Contents

WDL (World Digital Libraries)
Volume 3, Issue 1 • June 2010

ν Editorial

vii Issue editorial

1 The German National Library and digital preservation: challenges and opportunities
Reinhard Altenhöner

11 Cross-institutional cooperation on a shared bit repository
Eld Zierau and Ulla Bøgvad Kejser

23 Piloting an institutional repository at a research-intensive university: strategies for content recruitment and the role of the library
A Abrizah

41 Ontology-based visualization and navigation in an online digital library
Surjeet Mishra and Hiranmay Ghosh

51 Harvesting capability of Google scholar: a comparative study of three major journal lists—Directory of Open Access Journals, Highwire, and BioMed Central
Rosy Jan and Nadim Akhtar Khan

63 A new framework to preserve Tagore songs
Chandan Misra, Baidurya Bhattacharya, and Anupam Basu

73 Diatomscapes exposé: how faculty and digital librarians collaborate to promote and preserve the passion of research (CP3R) for digital future
Plato L Smith II

85 Book review
Digital Information: order or anarchy?
Hazel Woodward and Lorraine Estelle (Editors)
Reviewed by Shantanu Ganguly

87 News

93 Forthcoming events
Editorial

N BALAKRISHNAN
Indian Institute of Science, Bangalore

There are several challenges we come across in our endeavour to provide effective access to and use of scientific data and information in the developing world. These are over and above the difficulties associated with the information and communication technology (ICT) infrastructure. Although these challenges are serious, they can be overcome through sustained focus and joint action. In particular, there are numerous examples of proven and successful approaches that can usefully be adopted to solve most, if not all, of the recognized problems. For instance, different developed and developing countries have established a variety of novel and effective mechanisms to reduce barriers and promote the production, access, and use of digital scientific information. Models such as open-source software, federated open data networks, open-access journals, and collaborative websites are becoming increasingly accepted and useful tools for the advancement of public research and education. One particularly important mechanism – managing open institutional repositories – is the focus of this document. Digitally networked open institutional repositories for scholarly literature – for articles from refereed journals, as well as the grey literature and other research output – are coming to be recognized as a fundamental component of the scientific and educational systems, and of the broader knowledge society, in many countries.

It greatly enhances and facilitates scientific inquiry, diversity of analysis and opinion, new types of research, and methods of analysis, and generally prevents duplicative work and research inefficiencies generally. It eliminates, to a large extent, economic barriers on the transfer of information to researchers and students who are unable to afford high (or any) access cost. Overall, the different forms of online open access promote interdisciplinary and international research, particularly in integrating scientists in the less economically developed countries into the global research system. It truly democratizes access to information and tends to provide a level playing field for innovation.

The experts in the field of information sciences attempt to address some of these challenges in this issue. Abrizah, in her paper, gives a complete overview of her pilot study conducted at a research-intensive university. She focuses on the strategies to populate the IR, motivating the researchers to self-archive papers, capture pre-existing collection of theses and dissertations, and build the institutional repository.

Zierau and Kejser explore the usage of the Open Archival Information Systems (OAIS) reference model to provide the framework for systematically analyzing an institution’s technical and organizational requirement for a remote bit repository. The study further finds
that the institutions have varying requirements, for example, for safety, accessibility, and confidentiality of enormous resources.

The Nobel Laureate Rabindranath Tagore’s Rabindra Sangeet, published in Geetanjali, is among the most famous of Indian literary creations. Misra, Bhattacharya, and Basu have given a brief description of the new framework and a simple application for archiving Tagore’s songs in MIDI.

Altenhoner’s paper highlights the latest developments in the German National Library (Deutsche Nationalbibliothek, DNB), which was initiated in the 1990s. Mishra and Ghosh present, in their paper, a novel approach to visualize and navigate through the document collection of a DL using ontology. They present an abstract view of the knowledge map in a clear and concise manner.

While Smith, in his research paper, opts for an experimental approach of collaborative open access institutional repository digital collection building and digital preservation strategy for Diatomscapes, which is a small representative sample image of biological silica, Jan and Khan present a detailed study of Google Scholar’s credibility and harvesting capability by submitting select titles from the field of microbiology in three major, renowned, and open access journal lists.

We do hope that the readers would benefit from the scholarly articles presented in this issue and compare with their own experiences in facing the challenges in the world of digital information.
Issue editorial

SHANTANU GANGULY
Issue editor, World Digital Libraries

‘Digital Libraries’ can mean many things, but we consider them to be libraries first and foremost, built upon the enduring principles of information management, which lay at the heart of the practice of librarianship for decades. Libraries, themselves repositories, have always dealt with the management of repositories for the benefit of their users. With libraries now routinely managing repositories of various types in digital formats, does it auger well for the quality of repositories and institutions? There are considerable implications for librarians who manage digital material as full-text, where the digital items are accompanied by digital metadata.

A growing percentage of researchers have discovered the utility of the web for the dissemination of their research outputs, and have now been using it, for many years in the case of some disciplines. The library community, increasingly focused on a digital library agenda, has understood the need to ensure that the material being disseminated is managed successfully through proper description, indexing, and storage for long-term preservation. The approach, which has now proved its worth is being used for other types of material generated within institutions.

The marriage of research generated by academics, with output management by librarians, has created a new form of publication with open values, which presents a growing challenge to commercial publishers who have controlled the publication of research for many decades.

In order to address some of these challenges, The Energy and Resources Institute (TERI), in partnership with the Indira Gandhi National Open University (IGNOU) organized a mega conference – International Conference on Digital Libraries (ICDL) 2010 – the theme of which was ‘Shaping the Information Paradigm’. Held every three years at New Delhi, the four-day conference attracted 650 participants from 40 countries. Some selected papers from the conference have been published in this issue. Abdullah, in her paper, provides a comprehensive overview of the pilot study she concluded at a leading research-intensive university. She emphasizes on methods to populate the Institutional Repository (IR), in order to enable the researchers to self-archive papers and utilize the pre-existing collection of theses and dissertations. She also puts forth suggestions for new roles in content recruitment and interpreting policies.

Eld Zierau and Ulla Bøgvad Kejser study the use of the Open Archival Information Systems (OAIS) reference model to provide an overall framework for systematic analysis of the technical and organizational requirements of institutions for a remote bit repository. The study further shows that, depending on the collections
they hold, institutions have expressed varying needs for safety, accessibility, and confidentiality of enormous amounts of data and information.

Indian music and culture is regarded highly throughout the world, especially Rabindra Sangeet, which was composed and published by Nobel Laureate Rabindranath Tagore, is a worldwide phenomenon. Chandan Misra, Baidurya Bhattacharya, and Anupam Basu have given a brief description of the new framework and a simple application for archiving Tagore songs in Musical Instrument Digital Interface (MIDI) and notational formats.

Reinhard Altenhöner’s paper chronicles 10 years of activity and focuses on the latest developments in the German National Library or Deutsche Nationalbibliothek (DNB), established in the 1990s. DNB has the legal mandate to collect, archive, index, and enable access to digital publications.

A novel approach to visualizing and navigating the document collection of a digital library using ontology is presented by Surjeet Mishra and Hiranmay Ghosh. They further give an abstract view of the knowledge map, clear and concise, which can be progressively ‘zoomed in’ or ‘zoomed out’ to navigate the knowledge space.

Plato L Smith II, in his research paper, opts for an experimental approach of collaborative, open-access institutional repository digital collection building and digital preservation strategy for Diatomscapes—a small, representative sample image of biological silica.

Rosy Jan and Nadeem Akhtar Khan make a detailed study of Google Scholar’s credibility and harvesting capability by submitting select titles from the field of microbiology in three major, renowned, and open-access journal lists—Directory of Open Access Journals (DOAJ), Highwire, and BioMed Central.

Open Access Initiatives (OAI) and open digital repositories are the key elements for discussion across the globe. Issues that are of main concern are—interoperability, multilinguality, and digital rights management. I am sure the readers will get a new direction towards digitization of resources and institutional repositories, of institutional and national memories.
The German National Library and digital preservation: challenges and opportunities

Reinhard Altenhöner, Director – IT, Deutsche Nationalbibliothek
Adickesallee 1, D-60322 Frankfurt am Main, Germany
r.altenhoener@d-nb.de

Abstract
The increasing production and use of digital information immediately lead to the challenging task of ensuring the long-term accessibility of digital resources. The changing information technology (IT) development becomes more important for memory institutions. This is a challenge not only for institutions such as libraries, archives, and museums, but also for public administration and industrial enterprises.

Beginning in the 1990s, the German National Library (Deutsche Nationalbibliothek, DNB) started collecting and archiving digital publications. It became clearer that long-term preservation has to be done cooperatively in order to share experiences and reduce efforts. Since 2006, the DNB has the legal mandate to collect, archive, index, and enable access to digital publications published in the so-called German Web. This means that the library has to deal with a huge amount of objects. In consequence, the DNB is obliged by these new responsibilities to re-establish the workflows internally and to redefine the role and the tasks of the library within the national system of library and information services. This contribution refers to 10 years of activity and particularly highlights the latest developments in Germany.

Keywords
German National Library, digital preservation, Universal Object Format, Open Archival Initiative, Germany
Introduction
Germany as a federal country is characterized by shared power and responsibility on different regional/national levels. There are about 30 institutions with a distinctive responsibility for the legal deposit in their respective region. Keeping this in mind, it becomes clear that the libraries with a mandate to identify, collect, index, and archive digital objects have difficulty in fulfilling their task in an appropriate manner because of the lack of adequate resources and methods.

Therefore, the importance of digital preservation (DP) in the portfolio of cultural heritage organizations and data centres is being realized increasingly. DP describes the process of securing sustainable access for use and reuse of all kinds of digital data and publications. In this sense, DP comprises not only bit stream preservation but also continuous activities, starting with ingest (validation of the objects, extraction of relevant technical metadata, and so on) and is characterized by recurring interventions to migrate the object. DP is probably a broader view and has prevailed in the international debate by replacing the term long-term archive.

DP requires a large investment to meet personal and material costs. Therefore, it is clear that DP has to be carried out in a cooperative way, which integrates the experiences of dedicated communities and their feedback. One good example in Germany is the cooperative workflow with comprising than 90 participating institutions. They share the work for electronic theses and dissertations since 1998 by building a cooperative collection of more than 80 000 e-theses in the long-term archive of the DNB. Additionally, the cooperative approach was helpful because the amount of electronic publications per publication year is increasing constantly.\(^1\)

Also, in the global perspective, there exist approaches for cooperatively shared activities, for example, for format registries or automated extraction of technical meta-information. Despite the level of success of these initiatives, the principal requirement to establish trusted repositories with well-defined workflows for long-term preservation of digital objects remains. In 2003, the DNB initiated cooperative approaches with partners, for long-term preservation on a national scale. On the one hand, a national network of expertise on long-term preservation (\('\text{nestor}'\)) was initiated, and, on the other hand, a practical asset to implement an operational solution (\('\text{kopal}'\)) was initiated.

Nestor\(^2\)
The aim of the two nestor (the German network of expertise in long-term preservation and long-term availability of digital resources) project phases (2003–06, 2006–09)\(^3\) was to build a network of expertise in digital long-term preservation and to find a sustainable form of organization that allows the maintenance of the network.

The ultimate goal envisaged for the network throughout the project phases was ‘to secure the preservation of digital resources in Germany and to work with others internationally to secure global digital memory and knowledge base’. However, the intention was not for nestor to be a preservation organization in itself, or to build or plan a digital archive for Germany or anything similar. Most of the initial nestor partners were planning or implementing digital archives for their respective organizations at the time they entered into the cooperation, and nestor was created to help in exchanging knowledge, distributing best

\(^1\) \(<\text{http://www.dissonline.de/eng/index.htm}>\)
\(^2\) Thanks to my colleague Sabine Schrimpf in DNB for valuable inputs.
\(^3\) \(<\text{www.digitalpreservation.de}>\)
practices, and avoiding isolated applications. The nestor structures should foster communication between the partners and other relevant stakeholders and facilitate the development and implementation of a national cooperative digital preservation strategy. Thus, nestor is a platform to raise public awareness on digital preservation in the communities of memory institutions. It is a much broader platform in the area of science and research, especially in the sectors of finance and industry, and is also used to raise awareness amongst the general public, with special attention to politicians.

Other important missions of nestor have been to advance standardization in the area of DP and develop a concept for training, education, and instruction, together with academic partners. Synergies between grid technology and DP should be explored and targeted outreach activities for the archival community and the museum community should be organized.

When the second project phase came to an end in mid-2009, it was time to ask which of these missions were still valid in the post-project period and if there were new ones to be adopted. These questions were closely related to the planning for a sustainable, self-funded form of organization.

On the one hand, each of the partners who decided to invest their own funds in nestor expected that the investment pays back in some way or the other. Each of the involved institutions wanted to gain something from participating in the network. Important aspects were information exchange, common technology watch, sharing of best practices, and short lines of communication between the partners. On the other hand, the partners had to make clear what they, as a consortium, could offer to new partners. Consulting and guidance were important aspects there.

It has become clear that nestor, with its focus on information and communication, cannot on its own ‘secure the preservation of digital resources in Germany.’ But nestor with its network structures can, however, contribute substantially to securing digital resources.

Finally, the partners did not arrive at a completely new mission statement and a new form of organization: With six of the seven project partners and two additional organizations entering into a consortium agreement in June 2009, nestor transformed into a sustainable partner consortium and could continue its work.

The project goal of building a network of expertise was met. The cooperation agreement was an important step towards a sustainable organization. The mission remains to sustain and enlarge the network and its services and to continually integrate more expertise in order to secure digital resources in Germany.

