002.

Suleman, H. (2002) "Building Interoperable and Accessible ETD Collections: A Practical Guide to Creating Open Archives", *Fifth International Symposium on Electronic Theses and Dissertations* (ETD2002), Provo, Utah, USA, 30 May-1 June 2002.

Suleman, H. (2001) "Building Interoperable Digital Libraries: A Practical Guide to Creating Open Archives", half-day tutorial at the *First ACM-IEEE Joint Conference on Digital Libraries*, Roanoke, Virginia, USA, June 2001.

Workshops

Co-organised *Open Archives: Communities, Interoperability and Services*, held in conjunction with ACM SIGIR 2001, New Orleans, September 2001.

Co-chaired *Extending Interoperability of Digital Libraries: Building on the Open Archives Initiative*, held in conjunction with Fourth European Conference on Research and Advanced Technology for Digital Libraries, Lisbon, Portugal, September 2000.

Co-chaired *Extending Interoperability of Digital Libraries: Building on the Open Archives Initiative*, held in conjunction with ACM Hypertext'2000 and ACM Digital Libraries'2000, San Antonio, June 2000.

Standards Body Participation

Open Archives Initiative (OAI) Technical Committee
Period: September 2000 – June 2002
Role: Contributed to developing versions 1.0, 1.1 and 2.0 of the OAI Protocol for Metadata Harvesting, a network protocol for transferring metadata between digital libraries. Provided community support by implementing and disseminating software tools for development and testing.

Selected Software Tools

Web Characterization Repository, 1998-1999
http://repository.cs.vt.edu
This is a digital library of data sets, tools and publications relevant to the characterization of the WWW. Additional tools were created to specify the format of and validate Web log files. There were developed in the conjunction with the Web Characterization Activity working group of the World Wide Web Consortium.

Repository Explorer, 2000-present
http://purl.org/net/oai_explorer
This is a testing tool to enforce correctness of implementation of the OAI Protocol for Metadata Harvesting by prospective data providers. It was developed in consultation with the Open Archives Initiative's Technical Committee.

OAI Data Provider Tools, 2000-present
http://www.dlib.vt.edu/projects/OAI/
Tools were developed on an ongoing basis to enable experimentation and help developers to bootstrap implementations of the latest standards. The most widely used of these tools are:
- ETDdb extensions, to make an electronic thesis and dissertation archive into a Open Archive
- XMLFile, to create an Open Archive out of XML files
- VTOAI, a Perl template for OAI data provider implementations
- OAI/ODL Harvester, a protocol version-independent harvester for Open Archives

Open Digital Library (ODL) Components

http://oai.dlib.vt.edu/odl/

Various typical digital library services were cast as Open Digital Library components and reference implementations and sample user interface code was developed for each. This includes components to perform each of: searching, browsing, annotation, rating, recommendation, submission and peer review.