Kopal

‘Kopal’ (co-operative development of a long-term digital information archive) started in 2004 with the mission to practically prove and implement a cooperatively built long-term preservation system used for digital publications/resources. Within kopal, the partners have developed a technological solution for long-term archiving that includes not only the archiving and bit-stream preservation of digital documents but also the implementation of preservation planning mechanisms (especially migration) for digital documents to ensure their accessibility in the future. Kopal is based on the DIAS (Digital Information Archive System) solution of IBM, originally developed for the National Library of the Netherlands. Kopal leverages the commercial system DIAS with an underlying commercial software set of IBM-standard software, which was extended especially for remote access,

enhanced metadata administration, and extended machine-readable interfaces. In the project, an open source JAVA-based software library was implemented and used for automated ingesting routines (extracting of metadata, quality control, ingest and retrieval). Additionally, object validation and metadata extraction software were integrated and amended. The kopal Library for Retrieval and Ingest, koLibRI, is therefore important as the reuse and the possibility to integrate the features in other system environments have a crucial impact on the success of the complete solution. Furthermore, the software is used to migrate defined objects in the system in an automated workflow by governing the validation and access mechanism.

In order to maintain and reconstruct the archived data, a structured archive package format, the Universal Object Format (UOF) was employed by the project partners. The UOF describes a package structure, including metadata, and is suitable both as an archive format and as an exchange format for long-term archives. The metadata comply with the standards Metadata Encoding and Transmission Standard (METS), Long-term preservation Metadata for Electronic Resources (LMER), and Dublin Core. Furthermore, other XML metadata can also be used. The kopal system automatically extracts specific technical metadata using the software tool JHOVE, which was originally developed for the journal archive JSTOR.

As of June 2006, the DNB and its project partner, the State University Library Göttingen, have integrated parts of their digital collections into the system for test purposes. In mid-2007, the project was finalized, and now the operative phase of the project has started. In a cooperation contract between the libraries, IBM, and the data centre GWDG, which drives the operational service of the archive, all partners have agreed to continue their work and to enable other institutions to join the consortium. The kopal archival system has been geared into practical use and is ready to be adopted by further partners from the library and heritage community.

As stated before, one major goal in the project was the development in a cooperative environment, which allows multiple institutions to participate at different levels of involvement. This implies that both the work sharing (data centre responsible for the operational service and the bitstream preservation, the libraries for all the aspects of preservation planning and the ongoing functional extension of the solution) and the transparent integration into existing library (or similar institutions) systems and the reusability through memory institutions play a critical role. Considering the aspect of flexible reusability, international standards for long-term archiving and metadata were adopted. In this way, both sustainability and the ability to further develop the system are guaranteed.

The DP solution needs to be embedded in the working environment and dedicated workflows, in which cultural heritage organizations collect, share, disseminate and present digital objects. Basically, in kopal, a distinction of system users was made between ‘clients’ (at different stages) and those who were responsible for the complete system. In this sense, the possibilities of reusing kopal in a productive environment reach from the complete outsourcing to the self-driven in-house system solution inside an institution. The last possibility is rather expensive in terms of funding and staffing. Therefore, a more differentiated client-oriented solution was adopted.

Clients, that is, account owners, rent an account similar to the bank accounts we are familiar with. In kopal, this account is a virtual area on top of the system that is independent of other participating institutions. This means that every participating institution has its own

---

3 Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen
4 More details in a later segment of this paper under “dp4lib”
dedicated account, which can be administered for its own purposes. In consequence, these organizations assume the responsibility of curating the digital content they collect, for example, from the Web. So the role is extended to the obligation to run the ingest service and especially the curation activities as their own responsibility. An account-holding member uses the platform and additionally he is responsible for the normalization and evaluation of data. Even the presentation of the archived objects is part of the task. Also, the planning, the conceptual preparation, and the implementation require additional steps such as the systematic migration, for example, of dangerous formats. This takes place together with other account holders and needs investment in know-how, permanent monitoring, and qualification of personal.

Clients, that is, kopal-participants, assign the cited tasks to other institutions with the status of an account owner, who is responsible for the curation of digital objects and the services requiring DP. The participants are obliged to describe the policy that should be followed in the system for their own ingested entries; they select and describe the objects for the purpose of long-term preservation. From the perspective of a participant, the solution makes sure that the amount of effort is reduced in comparison to the needs of a full archival system. At the same time, it is possible to influence the rules and regulations in the archival system, to participate on the discussions and to take over dedicated responsibilities for specific tasks in the whole process of DP in a cooperative working scenario.

This model of operation and organization describes the range of possibilities and the potential options, where long-term preservation with this dedicated background could take place. The cost model basically developed within the kopal project is dependent on the degree of measures/services the leading organization (the ‘account owner’) is willing to assign.

Advantages of this approach in sharing the tasks and the degree of responsibility are as follows: resource sharing, shared licensing costs, and optimized use of distributed knowledge.

Planning the infrastructure

Beginning in 2006/07, the discussion on DP became more intense in Germany. One important aspect of the debate affects the value and need for permanent access to research data. On the national scale, nestor, the German network of expertise in digital long-term preservation, has deployed some activities to identify the state-of-the-art technology in data preservation and has defined the needs of scientists and researchers. Although the results are preliminary to some degree, they can be considered as first steps towards operational best practice recommendations: DP activities have to be planned and carried out in a community-specific way, and a serious cooperation between different actors for the information infrastructure is important. The risk of dividing the responsibility for research data and linked publications between the different types of organizations has to be avoided.

In addition to the strategic discussions and activities on the European level, similar efforts

7 Nestor has initiated two relevant studies: nestor – materialien 6 – Langzeitarchivierung von Rohdaten / Thomas Severiens, Eberhard R. Hilf, Frankfurt am Main: 2006 (nestor – materialien 6) see <http://nbn-resolving.de/urn:nbn:de:0008-20051114018> and Anforderungen von e-Science und Grid-Technologie an die Archivierung wissenschaftlicher Daten/Jens Klump, Frankfurt am Main: 2008 (nestor – materialien 9), see <http://nbn-resolving.de/urn:nbn:de:0008-2008040103>. Results are summarized 2009 in a nestor position paper: Digitale Forschungsdaten bewahren und nutzen: publication of nestor-WG Grid/eScience and LZA, see <http://nbn-resolving.de/nbn:de:0008-2009071031>. Similar activities from DINI, the German initiative for network information with a position paper on research data, see <http://nbn-resolving.de/urn:nbn:de:kobv:11-10098082>
are undertaken on the national level. Besides some preliminary papers in the DP-community, there are different initiatives to address the needs of science and research for a sustainable infrastructure for data and long-term access on a national scale. Beginning in 2005/06, the German Research Foundation included the topic of DP into their strategic founding guidelines for the timeframe up to 2015. In parallel, the paper ‘Neuausrichtung der öffentlich geförderten Informationseinrichtungen’ has been published by a joint taskforce of experts commissioned by the national Ministry for Education and Research and the ‘Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung’ (BLK). Facing the reorientation of the whole information infrastructure, DP was only briefly addressed in a single chapter. The need to preserve electronic publications in the first place, organizational issues (like the legal deposit), international collaboration, and a first step to define national responsibilities were stated. Therefore, the goal was addressed, but pragmatic recommendations for implementation were missing. Some years later, in 2009, a new approach was started by the follow-up organization of BLK, the ‘Gemeinsame Wissenschaftskonferenz des Bundes und der Länder.’ This group finished a state-of-the-art report of the situation of technical information infrastructure in Germany, which identified gaps and general needs of science and research. In the next phase of the working plan, the relevant organizations will be identified and suggestions for organizational task sharing will be made. Some preparatory tasks were undertaken in the initiative ‘Digital Information,’ a joint activity of the leading science organizations in Germany. Set up to function till 2012, the alliance is focusing on core activities like national licensing, a national hosting strategy, research data, open access, legal frameworks, and virtual research environments (VRE). Activities of working groups concentrate on recommendations and studies, for example, for a national hosting strategy.  

An interesting step was taken to evaluate the needs of a whole campus with the German project ‘KOLAWiss’, which recently published its results. The results based on former questionnaires show the different requirements of specific sciences and the urgent need for cooperation. In reflecting the state-of-the-art discussion on the national scale (the international discussion is similar), the topic DP is now a self-evident part of the discussion. This is done necessarily on a very conceptual level, but the stage has already been set. The discussion has switched from electronic publications as a ‘final’ result of research to the whole publication chain. In particular, the long-term availability of research data is a booster for new initiatives.

Regarding the relationship between national and European/international perspectives, most of the national discussion currently is driven by contributions and funding initiatives on the European level. Therefore, the DNB is participating in different European projects like PARSEInsight, SHAMAN, and Keep in order to share experiences and to enhance existent know-how. Another initiative is related to a long-term archiving solution within the library itself. The common need for a more flexible and service-oriented system, which consists of distinctive modules and which can be integrated into different system environments, is shared by a number of national libraries in Europe. These

---

9 <http://www.allianzinitiative.de/en/start/>
10 <http://kolawiss.uni-goettingen.de/?q=de/node/10> (unfortunately only in German) http://www.allianzinitiative.de/en/start/>
11 <http://www.parse-insight.eu>
12 <http://shaman-ip.eu/shaman/>
13 <http://www.keep-project.eu/ezpub2/index.php>
libraries have started an initiative for a next-
geneneration long-term archiving system hoping
that in 2012 a solution will be available.

**Long-term preservation infrastructure in the German National Library**

In its current state, the long-term preservation infrastructure consists of the following two components.

- A repository system for collecting and describing bibliographically electronic objects for all daily work on these data and for access through external users.
- A back-end archival system for long-term preservation.

Both elements are part of a combined system, which acts as a Digital Multimedia Archival System. It amalgamates the extended functions of the kopal system at the DNB in its role as a repository as well as an archival system and goes beyond the core kopal-DIAS solution.

As part of the repository system of the DNB, the kopal archive has the role of the archival back-end. Before digital objects are transferred to this back-end system, they are stored for the whole preparation phase in a dynamic storage component, operated using Java Content Repository (JCR) in the Apache Jackrabbit implementation on a local server cluster. This component can be seen as the working storage for the system. In addition to its storage function in the preparation of long-term archiving, it also has a caching function for performant access of archived objects, which cannot be provided by the backend archive kopal. The following functions are integrated.

- Publication harvesting
- Conversion of delivered metadata
- Generation of additional metadata
- Transfer of metadata to catalogue

- A processing filter
- URN-service
- Delivery management
- Issue tracking
- AIP generation and storage
- Multi-client support
- Data retrieval and access

The system’s architecture is tailored to the business processes of the DNB. All publications pass through the same process chain, although there are two ways in which electronic publications can be submitted to DNB.

- Via a web form
- Via Open Archival Initiative (OAI) harvesting for automated mass delivery

In both the cases, metadata that are delivered with the publications are converted/normalized into the in-house format. A processing filter is applied, which validates/enriches the metadata, checks the file type, validates the checksum, conducts a duplication check, and prepares the objects for ingest into the JCR. The JCR acts as an intermediate working storage before the objects are ready to be ingested into the long-term storage area. Before the objects are ingested into the JCR, a Universal Resource Namespace (URN) is registered for any object that does not have a URN already. Moreover, an Issue Tracker is applied to the submitted objects. The submission is recorded in the Deliverer Management. Finally, the metadata are imported into the catalogue system, and the objects are transferred into the JCR.

During this process, no long-term, preservation-related technical metadata are created yet. Only very basic technical metadata, like filename, file-size, mime-type, and checksum, are assembled and transferred to the JCR with the original object. A few bibliographical metadata like title and identifiers

---

14 With acknowledgements to colleagues in the DNB: J Kett, M Neubauer, S Schrimpf, T Wollschläger

15 <http://hul.harvard.edu/jhove/>
are also transferred to the JCR. The IDN of the catalogue entry (a permalink) works as a persistent link between catalogue entry and digital object.

In the JCR, technical metadata are generated with the help of JHOVE. Object and metadata are packaged together as an Archival Information Package following the Universal Object Format (UPF) and handed over to the long-term storage.

For communication with the outside, several interfaces are provided or are in preparation, including web forms, SRU, and services based on OAI, especially the OAI–PMH protocol. With these interfaces, the way is open to integrate the DNB repository system into subordinated infrastructure networks as it is planned in the integrated project SHAMAN so as to provide and exchange data and metadata within these networks—naturally in the limits of the copyright law.

The JCR works as a generic interface between the production area and the permanent data storage (the kopal-DIAS). It consists of an application component and one or more backends for data storage. For the coordination of the archival processes, a service-daemon has been developed that detects objects eligible for long-term archiving with the help of the OAI interface. This service work is integrated with the koLibRI workflow tool. The service daemon acts like a koLibRI process starter and can at the same time react on internal status information of the workflow tool. The information about the archival status of objects (for example, ‘object was transferred into the long term storage area’) is deposited in the repository and can enable advanced functionalities in a future access system. For the time being, copies of the objects that were transferred to the long-term storage are not deleted from the JCR to provide the mentioned caching abilities.

**Digital preservation for the library**

As mentioned above, one major goal in the kopal project was the processing in a cooperative environment, which allows multiple institutions to participate on different levels of involvement. This means work sharing (the data centre is responsible for the operational service and the bitstream preservation and the libraries’ respective memory institutions are responsible for all the aspects of preservation planning and the ongoing functional extension of the solution). The transparent integration into existing library systems and the reusability through memory institutions play a critical role. Considering the aspect of flexible reusability, international standards for long-term archiving and metadata were adopted. In this way, both sustainability and the ability to further develop the system are guaranteed.

In its life time, the kopal project has developed the basic functions to implement and fill in the roles and responsibilities described before. But in the area of automated communication between different systems and the practical level of operational organization, kopal still needs some additional development. This will be done in a second step with Digital Preservation for Libraries (DP4LIB).

The planning for the next steps started after it became obvious that the technical solution in kopal is not detailed enough to address the different needs of potential partners—especially smaller institutions, with a need for simple and integrated solutions.

During the discussion with several organizations and institutions, it has become clear that there is a need to have different models of implementation with a high degree of customizable options. Moreover, it has been recognized as a motivation to become involved in the basic principles of DP. The requirements for this engagement are very different, which means that the single services must fit the user needs. Therefore, a consortium structure will be set up with a documentation centre, different types of libraries and library service providers as well as a virtual consortium of other institutions basically located in the information infrastructure of...
science and research. This guarantees a broad range of potential requirements to make sure that the project covers most of the needs of those organizations in a prototypic way.

Existing services should be enhanced to create a real cooperative solution, not only in a technical sense but also operationally/organizationally. Therefore, the starting point is identified to initialize the next step of development for the cooperative kopal solution.

On the basis of the kopal results up to now, the partners wish to improve the practical reusability of the software development. In order to create a generic solution that can be implemented in many heterogeneous environments and integrated as a part of the working policy of cultural heritage organizations, there is a need to develop an open concept with modularized service packages. The general goals are as follows.

- Creation of a flexible long-term preservation infrastructure adapted to the needs of (smaller) cultural heritage organizations and their service providers.
- Technical enhancement of the existing solution, conforming to the partners’ requirements.
- Implementation of a reusable process model and preparation of a handbook to introduce long-term preservation in (smaller) cultural heritage organizations.

Technically, this means that there is a need for an enhanced rights management in order to provide different levels of ingest and retrieval. There is also a need for a seamless integration. Another goal is to obtain more information about the costs and amounts of work for the introduction of long-term preservation processing into different types of organizations. Finally, it will be possible to generate valuable estimates for a funding and investment model for a complete infrastructure solution.

The partners DNB and the State and University Library Göttingen will offer a package of services, which will allow re-users to choose between different levels of service and to customize the existing solution to their specific needs. Furthermore, the partners will deliver dedicated consulting and operational services. Identified positions/factors in the cost model are listed as below.

1. Consulting and support
2. Detailed planning
3. Hardware extension, licenses
4. Adoption/customization of SW components
5. Ingest
6. Operational service

Funded by the German Research Foundation, the project has just started, and first results are expected at the end of 2010.

Summary

Over 10 years of experience in the DNB can be summed up as follows. We have reached a notably good level of cooperation and exchange with different partners from heterogeneous communities. But we have not yet achieved the goal to serve cooperatively a distributed technically advanced infrastructure for DP in Germany. However, there are encouraging approaches to move forward up to a cooperatively driven organizational and operational task sharing.
Cross-institutional cooperation on a shared bit repository

Eld Zierau, The Royal Library of Denmark, P O Box 2149
DK-1016 Copenhagen K
elzi@kb.dk

Ulla Bøgvad Kejser, School of Conservation, Esplanaden 34
DK-1263 Copenhagen K
ubk@kb.dk

Abstract
This paper explores how independent institutions such as archives and libraries can cooperate on managing a shared bit repository with bit preservation in order to use their resources for preservation in a more cost-effective way. It uses the Open Archival Information System (OAIS) reference model to provide a framework for systematically analysing institutions’ technical and organizational requirements for a remote bit repository. Instead of viewing a bit repository simply as archival storage for the institutions, repositories, we argue for viewing it as consisting of a subset of functions from all entities defined by the OAIS reference model. The work is motivated by and used in a current Danish feasibility study for establishing a national bit repository. The study revealed that, depending on their missions and the collections they hold, the institutions have varying requirements, for example, for bit safety, accessibility, and confidentiality. This study furthermore revealed that requirements for the level of bit safety must be supplemented by risk analysis, which needs to involve elements of the architecture, for example, the number of copies and how independence between the copies is ensured. The paper describes the bit repository architecture and its strengths in being flexible in order to offer differentiated services with respect to, among other things, bit safety and cost. Furthermore, the challenges in formulating, for example, risk requirements are described.

Keywords
OAIS reference model, bit preservation, preservation level, risk analysis, bit repository
Introduction

There is growing awareness that operation and management of digital preservation systems require ongoing resources and highly specialized knowledge. As a result, institutions seek ways of outsourcing or sharing the responsibility for these activities in order to use their resources for digital preservation in a more cost-effective way, and potentially benefit from economies of scale. One obvious area to explore in this context is bit preservation, because the requirements for ensuring that bits remain intact and accessible are relatively well understood, compared to functional (logical) preservation, that is, ensuring that the bits remain understandable and usable. In practice, there are already organizations that offer bit preservation services, for example, the Online Computer Library Center Digital Archive\(^1\) and Iron Mountain\(^2\). Other examples that offer bit preservation solutions, but also some functional preservation, are Kopal\(^3\) and LOCKSS\(^4\).

Along these lines, the Danish Ministry of Culture has funded a project to investigate the feasibility of a joint bit repository for the national archives, libraries, and museums in Denmark. The project is driven by the Danish National Archives (SA-DK), the Royal Library (KB-DK), and the State and University Library (SB-DK), which are also envisioned to be the stakeholders of the bit repository, together with other institutions with long-term preservation needs. The aim is to design a common system that will provide secure large-scale means of ingesting, storing, auditing, and accessing bits, while the participating institutions retain responsibilities related to the logic of the content. Thus, institutions will remain in charge of all structuring and packaging of data and metadata, and for selecting appropriate preservation strategies. This is, for example, modelled in a Fedora\(^5\)-based system at SB-DK, which, however, still needs bit preservation for its objects. The business model for the bit repository will be based on pay per service.

At a first glance, setting requirements for a shared bit repository may seem to be a simple question of setting up common interfaces at ingest–storage and storage–access. However, an analysis of the stakeholders’ requirements revealed that one of the challenges in setting up a common bit repository is that the institutions have quite different requirements when it comes to preserving bits. The requirements differ according to their mission and the collections (data sources) they hold as well as national and international legislation such as copyright laws and archive and personal data protection acts. Examples are requirements for specific hardware in order to protect data or lower costs, and requirements for fast online access while others may be stored on off-line media. Collections may also vary with respect to their perceived value. Digitally born e-books may, for example, have a higher value than digital copies of printed books, because the latter may be re-digitized, if the digital copies are damaged or lost (Rieger 2008, p. 42). Therefore, institutions may also be willing to accept a lower bit safety level for the collection of digitized books. These service levels could be documented within the metadata standard PREMIS, which has an entity termed ‘preservation level’ for describing possible preservation services and the context in which these apply (PREMIS 2008, pp. 33–38). There are many challenges to meet differentiated requirements in one system. For example, can confidential data, even if encrypted, be stored together with, for example, web data, which will be processed for statistics? Another example

---

\(^1\) Online Computer Library Center Digital Archive <http://www.oclc.org/digitalarchive/default.htm>
\(^3\) Kopal Long-Term Digital Information Archive <http://kopal.langzeitarchivierung.de/>
\(^4\) LOCKSS, Lots of Copies Keep Stuff Safe, <http://www.lockss.org/lockss/Home>
\(^5\) Fedora <http://fedora-commons.org/>
is how to allow and make secure deletion or overwriting of data from one source, in case it is stored together with data from another source that must remain unaltered at all times?

This paper provides a systematic framework for analysing these quite complex requirements using the OAIS reference model (CCSDS 2002). We propose to use an OAIS-style view to describe the interfaces between participating institutions and the bit repository. This analysis highlights that one cannot equate the functionality of a bit repository with that of the OAIS functional entity archival storage. It shows that subsets of functions from all OAIS functional entities are required, which is why we denote it as a shared bit repository instead of just a storage facility.

The paper then presents the architecture for a Danish national bit repository, which must support multiple institutions with differentiated services: for bit integrity, confidentiality, accessibility, and data processing. This includes a description of the challenges in specifying requirements, for example, for bit safety. Finally, we discuss and conclude on the results and point to future work.

**Analysis of requirements based on the OAIS model**

The Danish institutions strive to be OAIS compliant, and it is, therefore, an important requirement that the bit repository is OAIS compliant as well. The OAIS model has proven useful for discussing issues related to long-term preservation and for analysing repository systems. As illustrated in Figure 1, OAIS is structured around six main functional entities: ingest, archival storage, data management, administration, preservation planning, and access; each of these consists of a series of functions.

Section 6 ‘Archive Interoperability’ in the OAIS standard discusses cooperation between multiple OAIS archives. It includes an example of how the functional entities archival storage and data management can be shared by standardizing the interfaces ingest-storage and access-storage where MANAGEMENT is set in charge of agreements made between the archives. A similar design would immediately seem to be an appropriate concept for analysing the requirements for a shared bit repository. However, more functions than the above-mentioned interfaces and those described under archival storage need to come into play. Archival storage receives storage requests and AIPs from ingest and selects media, prepares volumes, and returns a storage confirmation, including ID. Archival storage also conforms to special levels of service or security measures. However, these operations within archival storage depend on input from other functional entities: ingest may indicate the required data utilization frequency (media type) via the storage request. Administration provides storage management policies and disaster recovery policies, and, via preservation planning, it also provides recommendations on system evolution and media migration. Likewise, data management is
indirectly required for managing IDs (tokens) and possibly audit trail. This shows that archival storage is actually interrelated with a subset of functions from all functional entities in OAIS.

We, therefore, propose to perceive the bit repository as a subset of a full OAIS. We propose to view the bit repository (BR) as being embedded within the archival storage functional entity of each of the institutions repositories (IR) as illustrated in Figure 2. We denote this OAIS-style view as the IR-BR model. We show that the IR-BR model helps in analysing the interfaces and describes the interoperability between multiple institutions repositories and the bit repository, and determines which of the two layers the requirements belong to.

It is important to notice that the IR-archival storage does not assume that an IR-AIP is stored as one digital object. As illustrated in Figure 3, an IR-AIP produced from an IR-SIP can consist of more digital objects, that is, BR-SIPs. Each BR-SIP could for instance be representations of file components, metadata, and so on. It is the responsibility of the

Figure 2 The OAIS-style view where the BR is shared by multiple IRs by embedding the BR in the IR archival storage.

Figure 3 Example of how an IR-AIP may be stored as more BR-SIPs.
Cross-institutional cooperation on a shared bit repository

IR-data management to keep track of the relations between the stored BR-SIPs and their identification information for eventual access.

In order to analyse the IR-BR model in OAIS terms, we need to have a closer look at the interfaces between IR–ingest and BR–ingest and IR-access and BR-ingest, and how these influence the way the OAIS functional entities are broken down into functions. The flow between these interfaces is illustrated in Figure 4.

IR-ingest functions are all executed within the IR. However, when an IR-AIP is transferred to IR-archival storage, that is, ingested in the BR, it takes a slightly different path in the OAIS-style view than if the BR was just archival storage: Here it runs via the BR-ingest functions before it ends in the BR-archival storage. Likewise, the receipt for accepted data and completed storage is returned to the IR, as the BR acts as an OAIS-compliant bit archival storage. Thus, IR-ingest receives the storage confirmation from BR-ingest. Note that BR-ingest/generate AIP is relatively simple, as there is no logic attached to the BR-SIP, except from identification information and possible audit trail documentation, that is, no representation information is needed, as we only look at bit streams regardless of their interpretations.

Likewise, in IR-access, all functions involved are executed within the IR except at the interface between IR-access and BR-a0ccess. Thus, the IR-access/generate DIP function does not retrieve the IR-AIP directly, but goes via the BR-access function. Corresponding functions are run at the BR level.

Data management functions are administrated within the IR and the BR, respectively. If the institution wants information from the BR-data management to be part of the IR-data management, for example, audit trail information, then this information must pass through BR-MANAGEMENT.

Likewise, administration functions are managed within the IR and the BR. However, there is one exception for IR-administration, namely the receipt of reports on operational statistics from IR-archival storage/manage storage hierarchy, which under IR-ingest and IR-access goes via the BR. Furthermore, as already described, the requirements for storage management and disaster recovery policies set up by BR-administration are coordinated with IR-administration via MANAGEMENT.

Preservation planning is split between the IR and the BR. IR-preservation planning, among other things, handles technology watch regarding data format migration, while media migration is placed in the BR-preservation planning. Note that this also means that all activities related to functional preservation are solely up to the IR.

Figure 4 The interfaces between the IR-BR model
Observing the detailed aspects of sharing a BR, the BR-ingest must be able to receive BR-SIPs from different IRs. Furthermore, the individual IRs must be able to query BR-SIPs as a BR-DIP at a later stage, which means that the BR-SIPs must, as a minimum, contain the bits to be stored, identification of data source, and a unique identifier within the IR. Note that these requirements may also be needed for a single IR if it preserves more data sources. As a result, the BR-ingest/receive submission is required to distinguish from which disjunctive sources does the BR-SIPs come from, and the associated requirements for storing the digital object.

Generally speaking, all requirements that the institutions may have are based on directions from IR-MANAGEMENT and are dealt with by IR-administration. From the BR perspective, IR-administration represents BR-MANAGEMENT. Thus, it is at the interface between IR-administration and BR-administration that the mapping of the requirements for the BR takes place.

Requirements for the architecture for a national bit repository

In this section, we describe how the OAIS Model has impacted the architecture of the shared national bit repository (DK-BR). However, in order to understand the context, we start by a short description of the most important requirements for a DK-BR. Secondly, we describe the overall architecture and argue for the choices made on the basis of experiences. Seen in this context, we describe relevant issues noted by the OAIS analysis. Additionally, we describe the remaining challenges in expressing requirements for bit safety in general, and determining the right abstraction level.

Requirements

The DK-BR architecture is based on analysis of stakeholders requirements, the OAIS analysis, the partners own experiences with running bit preservation systems, and observing similar international institutions.

Both existing Danish institution’s BRs and the DK-BR build on general considerations for active bit preservation, that is, the number of copies, the degree of independence between copies, and the audit frequency of bit integrity (Rosenthal 2008). The audit frequency denotes the frequency of ongoing checksum control, which requires at least three checksums for the stored data, in order to compare and, if differences are detected, identify by voting which copy is erroneous and needs to be replaced (Christensen 2005).

From the requirements analysis, it was clear that a very flexible architecture was needed in order to meet varied requirements at different costs. For example, there were requirements for different levels of bit safety, where the number of copies has a direct influence on cost. The same applies for other requirements like access, where there were expensive requirements for having a copy stored on a distributed platform in order to enable processing for statistics on web data, for example, for linguistic analysis for researchers. This processing is required within the BR, as the large quantity of data makes it economically infeasible and technically hard to execute it outside the BR. Another example is that the SA-DK has one copy of data off-line on an optical media (DVD) because of infrequent access requirements and because this media is relatively cheap and useful for securing data confidentiality. Note also that requirements may conflict, for example, requirements for confidentiality and bit safety: while more copies increase bit safety, the risk of compromising confidentiality also rises.

A flexible architecture to meet various requirements

The architecture for the DK-BR has similarities to systems like LOCKSS and DuraCloud\(^6\). However, the DK-BR especially differs and has

---

\(^6\) DuraCloud http://duraspace.org/duracloud.php
its strengths in the way it meets differentiated requirements for bit safety and costs. A further strength is that it avoids having a central master index for the BR contents, which minimizes the risk of single point of failure. The overall DK-BR architecture is illustrated in Figure 5. Integrity audits (checksum checks) and corrective actions are also placed here. Client applications are configured to serve the individual SLAs.

Finally, the coordination layer coordinates communication between user applications and pillars. Its main functionality can be seen as being a data-bus for the communication.

The architecture is flexible to meet different SLAs for users by offering different combination of pillars that best serves the user’s requirements. The architecture support representation of different pillars, which can be combined to meet special requirements as, for example, for access speed or processing, to meet cost requirements, or to prevent specific risks. In other words, selection of pillars to an SLA is based on an analysis of the individual pillar characteristics and the combination of the ability of pillar characteristics to match the user’s requirements. For example, the distance between pillars influences the risk of two copies being destroyed at the same time. The diversity can be on many aspects, for example, technological (for example, hardware and software, on-line and off-line media, optical and magnetic media, employment of different storage techniques, and using different vendors), geographical, organizational, or data related (for example, checksums based on different checksum algorithms, different characteristics expressing policies of read only storage or no processing). Such considerations were used in the existing bit repository for Danish web material in Netarkivet\(^7\) (web-BR) (Christensen 2005), but not as broadly as those in the DK-BR.

Another difference is that a pillar may only contain a representation of a copy of the data in the form of a derived checksum. Storing a checksum does not make sense on its own, but it can provide one voter in a storage solution with bit integrity checks, but with less copies, and thus is more economical. In a risk assessment of the web-BRs storage, economy was an argument for

\(^7\) Netarkivet – Danish web archive www.netarkivet.dk

Figure 5 Overall architecture of the DK-BR
having only two pillars with full copies and one with a derived checksum to obtain three voters. It is however important to consider where the derived checksum is stored. This was evident from an assessment of the Netarkivet using DRAMBORA, a tool for internal audit of digital repositories (DCC and DPE 2007). It revealed that the checksum copy was positioned at the same physical location as one of the pillars, which meant that a fire or explosion could result in loss of two out of three voters.

The architecture avoids single point of failure by avoiding a central master index of the contents of the BR. Instead, it is based on cached information that can be updated or verified by information from the pillars. Thus, there is no single point that is indispensable in calculation of a master index. The architecture also avoids single point of failure by having duplicated parallel instances of coordination layers and user applications, although this is not depicted in Figure 5. This also meets requirements to scalability and uptime as they are designed to be replaceable.

Other risk-preventing actions are taken into account, including risk-preventing actions from standards required for the Danish institutions called DS 484, which corresponds to ISO/IEC 27001:2005 (ISO/IEC 2005). This involves, for example, restricted access to the physical servers, and separation of production and test environments.

**OAIS view**

Taking an OAIS view, all requirements from the IRs are in practice administrated by BR-administration, and translated into specific requirements for the different OAIS functions. Thus, specific requirements for archival storage, such as selection of storage media, throughput rate, and bit error rate, are handled in BR-administration by specifying an SLA defining appropriate combination of pillars and their characteristics.

BR-ingest and BR-access only operate on handling data in the form of bit sequences, leaving these functions close to the OAIS archival storage functions: receive data and provide Data. However, BR-ingest must provide BR-data management with unique identification of bit sequences and source to which the bit streams belong.

The more complex aspect of data management is the audit trail. In the DK-BR architecture, audit trails are mainly provided by individual pillars, and it must be considered in the final design whether, for example, information about media migration must adhere to AIPs or be part of the system management information. Additionally, requirements for how pillars must preserve audit trails must be stated.

As argued earlier, the BR-preservation planning takes place within the BR. In the DK-BR architecture with several independent pillars, it actually takes place within each pillar, which leaves technology watch for media to the individual pillar. In order to assure bit safety, the institutions have to rely on documentation from the BR-administration regarding IDs, checksums, and possibly audit trails. That is, it is reported via BR-administration information to the IR-administration passed to the IR-preservation planning, where actions may be initiated in case of alerts. As IR-preservation planning cannot directly require that a pillar, for example, performs media migration, the IR may instead change to another pillar within the BR, which better fulfills its requirements.

It is also noteworthy that requirements from IR-preservation planning may influence other requirements for the BR, for example, requirements for file format migration. The reason is that such a migration on large data volumes can require large CPU power for processing directly on the stored data, as capacity limits, for example, in bandwidth, can make it practically impossible to do it outside the repository via access and re-ingest.
The organization of DK-BR influences the implementation of BR-administration and the reporting functions needed via or in parallel with the client applications, but it is not yet decided how this organization will look for the DK-BR.

**Special requirements specifications and specification levels**

Requirements to bit safety, confidentiality, political, and economical requirements are hard to express explicitly without involving elements within the architecture. There are many challenges in specifying and controlling such requirements, which means that there will always be an element of trust between the IR and the BR.

It is generally challenging to express risks by quantitative elements, because this can result in a very simplistic view. In particular, the requirements for bit safety are hard to express, due to the lack of scientific ways to measure, verify, and control this (Rosenthal 2008). As part of the work with the DK-BR architecture, we tried to design a risk model. At the pillar level, risks are indicated in terms of pillar characteristics, which are made as explicit as possible. At the BR level, risks are indicated in terms of the combination and variation between pillars in order to express the independence between copies/pillars. Also, the number of copies and bit audit frequency must be taken into account. The final risk model could either be based on statistics or simulations as it was done for the Netarkivet (Christensen 2005). Additionally, aspects like single point of failure and aspects from the TRAC checklist (OCLC and CRL 2007) should be incorporated. This work was not finalized within the pre-study, but a number of related challenges were identified.

Also, the level of abstraction specifying requirements can be a challenge. When Netarkivet had a hardware migration after five years in production, it appeared that the original requirements were too specific, including, for example, the exact location of a pillar and which specific OS to use. The original idea was to require independence between pillars, for example, in the form of distance and different operating systems (OS), but this was not put in as requirements and not in terms of exact distances between pillars and how much the two OS must differ. Consequently, an abstraction level must be made that avoids dependencies of the ongoing evolution, or requirements should be regularly adjusted to match present time.

**Discussion**

Our analysis of OAIS highlights that one cannot equate the functionality of a BR with that of the OAIS functional entity archival storage, and this led to the IR-BR model. While there is no conflict in viewing the BR as a subset of an OAIS archive, the OAIS-style view where the BR is embedded in IR-archival storage may be subject to discussion. The reason is that some of the functions that communicate across the interfaces between the IR and the BR have to be redefined in order to take into account that the data and information flow takes a different path. However, the OAIS standard states that actual implementations may break out functionality differently; thus, in our opinion, we did not compromise the OAIS concept by these redefinitions.

The IR-BR model was useful for identifying interfaces, and for understanding the flow of data and documentation between IR and BR. In particular, the model helped understand how the audit trail should be managed. It also showed that we need to consider the nature of the digital object identifiers. Envisioning having a BR for the next 100 years or more, it may not be feasible to have historic information related to the source, for example, relating to an institution that no longer exists, hidden in the identifiers (Zierau and Johansen 2008). Therefore, having UUIDs for digital objects in the BR should be considered, along with the data source and identifier given by the IR.

If we observe other variants of the preservation system like LOCKSS or Kopal, the
split between their system and customers is placed differently compared to the one presented in this paper. LOCKSS, for example, also includes some functional preservation, whereas we looked at a BR without any knowledge of the digital objects apart from identification and audit trails. Kopal is also described in terms of OAIS, but leaves validation and packaging outside the OAIS repository. The question is whether our IR-BR model can be used on these cases as well. It has not been a part of our work to attempt this, but we would expect that the same model could help in identifying locations of functionality, interfaces, and services as we have done for a BR in this paper.

Although the DK-BR has a flexible architecture, the final possibilities of making SLAs with users will depend on the number and variation of pillars that are implemented within the DK-BR. Furthermore, the way in which different requirements contradict each other, especially looking at confidentiality, integrity, and costs, can mean limits on the possibilities.

However well the requirements may be expressed, the success of separating out bit preservation and sharing or outsourcing these activities will depend on the ability of the BR to demonstrate trustworthiness. Thus, specification of requirements should be extended by audit and certification initiatives, such as the audit and certification tool for trusted digital repositories TRAC, which extends on the OAIS standard. Furthermore, without sanctions or escalation procedures, it is not possible to enforce that the requirements are met.

Conclusions and future work
We conclude that it is helpful for institutions setting up differentiated requirements for a shared bit repository to view it as a subset of a full OAIS. In this sense, the bit repository becomes a repository within the institutions repositories as described in the IR-BR model. This analysis helped reveal locations of functionality and set up the requirements.

Besides, we found that requirements must specify how audit trails within the BR are managed and possibly sent to IR. Also, we found that BR must provide BR-SIP identification and source information.

For the DK-BR, we found that there are significant differences in requirements for different data sources. In order to make it economically beneficial to share such a BR, it is important to look for similarities in order to find the areas where economic gains can be achieved and define architecture that meets differentiated requirements.

In the work with the DK-BR architecture, we found that there is still a need for applying a risk model in order to express issues like bit safety and confidentiality as explicit requirements. However, even though they are expressed via a risk model or directly, the abstraction level of the parameters or requirements must be done with care in order to ensure the bit preservation, and avoid complicating evolution through, for example, migration or organizational changes over time. Both risk model and abstraction level for requirements specification could be subjects for further studies.

When expressing requirements for a BR, there will always be an element of trust involved, even if they are expressed via risk model or other means. Therefore, future work will include articulation of requirements and BR specifications in relation to certification and audit initiatives. This will also contribute to allow balancing of the requirements against the costs.

Acknowledgements
Fruitful information and discussions have come from KB-DKs involvement in the Danish feasibility study for a national bit repository. Also, valuable input from discussions with Andreas Rauber from Vienna University of Technology has contributed to the paper.
Cross-institutional cooperation on a shared bit repository

References


Piloting an institutional repository at a research-intensive university: strategies for content recruitment and the role of the library

A Abrizah, Digital Library Research Group, Library and Information Science Unit
Faculty of Computer Science and Information Technology
University of Malaya, Malaysia
abrizah@um.edu.my

Abstract
A research group at a research-intensive university, whose vested interest is in increasing the accessibility of the university’s research output to the world literature, has developed a self-sustainable and expandable open access institutional repository (IR).

The paper developed by the University of Malaya’s IR describes the strategies that the digital library research group employed. It also explains how the roles of academic librarians would change in the process of building institutional repository. There are extensions of existing roles in terms of system evaluation, advocacy, and reference services. New roles include content recruitment and interpreting policies. It also points out possible directions that can make the repository sustainable.

Keywords
Institutional repositories, open-access digital libraries, academic libraries, roles of librarians, Malaysia
Introduction

Institutional repositories (IRs) are now becoming a component of the technical infrastructure in research-intensive institutions, and a favoured option for providing open access to research output. Universities and research institutions throughout the world are investigating, piloting, and developing systems for building collections of digital resources and learning materials in the shape of open-access repositories (Rothery and Bell 2006). Foster and Gibbons (2005) define an institutional repository as ‘an electronic system that captures, preserves, and provides access to the digital work products of a community.’ Crow (2002) and Ware (2004) characterize the following features of an IR.

- It is institutionally defined and it captures only the intellectual property of the host institution such as purely scholarly work, or administrative, teaching and research materials, both published and unpublished.
- It is open and interoperable and the primary goal is to disseminate the institution’s intellectual output.
- It is cumulative and perpetual and carries with it a long-term obligation on the host institution to preserve IR content.
- It contributes to the process of scholarly communication in collecting, storing, and disseminating scholarly content. As such, authors and researchers can deposit materials in IRs, subject to copyright, with the host institution providing the infrastructure for these materials to be properly organized, archived, and disseminated.

This infrastructure has emerged since 2002 when major research universities in the US (such as MIT and Cornell University, using DSpace), and the UK (Southampton and Oxford University using E-print) launched their own IR systems. Over the past four years, an increasing number of research universities has implemented or plans to implement an IR (Markey et al. 2007). Lynch and Lippincott (2005) found that, in the US, out of the 97 universities categorized as ‘doctoral universities,’ 40% already operated IRs. Among non-implementers, 88% were found to be in the planning stage of IR implementation. In 2005, a survey was undertaken at universities in 10 European countries—Belgium, Denmark, Finland, France, Germany, Italy, the Netherlands, Norway, Sweden, and the United Kingdom (van Westrienen and Lynch 2007). It was found that the number of IRs varies from as low as 1.5% of universities in Finland to as high as 100% in Germany, Norway, and the Netherlands, with the focus almost exclusively on collecting faculty publications. By mid-2006, all Australian universities had established IRs for the purpose of giving researchers a vehicle to enhance the availability of their publications by making them available via open access (Henty 2007). On the other hand, studies on IRs in Asia revealed that open-access repositories are not widespread in China (Fang and Zhu 2006) and that universities in India are at present lacking in infrastructure for establishing IRs (Fernandez 2006).

A few research universities in Malaysia have established—or are on their way to establishing—IR services to enhance visibility and impact of research generated within that university. The development of IR services is related to the open-access movement in Malaysia, which seeks to make valued research outputs openly available by encouraging academics to place their publications into repositories, thereby enhancing their availability and visibility to the global academic community and increasing the chances for use and exchange of ideas among scholars of similar disciplines (Abrizah et al. 2007). At the same time, university research increasingly involves the use, generation, manipulation, sharing, and analysis of digital resources. However, not every institutional repository in Malaysia adopts the principle of open access and it is possible for the institution to restrict the access to its member (www.opendoar.org). The University
of Technology Malaysia (UTM) and the National University (UKM), for example, allow access to few theses and dissertations not only to members of the institution. This characteristic fits Clifford Lynch’s framework of institutional repositories: ‘a set of service that a university offer to the members of its community for the management and dissemination of digital materials created by the institution and its community members’ (Lynch 2003). Three out of the 15 Malaysian archives that have joined forces in the open access movement in Malaysia (Figure 1) and have registered with the Registry of Open Access Repositories (ROAR) (http://roar.eprints.org). These efforts have been initiated by the research-intensive university focused in this paper—the University of Malaya.

This research is concerned with understanding the strategies that advance and influence IR development. The aim was to learn about the conditions that facilitate—and challenges of—the IR initiative at the University of Malaya, and to understand the role of the academic library in providing this research infrastructure. There are some research studies that are close to this goal. In order to understand the requirements to provide an IR that will preserve and disseminate research materials created by or associated with a research-intensive university, the present study began with an extensive search for information concerning content recruitment through faculty’s contribution towards open-access publishing and institutional repositories. It was apparent from this review that there research has been conducted that focused on the needs and potential contribution of faculty, as well as librarians’ roles in this area.

**Literature review**

In research universities, IRs are predicated on contributions by their stakeholders, which include both academic and non-academic

---

**Figure 1** Malaysian Archives, initiated by the University of Malaya

staff: those involved in teaching and research, and both postgraduate and undergraduate students. Each of these groups contains potential authors and readers of materials in IR, and the contributions of authors are critical to the success of an IR. As such, whether or not IRs become a part of the intellectual infrastructure depends on the extent of the university’s community contribution. Shearer (2003) argues that the success of an IR should be determined by its use, and one of the measures of usefulness is the contribution of content. Faculties are ideal for making a major contribution to an IR by creating, not preserving, new knowledge. This is because they are becoming increasingly involved in producing scholarly works and participating in the evolving scholarly communication process. As IRs are flourishing to preserve scholarly output and make it openly accessible, more and more faculty members are in favour of providing open access to the universities’ research output, maintained either institutionally or on a subject basis. Faculty contribution is considered to be one of the success factors for an IR, even though several studies have found low rates of faculty submission (Chan 2004; Foster and Gibbons 2005; Pelizzari 2005; Davis and Connolly 2007). These studies found that the challenges for an IR are not in its technical implementation, but in affecting the culture changes necessary for it to become an integral part of activities of the research institution. Cultural, rather than technological, factors limit the use and development of IRs. Literature suggests that ingrained behaviour, inertia, indifference, and resistance to change hamper the adoption of working practices needed to support IR (Ware 2004).

**Institutional repository**

While institutional repositories are becoming more prevalent in academic life, the disappointingly small number of materials in them reflects worldwide trends. Davis and Connolly (2007) reported that Cornell’s IR is largely underpopulated and underused by its faculty because the Cornell faculty have little knowledge of, and little motivation to use, the repository. Van Westrienen and Lynch’s (2005) European survey also reported low faculty participation in IRs. Their article identified several reasons for non-participation from faculty, including the following: (1) difficulties in informing faculty and convincing them to participate; (2) confusion and uncertainty about intellectual property issues; (3) scholarly credit, and how the material in IRs would be used; (4) perception of open access content being of low quality, and (5) lack of mandatory policies for depositing manuscripts. Correspondingly, Swan and Brown (2005), who investigated author self-archiving behaviour, found that there was a substantial proportion of authors unaware of the possibility of providing open access to their work. Only 30% of the 1296 respondents using specialised open archive initiative (OAI) search engines to navigate the open access repository, and only 10% of authors knew of the SHERPA/RoMEO list of publishers’ permissions policy with respect to self-archiving. More people opted for putting their work on a website than those that used institutional or subject-based repository. However, a vast majority of authors would willingly comply with a mandate from their employer or research funder to deposit copies of their articles in an institutional or subject-based repository. Swan and Brown (2005) found that authors’ reluctance to self-archiving their work was due to the perceived time required and technical difficulties in carrying out the activity.

**Open access initiative**

Although a number of studies have investigated the attitudes of authors with respect to open access publishing and IRs, none have, however, viewed other institutional stakeholders. Academic libraries, in particular, are a group that can make a major contribution to an IR. Academic libraries are increasingly becoming
involved in managing electronic scholarly products and participating in the evolving scholarly communication process. They not only acquire electronic resources but also create them. Libraries are being funded to digitize valuable parts of their special collections, especially theses and dissertations, both to preserve the original and to make the content readily accessible. As IRs are flourishing, more and more academic libraries are in favour of providing open access to the universities’ research output, maintained either institutionally or on a subject basis.

Open access and IRs may result in considerable savings for libraries besides the potential benefit of greater exposure to authors’ works. Although the future shape of scholarly communication in IRs remains unclear due to lack of contribution by the stakeholders, it is clear that library and information professionals have key roles to play (Chang 2003; Allard, Mack and Feltner-Reichert 2005; Chan, Kwok, and Yip 2005). Chang (2003) proposes that it is necessary for librarians to be conversant with digital collection management and open archive information system management skills. Library staff need to be trained to prepare documents in an acceptable format and submit content to the repository. Allard, Mack and Feltner-Reichert (2005) found that IRs provide librarians with new challenges because self-archiving makes authors more active partners in collection development. Also the librarian might become the steward of unique collections that grow rapidly because of author contributions. In her analysis of 30 scholarly literature on IR, the author found that nearly one-third of the articles did not mention how a library is involved in the IR effort. Even those who did note that libraries had a role in the process did not always explain the role of library. The areas that were mentioned as having involved the library included IR creation and maintenance. Chan, Kwok, and Yip (2005), on the other hand, illustrate how the roles of reference librarians are changed in the process of building institutional repository.

There are extensions of existing roles in terms of system evaluation, advocacy, and reference services. Brand new roles include content recruitment and interpreting publishers’ policies.

Librarians as change agents
Librarians are in an ideal position to act as change agents in the promotion of their own university’s IR, as well as other IRs as potentially valuable sources of information for their clients. Bauer (2005) points out that marketing IR is critical. Also marketing it to librarians should be the first step in promoting it on campus because, without their support, it will be difficult to achieve broader acceptance on campus. One responsibility of the academic library reference team is to show the academic staff and students in their university how to find and use information. Hence, librarians are ideally placed to act as change agents promoting IR as an information source (Revell and Dorner 2009).

Objectives and method
The objective of this research project was to identify strategies and conditions that advance and influence IR development. The aim was to learn about the conditions that facilitate, and challenges of, IR initiative at the university that had made substantial commitments to developing and sustaining an IR. Questions to be answered are as follows: (1) What does faculty think about making their intellectual output available through IR? (2) What are the strategies that the university can employ to populate IR with its research output? (3) How do librarian intermediaries contribute to the IR development process?

The study has provided fruitful territory for exploring this question, with data collected directly from the experience of the repository developers, the faculty and librarians associated with IR efforts. Data-gathering techniques employed to answer the research questions are mentioned in the following list.

- A web-based survey to investigate (a) the issues in establishing a facility to provide
open access to research materials such as level of knowledge, motivation, participation, partnership, ownership, and management and (b) the potential of an IR and the requirements of a good digital repository in allowing faculties to cooperatively develop and upload the resources to the IR. An e-mail invitation to participate in the survey was sent out internally to all academics (around 800 of them) within the university, which piloted an IR in 2008. The e-mail, which contained a hypertext link, enables the participants to link to the survey database hosted by SurveyPro (www.surveypro.com). The survey instrument consisted of six sections: (a) awareness and knowledge of IR as well as current IR contribution; (b) usefulness and importance of IR; (c) self-archiving experience; (d) future IR contribution; (e) reasons and concerns for contribution, and (f) demographic.

- Interviews and discussions during formal and informal meetings with groups of people have played important roles in building the IR. Information on the additional features that the IR should have, and the roles that librarians would want to play in an IR service was obtained through a library workshop on IR collaborative content development.

- Personal observations and experiences through formal and informal encounters with individuals who are potential users and submitters of the IR. Information on issues such as access and copyrights was obtained through telephone calls and e-mails from these people.

As a case study, the analysis was designed to be illustrative and capable of capturing a range of development approaches and experiences to suggest areas for further research. The results presented here are not intended to be representative of the full range of IR development activities. However, the analysis does provide a useful base of findings that can inform and guide academic libraries as they make decisions about priorities and approaches to developing their own IR initiatives.

**Setting up the infrastructure**

The university’s IR started with the basic premise that the scholarly output of researchers is an institutional intellectual asset, one that should be carefully guarded and preserved for posterity. Although, in most cases, researchers have transferred the copyright of their publications to the publishers, they may still exercise their self-archiving rights to make their scholarly work openly accessible if all publisher requirements are complied with. Once this principle has been agreed upon, gathering university and library-wide support for the project was easy. The following groups of people have played important roles in building the IR. They have been engaged in all stages of its development: the definition of goals and scope, evaluation of system and content, forming strategies and procedures, interpreting submission policies, and contacting and servicing faculty.

- The Digital Library Research Group: The Digital Library Research Group (DLRG), positioned under the Library and Information Science Unit, Faculty of Computer Science & Information Technology, University of Malaya (http://www.fsktm.um.edu.my), is committed to research and provision of information infrastructure for digital resources to be properly organized, archived and disseminated. Some of its ongoing research projects are the Malaysian Abstracting and Indexing System (MyAIS) (http://myais.fsktm.um.edu.my), the Digital Library of Malay Manuscripts (MyManuskrip) (http://mymanuskrip.fsktm.um.edu.my) and the university’s Institutional Repository (UMDSpace) (http://dspace.fsktm.um.edu.my) (Figure 1). The research group leader, Professor Zainab Awang Ngah, spoke about the need for universities and research institutions to handle their own digital
repositories in her inaugural lecture in 2006 titled ‘Scholarly Skywriting: E-print Archives and E-Journals, Panacea or Problem?’ The idea of establishing an IR at the university then grew from a funded research in 2008 to investigate the issues in establishing a facility to provide open access to research materials such as level of knowledge, participation, partnership, ownership and management, and the potential of an IR and the requirements of a good digital repository in allowing faculties to cooperatively develop and upload the resources to the IR. An outcome of the research is an IR to support a new pattern for scholarly communication, apart from surfacing UM’s ‘hidden resources’ (which have not been tapped on a large scale so far) and low-cost interoperability among various faculties’ web portals. It is expected that this IR will increase the accessibility of scholarly work, which exist in digital format, and make UM’s contribution to world literature more visible. Materials can be accessed more widely and also utilized for purposes different from those that originally motivated the creation of the repositories. An additional incentive is the potential for cost saving inherent in new models of the scholarly communication process that could be supported with this approach. Such a system is becoming necessary for the university to gauge her publication performance at both individual and institutional levels, and measure the research output. It can provide a better knowledge of the university’s output from research activities funded by various national and international organizations and improved information utilization for end users.

- Department of Systems and Network, Faculty of Computer Science and Information Technology: There are many routes to engagement and many strategies for service deployment to support repository services. Among others are in-house development and deployment, collaborative or partnering approaches, and contracting for services (Association of Research Libraries 2009). This research project employs in-house development and deployment of the IR as open source applications are available (for example, DSpace, E-print or Fedora), making it possible to get a repository up and running with quite modest resource investments. In the early stage of the project, the DLRG partnered with their systems and network colleagues to evaluate two digital library softwares, Eprints and DSpace. Both systems were evaluated by comparing and contrasting them on a number of criteria: database structure, interface, search capabilities, special features, software requirements, speed and reliability, and export options. Advanced search and sort options were only available in Eprints, but DSpace supported browsing by more search fields. Eventually, DSpace was DSpace, mainly because it represents knowledge through communities and collections. In other words, DSpace is organized into communities, a high-level organizational structure whose only purpose is to divide collections into related groups, and each community contains one or more collections, which are containers for related items (deposited object of any type: a published article, an image, audio, or video file, notes, a presentation, and so on.). Both groups found that this structure for organizing information is very suitable for a research institution. Moreover, it was reported that nearly 80% of those with IRs were using a local deployment strategy and more than half of them were using DSpace (Lynch and Lippincott 2005).

- The University Management: Institutional support is vital for the sustainability of any IR. The university management has agreed to mandate open access self-archiving. A letter from the vice-chancellor to all deans and directors, dated 2 March 2009, required the
faculties, academies, and institutes to submit soft copies of theses and dissertations to the head of information systems department at the library to be deposited in the IR developed by the DLRG.

- The University Library: The librarians have been engaged in various stages of the IR development: the definition of goals and scope, evaluation of system and content, forming strategies and procedures, interpreting publishers’ policies, contacting and servicing faculty members, acquisition of content, and promotional efforts. Content recruitment for the IR has been set up in liaison with the university library. The Chief Librarian and her Deputy stepped up and volunteered to be part of the IR initiative, motivated by their belief in open access principles. They participated in advisory boards that crafted guiding policies on preservation, submission procedures and collection scope, and they developed and implemented plans for generating repository awareness. A few librarians were also involved in evaluation and testing of repository functionality and interface design through a special workshop organized by the library.

- The Institute of Postgraduate Studies: The Institute of Postgraduate Studies (IPS) is engaged in digital repository to surface the university’s postgraduate students’ research output. The IPS accords high priority for this activity as it is essential for maintaining the institute’s value in a research university. The IPS has facilitated submission of digital copies of theses and dissertations on CDs from faculties to be passed to the library for archival purpose in the IR.

**Issues and strategies for content recruitment**

After setting up the infrastructure, the main focus is on getting content into the IR. After continuous content recruitment of the IR’s testbed of theses and dissertations by the Faculty of Computer Science and Information Technology community, the DLRG works on a wider content recruitment approach and user-support structure. The web-based survey concerning faculty’s self-archiving experience and IR adoption (Abrizah 2009) revealed that 73 (55.7%, n = 131) academics of the university had deposited their research/teaching materials on publicly accessible web sites as well as other open access digital repositories. All these self-archiving respondents knew what open access meant, 47 (64.4%) were aware of the university’s IR and the majority (65, 89.0%) planned to contribute to it. Therefore, most respondents had some IR awareness, and a majority of those who planned to contribute already had experience with self-archiving. Out of the 73 respondents who had self-archiving experience, 3 (4.1%) had self-archived their work for more than five years, nine (12.3%) had done so for three to five years and an additional 14 (18.2%) had one to three years’ experience. The majority (47, 64.4%) had deposited their work on publicly accessible web sites for the past one year. When asked about the frequency of contribution to IRs in an open-ended question, one respondent stated, ‘I have been contributing through my faculty’s web site for years.’ Nine respondents who were aware of the university’s IR planned to contribute in the future, and already had self-archiving experience in other open access venues such as arxiv.org, E-LIS, and MyAIS (myais.fsktm.um.edu.my), the open access system for Malaysian scholarly publications. One professor indicated depositing various versions of his scholarly articles to a particular open access repository whenever the papers have been submitted for review and have been revised.

On the basis of the findings, it has been decided that the DLRG replicate Foster and Gibbon’s (2005) strategy for content by working with a small IR early-adopter group from other faculties, and networking from them to their colleagues. Ten academics from science-based faculties were invited to participate in
submission and to evaluate the DSpace IR. However, only one responded and agreed to participate. This science researcher set up his research sub-community, and populated it with 12 items under a collection in two weeks’ time. However, for unknown reasons, the collection was deleted a few months later. It was found that the researcher did not want to ‘make his collection publicly available any longer’ (anonymous, personal communication, 26 November 2008) and decided to remove the items. This demonstrates a misunderstanding in the way an IR operates and the ignorance that the system has the feature to allow access to items and collections to members of the institution or to the community only.

In the web-based survey (Abrizah 2009), the researcher elicits 131 faculty responses to a requirement from the university or research funder to make their work available via open access by self-archiving in the university’s IR. A total of 52.7% (69) respondents complied willingly, while 47.3% (62) complied reluctantly. There were none who did not comply. The finding clearly indicated that a mandate from an institutional employer or a research funder to self-archive would meet with very little resentment and even less resistance from the respondents. This study supports those by Kim (2007) and Swan and Brown (2005) who opined that if employers and grant funders encourage self-archiving, authors or researchers would consider depositing their work into IRs. If not, they would not be motivated to do the same. The researcher’s views are that it would also make a tremendous difference if the university management could strongly encourage faculty members to self-archive, instead of submission done by a librarian intermediary.

What makes faculty reluctant to contribute to IRs? In order to investigate this question, the 65 respondents motivated to contribute to the IR in the web-based survey had to respond to 28 statements regarding their concerns about self-archiving. Overall, many faculty members disagreed with the statements presented as ‘deterrents of self-archiving’ (Pickton 2005). The top three deterrents for more than 70% respondents include: ‘I am concerned about other publishers owning the copyright of previously published material’ (75.4%); ‘I am concerned about plagiarism’; and ‘I am concerned that others might copy my work without my permission’ (73.8% respectively). As such, concerns about copyrights and plagiarism might impede self-archiving. This issue has also been expressed by faculty and students who contacted the researcher through e-mails indicating that ‘the theses should be accessed only in campus networking, using username and password to eliminate copy and paste and for security reasons’ (Chin H L, personal communication, 5 October 2009) and ‘it is not a good idea to have the whole dissertation available full-text as students can easily plagiarize the contents’ (Yu M L, personal communication, 5 May 2009).

In addition, more than half of the respondents in the survey disagreed with the following statements reflecting pre-print culture, publishers’ policy, trust of readers, and preservation as the reasons for not contributing to IR.
- I do not want to put my work with work that has not been peer reviewed (36, 55.4%)
- I might want to change or delete my work (43, 66.2%)
- I am concerned that if I deposit my work in the University’s Repository I may not be able to publish it elsewhere later (36, 55.4%)
- I am concerned about the effect of open-access repositories on journal publishers (44, 67.7%)
- I am concerned that others might alter my work without my permission (44, 67.7%)
- I am concerned about the long-term feasibility of the repository (43, 66.2%)
- I am concerned that my work might not be preserved in the long term (41, 63.1%).
This result suggests that the respondents might be more concerned or skeptical about the quality and secure maintenance of open access materials. As such, IRs might have to emphasize their function of facilitating the pre-print culture and of long-term preservation and explain how these would be accomplished. Although the quantity of deposited content is one indicator of successful recruitment, quality is also necessary for repositories to further the cause of open access. Students, as well as their research supervisors, were concerned about the quality of their work that is available on open access. A faculty member, during an academic meeting, expressed her concerns that the ‘quality of some students work [dissertations] might not be appropriate for it to be available on open access.’

The researcher had also expected that submission barriers may inhibit the input activities of an IR. The major distinction here is between self-archiving, which refers to authors (or author’s representative) depositing their own work, and mediated archiving, which refers to authors submitting articles to IR staff. However, the vice-chancellor has made it a requirement that each faculty member submits the soft copies of theses and dissertations to the library and the library deposits the resources in the IR developed by the DLRG. As such, it has been decided that the IR will only include theses and dissertations.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Results of piloting the institutional repository (IR) after one year</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type of content</strong></td>
<td>The main focus of the IR holdings is on textual material. Contains theses, dissertations, conference proceedings (organized by the university), and articles deposited by research groups</td>
</tr>
<tr>
<td><strong>Copyright policies</strong></td>
<td>Submitters must ‘sign’ (that is, click through) a distribution rights agreement that affirms that the submitted item does not infringe on copyright or that copyright clearance has been obtained</td>
</tr>
</tbody>
</table>
| **Archiving policies** | 1. Self-archiving by faculty who are members of a recognized ‘community’ within the University.  
2. Opt-in versus Opt-out policy—to secure their consent, students are given a form to include theses and dissertations in the IR, and the choice of making them either private or accessible to the public  
3. Governance licences |
| **Advocacy** | Demonstrations performed at the internal level to faculty members and workshops to librarians  
Also planning to put more information on the library web site regarding IR and open access  
Presentations about IR to library and university committees and to departments identified as potential early adopters  
IR staff from the library will begin solicitation of other faculties |
| **Involvement of academics** | None so far |
| **Delivery process** | Populating the IR is not yet incorporated in the standing workflow and processes of the university. Depositing is currently done by the DLRG, and is likely be taken care of by the librarians later |
Through a discussion with the librarians, it has been agreed upon that populating the IR is not yet incorporated in the standing workflow and processes of the library, as ‘the system is not yet owned by the library’ (Deputy Chief Librarian, personal communication, 27 February 2009). The library maintains an electronic catalogue of theses, which provides the metadata and abstracts of all the Masters and PhD theses produced at the university. To simplify workflow, the IR librarian suggested that ‘the DOI be linked to our Pendeta WebPAC when our catalogers do the inputting of each thesis in the library catalogue’ (Head of Information Systems Department, personal communication, 8 August 2009). A mechanism has been established to capture new theses on a regular basis. Depositing is currently being done by the DLRG, and is likely to be taken care of by the librarians later. Table 1 presents the results of piloting the university’s IR, which has led to the strategies for content recruitment.

### Implication on librarians’ roles

Although IRs are gaining in momentum throughout academia, the faculty in this study seems to be cautious regarding IR contribution. The concerns related to IR among the faculties reflect to some degree the way in which repositories have developed in Malaysia, where for the most part IRs have been introduced for the worthy purpose of giving researchers a vehicle to enhance the availability of their publications by making them available via open access. The study has identified the following roles of librarians to make the IR service and content recruitment a success.

- **Understanding the software used:** Although IR technology was not a strong focus in the IR literature (Allard, Mack and Feltner-Reichert 2005), it is externally important that librarians have a full working knowledge of the software features. Tasks such as a database evaluation by comparing and contrasting the IR systems available based on criteria such as database structure, interface, search capabilities, special features, software requirements, speed and reliability, and export options need to be done before the final selection of the software.
- **Publicity and advocacy of IR:** The success of open access archiving in expanding access to scholarly works depends significantly on the author’s knowledge of open access, and the ready availability and accessibility of archives to authors. As Papin-Ramcharan and Dawe (2006) explain ‘If authors are unaware of the existence and benefits of archives then they cannot self-archive.’ The faculties in this study are poorly informed on IRs. Almost two-thirds do not know if their institution has one. This low level of awareness may result from one current strategy used by the university library to populate its repository in which librarians collect and deposit materials on behalf of faculty members. The deposited items are generally post-prints such as conference papers and journal articles. Therefore, faculty members may not realize that their materials are already in the library’s repository. The other reason is that the IR of the university has just been deployed and has not been widely publicized. As such, the librarians need to approach the faculty in a number of ways: on a one-to-one basis through informal conversations, small group discussions, departmental or faculty visits through the liaison librarians, and campus-wide promotion.
- **Establishing an institutional mandate:** All faculty respondents in this survey would comply with the university or research funder that required them to deposit copies of their scholarly work in the university’s repository. As IRs exist to serve the institution and funding bodies, rather than the individual, several institutions around the world have implemented such a mandate as recorded in the Registry of Open Access Repository Material Archiving Policies (ROARMAP). An institutional mandate might be successful in producing open access for the research-
intensive university in this study. There have been evidences demonstrating that institutions that have a mandatory policy have high proportion of published articles self-archived (Sale 2006), compared to those that have only voluntary policies (Suber 2006). The library needs to discuss the mandating of submissions in the IR and establishment of the self-archiving policies with the university’s top management. Genoni (2004) suggests that librarians should play the role of a content manager and establish relevant policies that cover what to include and exclude and how to determine copyright ownership for different types of documents.

- Educating faculty regarding self-archiving issues: The commonly expressed concerns regarding self-archiving are copyright and plagiarism. Considerable work has been done on copyright in association with the use of repositories to enhance open access for research outputs, especially published articles. Librarians need to seize every opportunity to inform the faculty members of the open access movement, the trends of open access publishing and increasing governmental and organizational support for IRs. Faculties need to be informed that over 90% of journals explicitly permit authors to self-archive their articles (Swan and Brown 2005), in most cases as post-prints (after peer review, in the form of the author’s final submitted manuscript). Educating the faculty regarding self-archiving issues needs to be undertaken to highlight the motivations for using the IR and to reassure faculty who may be worried about the deterrents. As such, to facilitate faculty to make an informed decision to deposit their work, the university’s IR would provide FAQs covering the following areas: ownership of copyright, protection of rights using Creative Commons license, plagiarism and file security. The IR would also need to provide a link to the SHERPA/RoMEO list of journals ‘publishers’ self-archiving policies (http://romeo.eprints.org).

- Submission review for content and metadata: When faculty self archive, they will also submit metadata. In an IR environment, librarians have to be responsible to determine the acceptable resources, metadata standards, review the content as well as the quality of metadata described by the authors. The number of staff assigned to manage the IR is likely to affect the visibility and growth of the repository, resulting in greater input and perhaps greater use. As such, librarians have to be included in the standing workflow and processes of populating the IR.

- Training of authors: Authors are a very important aspect of the IR. Librarians must actively pursue their role as educators to work with authors of intellectual works that will be contributing to the IR. This is a natural extension of the user training that librarians have provided for decades. Education would include helping the university community learn to use IR software for self-archiving. In addition, the training should include topics related to creating documents that can be more easily maintained in a digital environment, to issues surrounding digital preservation and providing guidance concerning metadata.

Populating the university’s IR through self-archiving has been a painfully slow and uphill process, similar to the process described in Chan, Kwok and Yip (2005). The IR started in January 2008 with a small collection of computer science theses and dissertations. They were mainly PDF files. Faculty members were invited to submit to the IR through official letters signed by the vice-chancellor. The response was not encouraging; only 12 submissions were received in the first three months after the IR was implemented at the faculty level in June 2008. The total number of direct submissions reached 64 by the end of 2008, which is by no means an encouraging figure. The author has reasons to believe that the academics have busy schedules, and will
consider self-archiving extra administrative work; however, many understand and support the idea of open access and IRs. Therefore, more aggressive strategies were adopted to populate the IR when faculty made it mandatory for students to submit an electronic copy of their theses and dissertations, and the DLRG initiated the submission of these resources in the IR. At the time of this study, the university’s DSpace held 338 items, which were organized into 17 collections within 18 communities and sub-communities. The IR has been piloted as a university-wide structure and publicized with an established skeleton of communities and collections already in place. However, only four faculty communities have been populated and 78% of these communities remain empty. Known as DSpace@UM, it is now available at http://dspace.fsktm.um.edu.my.

**Conclusion**

This paper reports on the strategies to establish a facility to provide open access to digital resources, and allow faculties, organizations, and individuals to cooperatively develop and upload resources to the IR. The results identify the approaches used by development teams in advancing the IR system and services in the university and the roles and competencies required in this new kind of professional library work. The analysis also includes important faculty perspectives on the value of IRs within the changing scholarly communication landscape, drawn from web-based survey with faculty depositors and interviews with librarians affiliated to a range of academic departments. The bottom-up strategy initiated by the research group involved in the implementation was to first publicize the potential benefit of the IR, apart from surfacing UM’s ‘hidden resources’ (which have not been tapped on a large scale so far) and low-cost interoperability among various faculties’ web portals. As the IR is indexed by Google and GoogleScholar, it helps increase the accessibility of scholarly works, which exist in a digital format, and make UM’s contributions to world literature more visible. Materials can be accessed more widely and also exploited for purposes different from those that originally motivated the creation of the repositories. Moreover, the possibility of accessing this repository enables the construction of new kinds of services that can better serve the needs of the learning community. The academic librarians will face a new challenge with IR because acquisition, a key aspect of the collection management process, is placed in the hands of the university community through the act of self-archiving. In fact, to populate an IR with content, faculty and students are actively encouraged to contribute their intellectual works. This characteristic of the IR suggests a fundamental shift in the collection management process, as authors of intellectual works are now responsible for selecting items that will be added to the collection, and for creating the initial surrogate record that facilitates access.

This study, based on a small set of survey data, has presented findings on faculty awareness and their use of open access repositories, the advocacy undertaken, and issues that may influence faculty’s motivation for IR contribution, which will lead to the actual deposit into the IR. Findings suggest that over one-third of the faculty respondents is unaware of open access and IR, or is aware of its existence but remain detached from it. However, faculties’ attitudes to the open access movement and IRs are generally positive—the majority acknowledges the importance of an IR and likes the idea of making intellectual output available through the university’s IR. The concerns that faculty members have regarding IR contribution imply that librarians have an important role to play with regard to relationship with self-archiving authors. This paper has identified the following roles and responsibilities of librarians in an IR environment: (1) understanding the IR software used; (2) publicity and advocacy; (3) establishing an institutional mandate;
(4) educating faculty regarding self-archiving issues; (5) submission review for content and metadata; and (6) training of authors.

Methodical IR development informed by best practices in the open access community, as well as findings from this study have been used for repository design customizations and functionality enhancements that complement the needs, interests, and concerns of the faculty. The IR development has been aimed at achieving near-term goals for building content and services in close consultation with faculty. The testbed is a collection of theses, dissertations, and articles by the Faculty of Computer Science and Information Technology, Faculty of Education, Faculty of Languages and Linguistics, and Faculty of Law community. Future work will involve promoting DSpace@UM to other faculties in the university to create communities and collaboratively develop collections of peer-reviewed post-prints of scholarly articles, image and video objects, learning objects, theses, and dissertations. It is hoped that more communities, especially within the science-based faculties, will be established in the IR. This is due to the fact that those publishing in the sciences such as molecular biology, physics, mathematics, and computer science were most likely to have published their work via an open access repository, as reflected from in open-ended responses in the web-based survey, which listed Genbank, EMBL, MiRbase, arxiv.org, and E-LIS as the avenues for archiving (Abrizah 2009).

Preliminary findings have shown that an IR is an extremely worthwhile endeavour, and is a viable proposition for the university’s support for a new pattern of scholarly communication, apart from surfacing its scholarly works and low-cost interoperability among various faculties’ web portals. It is hoped that this IR will increase the accessibility of scholarly works, which exist in digital format and make the university’s contributions to world literature more visible. However, as evidenced by other studies (Davis and Conolly 2007) and verified again by this initiative, faculty output is not finding its way into the university’s IR in large numbers (see http://dspace.fsktm.um.edu.my). The prevalence of peer-reviewed work nationwide and the well-documented difficulty in recruiting works of any type are not currently facilitating significant inroads into the open access movement. However, at this stage, the success of the institution in implementing an IR, as gauged by the criteria in this study, should provide hope to later entrants into the community and should influence the way we evaluate the potential of these repositories in Malaysia.

Acknowledgements

This research was conducted with the support of an UMRG grant (RG001/09ICT) from the University of Malaya (UM). The author thanks colleagues from various faculties for their participation, the UM Library, and members of the DGLR Group for their invaluable contribution in this research project.

References


Piloting an institutional repository at a research-intensive university


Pelizzari E. 2005. **Harvesting for disseminating: open archives and the role of academic libraries.** *The Acquisitions Librarian* **33/34**: 35–51


Suber P. 2006. Ten lessons from the funding agency open access policies. SPARC open access newsletter, Issue #10. Details available at <http://www.earlham.edu/~peters/fos/ newsletter/ 08-02-06.htm#lessons>, last accessed on 13 July 2006


Ware M. 2004. Institutional repository and scholarly publishing. Learned publishing 17: 115–112
Ontology-based visualization and navigation in an online digital library

Surjeet Mishra, Innovation Labs (Delhi), Tata Consultancy Services, Plot No.249, D&E Udyog Vihar, Gurgaon – 122 001, India
surjeet.mishra@tcs.com

Hiranmay Ghosh, Innovation Labs (Delhi), Tata Consultancy Services, Plot No.249, D&E Udyog Vihar, Gurgaon – 122 001, India
hiranmay.ghosh@tcs.com

Abstract
In this paper, we present a novel approach to visualize and navigate the document collection of a DL using ontology. We assume a DL corpus as a network of a large set of information nodes. We use domain ontology to depict the background knowledge organization and map the multimedia information nodes on that knowledge map, in order to make the implicit knowledge organization in a collection explicit. The ontology is automatically created by exploring the Wikipedia knowledge hierarchy, and is delimited to tightly cover the information nodes in the library’s collection. We present an abstract view of the knowledge map, clear and concise, which can be progressively ‘zoomed in’ or ‘zoomed out’ to navigate the knowledge space. We place the ontology nodes on the screen, based on their mutual similarity scores for aiding the cognitive process during navigation. The system provides an interactive navigation interface to the users.

Keywords
Effective visualization, navigation in multimedia collection, knowledge organization, semantic web, Wikipedia-based ontology creation, semantic DL.
Introduction

There are a large number of digital documents (text, video, and so on) available on the Internet. As this number is increasing fast with each passing day, it has become crucial to organize these multimedia documents and provide effective tools to browse through the collections. Digital libraries (DLs) serve the purpose of organizing and preserving the digital artefacts, and also make them easily accessible to the users. DLs have evolved a lot from their early forms in every aspect, whether it is in their collection type (from text files to video and now any multimedia format) or the type of services they provide to their users, which include various access paradigms. The oldest DLs were digital text archives like the Oxford Text Archive,¹ which consisted of text documents only. But currently the modern public libraries, such as the European Navigator² and the New York Public Library,³ hold vast amounts of information in different multimedia forms along with sufficient metadata.

Electronic libraries can be broadly classified on the basis of the domains and the documents present in their collection, as domain-specific DLs and broad-spectrum DLs. The domain-specific libraries contain documents on some specific domain or subject. The users of such a library are a specific group of professionals (for example, historians and doctors), which has expertise in the domain knowledge. However, a broad-spectrum or non-domain-specific library is more assorted in the domain or the documents present in it. Most of the widely used libraries are broad-spectrum libraries. Their users are much more diversified and may not have enough domain knowledge to query the system. Access to digital materials continues to be an issue of great significance in the development of DLs. Many online libraries provide some mature retrieval tools and classification hierarchy for accessing the information nodes (documents). While these aids support access to specific information, they do not provide an overview of the collection, thus making the exploration difficult for novice or casual users.

In the scenarios discussed above, an information visualization system allows the user to express his/her needs by directly interacting with the view, unlike in a retrieval process. Earlier works on collection and visualization reorganize the document collection to create different views of the collection, based on classical distance measures of information retrieval theories (Paulovich and Minghim 2008; Berendonck and Jacobs 2003). These distance measures can, however, be computed for text documents alone and thus the algorithms cannot be extended to predominantly multimedia collections of DLs like the European Navigator and the New York Public Library.

We consider that the underlying knowledge organization structure is important for visualizing a collection. We present a novel user interface that facilitates visualization of and navigation in a large multimedia collection constituting a DL. A document collection can be assumed to be a network of large number of information nodes of different media forms. For the successful implementation and use of any DL, the knowledge structure needs to be organized. The knowledge organization of a library is done using one or more knowledge organization techniques for organizing information and promoting knowledge management. Knowledge organization systems include highly structured vocabularies (such as thesauri) and less traditional schemes (such as semantic networks). Since these knowledge organization systems are means to organize information, they are at the heart of every library (and archive). The semantic web, which structures the concepts in a

¹http://www.ota.ox.ac.uk/ (last accessed on 15 October 2009)
²http://www.ena.lu (last accessed on 15 October 2009)
³http://www.nypl.org (last accessed on 15 October 2009)
document as a network, is one such mechanism for knowledge organization. We have used ontology as the mechanism to organize the knowledge structure. Ontology is a formal representation of a group of concepts on a domain and the relationships between them. It can be used to define the domain. Earlier works has been done to visualize a DL collection using manually created ontology (Abascal-Mena and Rumpler 2007), which imposes the burden of creation on the user (content creator), as well as the set of concepts (information nodes) get limited to the creators’ view and understanding of the documents in the collection.

We propose to construct an ontological map of the collection (domain) and map the information nodes (individual document) on that ontological map, thereby making the implicit knowledge organization in a collection explicit. The ontology is automatically created by analyzing the links in a large public knowledge resource, namely the Wikipedia, and is delimited to tightly cover the information nodes in the collection. Even with such constraints, the ontology contains several thousand nodes, which cannot be visualized together clearly on a knowledge map. Several approaches to large ontology visualization have been presented by Katifori, Halatsis, Lepouras, et al. (2007) and Geroimenko and Chen (2006). Some semantic DLs, such as Queens Library4 (Figure 1), use the semantic web to present the localized view of the knowledge map around a central concept rather than an overview of the collection to the user. We have developed a new method to present an abstraction of the knowledge map, which can be progressively ‘zoomed in’ or ‘zoomed out’ to navigate the knowledge space of the collection, as on a geographical map. Our approach is unique, in that it presents abstract views of the knowledge structure at different levels, does not clutter the display with too many nodes, and enables flexible zoom and navigation operations.

We have developed a new method to present an abstraction of the knowledge map, which can be progressively ‘zoomed in’ or ‘zoomed out’ to navigate the knowledge space of the collection, as on a geographical map. Our approach is unique, in that it presents abstract views of the knowledge structure at different levels, does not clutter the display with too many nodes, and enables flexible zoom and navigation operations.

Figure 1 Localized view of knowledge organization in Queens Library

In our approach, a casual visitor is greeted with the overall structure of the collection, enabling him to deep-dive into some broad areas of interest, progressively refine his information needs, and contextually discover the documents in the collection. The user interacts directly with the graph to express his needs. Lateral navigation in the knowledge map enables the user to explore the relationship between the topics and to discover related information nodes. Our main claims in this paper are (1) automatic organization of the information nodes present in a collection in an ontological structure and (2) intelligent creation and presentation of suitable abstractions of information structure to provide a browsing interface that facilitates the navigation process in the e-libraries (electronic/DLs).

4 http://www.queenslibrary.org (accessed on 15 October, 2009)
This paper describes our work in detail and provides information of implementation as well with some illustrative examples. It also discusses the real-life applicability of the system. The last section concludes the paper.

Description of work
Our work has two major components, namely (1) to create ontology covering all information nodes (documents) in a collection from Wikipedia and (2) provide abstract views and navigation facilities in the collection with the ontology. We describe these components in detail below.

Ontology creation
The documents in the collection need to be inter-related on a knowledge space for effective visualization and navigation. The ontology is a formal tool to represent a bounded knowledge domain. Creating a formal ontology for an ill-defined domain such as the collection of a video library or bookstore is a challenging task. Of late, there has been significant research interest in creating informal ontology (often called folksonomy) by harnessing the links in large social networking applications (Hepp, Bachlechner, and Siorpaes 2008). Wikipedia is an encyclopedia on the web, created, and maintained by public at large, and its categorization structure represents a knowledge organization, evolving over time. Every article in Wikipedia represents a topic, which belongs to one or more categories. Each category (except the root of the hierarchy, Category: Contents) is in turn defined as sub-categories of other categories, facilitating navigation across related subject areas. Our use of Wikipedia to create an informal ontology is motivated by Suchanek, Kasneci, and Weikum (2008). However, we have created a constrained view of the ontology by restricting it to contain the category nodes that are just sufficient to cover the information nodes in the collection. There are two top categories to start browsing the Wikipedia knowledge network: (1) fundamental and (2) main topic classifications (MTC). The second one is more detailed and covers enough main topics so that all single articles fit into one of the topics under this category. Therefore, we treat the category node MTC as the root element of our ontology.

Generally, some metadata is associated with the documents in a collection. In this work, we assume sufficient metadata to exist to relate a document to one or more Wikipedia topics. The important topics related to a document are sometimes explicitly given as the tags attached to the document. In the absence of such explicit association, the related topics can be extracted from the metadata (free text annotation or description) of the documents using different text-processing techniques. Once the topics pertaining to a document have been identified, we create the ontology superstructure for that document in a bottom-up manner, starting from a topic and going up the Wikipedia category hierarchy ladder using depth first traversal (DFT), till we reach the MTC node. This category node is the lowermost ancestor of all topic nodes in Wikipedia and we treat this as the root node for constructing the ontology. This process is repeated for every topic in a document and for every document in the collection, merging the common sub-graph as soon as it is discovered. Thus, a document is associated with one or more leaf category nodes in the ontology (Figure 2). The incremental method of building helps in updating the ontology when new documents are added to a collection.

We represent the knowledge graph as a singly rooted directed acyclic graph (DAG), with the MTC node as the root, to restrict the user from looping in cycles during navigation. The set of vertices comprise a set of documents D and a set of categories C. The set of edges comprise two types of edges: (1) ‘is-a’, which connects a category node to another, and (2) ‘instance-of’, which connects a document node to a category node. We represent these relations by the symbols \( \subset \) and \(<\), respectively. Note that the category nodes are
Ontology-based visualization and navigation in an online digital library

We have observed the following two types of transitive properties.

1. Let $c_1, c_2, c_3 \in C$. If $c_1 \subset c_2$ and $c_2 \subset c_3$, then $c_1 \subset c_3$

2. Let $c_1, c_2 \in C$, $d \in D$. If $d < c_1$ and $c_1 \subset c_2$, then $d < c_2$

These properties have been used to abstract the knowledge map at multiple levels and to allow 'zoom-in' and 'zoom-out' operations, as explained below.

**Navigation interface**

The main objective of creating the knowledge map or the ontology is to use it as the base to visualize the document collection to the DL users. But the knowledge map so created contains a few thousands of category nodes, posing a challenge to visualization. Several techniques have been developed to visualize very large (ontology) graphs. The folding and unfolding of subject taxonomies provide a shallow overview of the knowledge organization and may pose a significant challenge to a casual user to navigate (Katifori, Halatsis, Lepouras, et al. 2007). Hyperbolic distortion methods present nodes of interest at the centre of view with magnification while pushing the other nodes to the periphery as shown in Figure 3(a) (Katifori, Halatsis, Lepouras, et al. 2007; Heymann, Tham, Kilian, et al. 2002). The distorted view prevents a user from visualizing the overall knowledge structure. Another common approach of visualizing a large graph is clustering of nodes as shown in Figure 3(b) (Berendonck and Jacobs 2003; Katifori, Halatsis, Lepouras, et al. 2007). But, semantic labelling of the clustered nodes is a challenge.

**Figure 2** Knowledge organization in ontology

**Figure 3** Graph visualization techniques: (a) Hyperbolic distortion or fisheye layout (b) Clustered layout
For providing a better cognition to the user, we present an abstract view of the ontology by creating a view with a few selected category nodes and edges that depict the overall structure of the graph. A graph structure is characterized by the sub-graph of each of its nodes. We use Strahler Score (Delest, Herman, and Melancon 1998), which is a measure of the structural complexity of a node’s sub-graph (Auber 2002), to select the significant nodes. These set of nodes is called the summary nodes and is designated by $V'$. The number of nodes to be depicted is guided by the screen area allocated to the graph. We construct the summary graph by drawing some edges to connect these nodes. There is an edge $v_1 \rightarrow v_2$ ($v_1, v_2 \in V'$) if there is a path from $v_1$ to $v_2$ as the semantic relationships hold transitivity.

Wikipedia category nodes show high degree of interconnectivity. The large amount of edges present in the summary graph clutters the view. Therefore, we create a minimum weighted DAG around the selected nodes by using Chu-Liu/Edmond’s algorithm to get the abstract view graph (Chu and Liu 1965; Edmonds 1967). Visualization is further aided by placing the category nodes with higher semantic similarities closer to each other. The user can click on any of the nodes depicted on the screen to navigate on the ontology graph. As a user zooms in, the topics below the selected node get magnified, the topics in the other regions shrink, and some of the dynamically created document clusters split. Thus, as a user navigates the collection a contextual view of the collection is dynamically created. Completeness of the summary view demands that all the category nodes in the original graph should be navigable. To guarantee completeness, we add the immediate parent and child nodes of the node currently being explored to the dynamic knowledge graph. Once the category graph is organized, the cluster of document nodes is attached to each category node signifying the number of documents related to that category. Figure 4(a) and (c) depict the overview of the collection and a view when an intermediate node is being explored. Some navigation examples are given later in this paper.

**Figure 4** Abstract views of: (a) overview of ontology, (b) final view on ‘Tourism in India,’ and (c) intermediate navigation stage.
Implementation details
This work has been prototyped with a corpus of over 650 documents. The domain ontology is created using JWPL\(^5\) (Java Wikipedia Library) Application Programming Interface (API) to work on the Wikipedia dumps. The graph visualization has been implemented with JUNG 2.0\(^6\) (Java Universal Network/Graph Framework)—a software library that provides an extendible language for the modelling, analysis, and visualization of data that can be represented as a graph or network for graph visualization. The complete application has been developed in Java. Figure 5 shows the trend in the increase in number of ontology nodes with an increase in the number of documents in the collection.

Though there is an initial quick rise in the number of ontology nodes, but it tends to saturate at a finite value, after sufficient numbers of documents are inducted, thereby ascertaining tractability of the problem.

Four illustrative examples of navigation
We illustrate the navigational steps with the following few illustrative examples.
1. To discover the documents (‘History and Ideology’ and ‘India’) related to tapestry, a user may traverse the path:
   MTC → Arts → Visual arts → Textile arts → Tapestries
2. To discover documents on Tourism in India, a user traverses the path:
   MTC → Geography → Geography by place → Geography by countries → Members of the Commonwealth of Nation → India → Tourism in India

Some of the navigational steps are shown in Figure 4.

The black bubbles indicate the category nodes and the grey bubbles indicate the cluster of documents associated to that category, labelled by the number of documents in the cluster. The cluster expands on clicking and each document can then be accessed.

Applications
This work can be used in the DL visualization systems. As this work handles all types of multimedia contents, it can also be used to build various applications dealing with multimedia assets in online multimedia (books, movies, and so on.) stores, video sharing portals, and so on. Many ontology-based works are done with manually created ontology and thus the semantics is not used to the fullest, that is, the semantic relations are limited to the ontology creator’s perspective of the documents in the collection. The ontology creation module of the presented work can be utilized in ontology systems to create domain ontology from

---

\(^5\) http://www.ukp.tu-darmstadt.de/software/jwpl/
\(^6\) http://jung.sourceforge.net/index.html
Wikipedia. The system can be extended to work with document segments (chapters from literature artefacts and scenes of videos) as it can handle any form of information nodes, provided there is metadata attached with the segments.

**Conclusion**

The work presented in this paper provides a novel interface for visualization of and navigation in the document collection of a DL, consisting of different types of documents and multimedia content, using enriched ontology created automatically from Wikipedia. This work is applicable to any multimedia collection as it can handle all types of multimedia contents and is suitable for all DLs. The system does not require the users to have domain knowledge and hence it serves the diversified user base that actually uses the electronic libraries available over the Internet. As the system allows hopping from one concept to another related concept, which may not be directly connected in the hierarchy, the navigation is faster as compared to traditional taxonomical (hierarchy-based) navigation as it permits.

**Acknowledgement**

We thank ‘Scholars without Borders’, an online store for scholarly publications, for making its collection available for this research.

**References**


Ontology-based visualization and navigation in an online digital library


Harvesting capability of Google scholar: a comparative study of three major journal lists—Directory of Open Access Journals, Highwire, and BioMed Central

Rosy Jan, Assistant Professor, Department of Library and Information Science, University of Kashmir
rozy@kashmiruniversity.ac.in

Nadim Akhtar Khan, Assistant Professor, Department of Library and Information Science, University of Kashmir
nadim@kashmiruniversity.ac.in

Abstract
Google is considered to be a monolithic Internet power. Among its various existing products, Google Scholar is an incredible tool that allows researchers to search for a wide array of scholarly literature across the web. The study is a sample test of Google Scholar’s credibility and harvesting capability by submitting select titles from the field of microbiology in three major, renowned, and open-access journal lists—Directory of Open Access Journals (DOAJ), Highwire, and BioMed Central. The findings reveal that out of the 354 submitted titles, only 16.38% were clearly identified. The highest number of titles, that is, 52.32%, was retrieved from BioMed Central, followed by 25.37% from Highwire, and 21.7% from DOAJ. They provide most (65.50%) of the record as full text, followed by links (32.71%) and a negligible ratio of records as citations (1.78%). The study further reveals that some results are available under multiple URLs, thus leading to duplication. The highest percentage of results is from open-access web servers (49.64%). A total of 10.63% of web servers were found to be present in foreign languages, followed by an organizational contribution of 9.92% and 2.12% from conventional data providers.

Keywords
Google Scholar, scholarly literature, DOAJ, BioMed Central, Highwire
**Introduction**

Google Scholar is a subset of the larger Google search engine. It is a freely accessible web search engine that indexes the full text of scholarly literature across an array of publishing formats and disciplines. Still in Beta version, Google Scholar had its debut started in November 2004 (Google 2009).

Google Scholar covers a great range of topical areas. It appears to be strongest in the sciences, especially medicine and social sciences (Vine 2006). Google Scholar enables the user to search specifically for scholarly literature, including peer-reviewed papers, theses, books, preprints, abstracts, and technical reports from all areas of research in a simple way. It allows one to find articles from a wide variety of academic publishers, professional societies, preprint repositories and universities, as well as scholarly articles available on the web (Mayr and Kathrin 2007).

Google Scholar searches for terms from the full text of scholarly sources, not just abstracts or titles, and provides a high level of knowledge discoverability for many interdisciplinary topics. Google Scholar is a portal to both commercial and open-access materials on the open web and repositories (Hartman, Mullen, and Bowering 2008).


**Selected journal lists**

The journal lists contain collections of scholarly full-text documents maintained by government agencies, associations, universities, professional volunteer groups, as well as new and traditional scholarly publishers (directly or indirectly through their digital facilitators). The selected journal lists include Directory of Open Access Journals (DOAJ), BioMed Central, and Highwire.

**Directory of Open Access Journal**

The DOAJ is a wonderful research tool for users wishing to quickly search many open-access journal articles. The vision is to contribute to the movement towards free unlimited access to scientific results. The service aims to give users a single window to quality-controlled, scientific open-access journals in full text (DOAJ 2009).

**BioMed Central**

BMC (BioMed Central) is a United Kingdom-based, peer-reviewed, open-access, online publisher specializing in medical research and biology. The BMC publishes over 180 scientific journals, and describes itself as the first and largest open-access science publisher (BioMed Central 2009).

**Highwire**

Highwire produces the online versions of peer-reviewed journals and other scholarly content. Highwire partners with influential scholarly societies, university presses, and publishers create a collection of the finest, fully searchable research and clinical literature online. Together, these partners produce 71 of the 200 most frequently cited journals published in science. As of May 2009, they host the largest repository of high impact, peer-reviewed content, with 1245 journals and 5,909,326 full-text articles from over 140 scholarly publishers. Highwire-hosted publishers have collectively made 1,894,133 articles free (Highwire 2009).

**The problem**

Googlism has emerged as a synonymous term for searching the web. The scholarly community relies on Google, particularly Google Scholar, which is intended for and claims to search exclusively for scholarly documents. This study has been undertaken to understand and
investigate the coverage of scholarly literature by Google Scholar and to explore the deficiencies in the coverage. It intends to find out (1) the web server that is the most important data provider and (2) the information sources that are well represented, and their open-access content accessibility.

**Scope**
The study undertaken has been confined to a freely accessible web search engine from the Google’s scholarly index—Google Scholar. Three major open-access journal lists – BioMed Central, DOAJ, and Highwire – were used as a platform for accessing the journal titles to ascertain the harvesting capability of Google Scholar. The field of microbiology is chosen from all the three repositories.

**Objectives**
- Determine the harvesting capability of Google Scholar from three major journal lists
- Find out the number of journal titles in the first 10 or above 10 results found in Google Scholar
- Find out the document type of the record
- Analyse the web server of the record as to whether it is a commercial or a non-commercial one

**Methodology**
Google Scholar offers a simple way to index the contents of a large number of journal lists. As it was impossible to query all the available journal lists, the selection was made from three major journal lists, that is, DOAJ, BioMed Central, and Highwire. Precision was ensured by the use of ‘Phrase’ search as a base for extracting the titles from Google Scholar.

The study was further divided into the following steps.

**Step 1: Querying the journal title** Titles from three selected journal lists were queried to determine the coverage of Google Scholar.

**Step 2: Downloading the Google Scholar research page** A maximum of first hundred records were analysed and downloaded for every journal title to be processed.

**Step 3: Analysis and aggregation of extracted data** The extracted data was aggregated using simple counts. Journal title that could be clearly identified and not identified were counted. For each result matched to the journal, the following fields were identified.
- Title and document type of the record, that is, link, citation, PDF-Link and in other formats, such as PS, DOC, RTF.
- Domain of the web server, that is, the main source for each linked record.

**Literature review**
A comprehensive literature search was undertaken multi-dimensionally. The search reflected various facets of Google Scholar, visualizing its various potentialities as a scholarly tool accompanied with some other deficiencies. However, only the titles that coincided with the topic discussed here were incorporated in the literature review, seen below

Noruzi (2005) revealed that Google Scholar provides a free alternative or complement to other citation indexes or databases. The study explained that Google Scholar has the advantage of giving multi-disciplinary coverage. In this study, several suggestions have been given for improving Google Scholar. Pomerantz (2006) analysed that Google Scholar is the latest tool information-seeking technologies that increasingly realizes the goal of achieving 100% availability of information. However, Google Scholar does not provide access to 100% of information resources in existence; but rather enables the discovery of information resources, and allows for the possibility that these resources will be discoverable by the user. Another study evaluated the strengths and weaknesses of Google Scholar in comparison to that studied by Pubmed and Scirus by running simple tests.

of its coverage. The study reveals that Google Scholar cannot be dismissed outright. The study of Sadeh (2006) accords Google Scholar a unique position among other scholarly resources when compared to other metasearch systems. However, metasearch systems have several advantages over Google Scholar. Walters (2007) analysed that the Google Scholar indexed the greatest number of core articles (93%) and provided the most uniform publishers and data coverage by evaluating its content with seven other databases (Academic Search Elite, Ageline, Article First, GEOBASE, POPENLINE, Social Science Abstracts, and Social Science Citation Index). Haya, Nygren, and Widmark (2007) analyzed that Google Scholar performed better in almost all measures. Responses to Google Scholar were more positive than those to Metalib. However, the students were not overwhelmingly enthusiastic about either of the tools. Norris, Oppenheim, and Rowland (2008) demonstrated the relative effectiveness of the search tools in finding out open-access versions of peer-reviewed academic articles and found that among the general search engines, Google and Google Scholar ranked first in their retrieval as compared to Open DOER and OAIster.

Bakkalbasi et al. (2006) carried out citation analysis in an observational study examining three databases, that is, Google Scholar, Scopus, and Web of Science, and did not identify any of the above as an answer to all citation tracking needs. The study indicated that Google Scholar alone might not replace other scholarly search tools. Mayr and Walter (2007) showed the deficiencies in the coverage and up-to-dateness of the Google Scholar index by putting the queries against five journal lists. The study identified noted the most important web servers as the data providers for this search engine and the information sources that are well represented. The conclusion that was drawn revealed weaknesses in the accessibility of the open-access content and lack of transparency and completeness. Shultz (2007) observed that Google Scholar does not appear to be a replacement for PubMed, though it may serve effectively as an adjunct resource to complement databases with more fully developed searching features by comparing test searches between Pubmed and Google Scholar. However, the suggestion was made to repeat the study in one or two years to determine if further refinements have improved their performance. The study by Neuhaus and Daniel (2008) provided an overview of the citation-enhanced databases, such as Chemical Abstract, Scopus, and Google Scholar, when used as a data source for performing citation analysis. The study did not find Google Scholar to be a useful tool for citation analysis in its beta version.

Data analysis

Identification of journal titles in Google Scholar

A general overview is presented in Table 1 for all the selected journal lists showing the number of journal titles being identified and not identified by Google Scholar from the results received. The total number of titles selected from the directories/repositories is classified as ‘titles selected’. The titles identified by Google Scholar are represented as ‘clearly identified’. All other title that are not identifiable by Google Scholar art shown as ‘not clearly identified’.

Table 1 represents the total number of journal titles found in three journal lists: Highwire, representing the maximum number of 282 titles; BioMed Central, having 37; and DOAJ with 35 titles in the selected field microbiology. Of the total titles (354) submitted to Google Scholar, only 16.38% were clearly identified. The highest percentage (53.44%) was identified from BioMed Central followed by Highwire (24.13%), while the least percentage (22.41%) was indexed from DOAJ. However, most of the journal titles (83.61%) could not be identified from the results delivered from Google Scholar.
Harvesting capability of a Google scholar: a comparative study of three major journal lists

Table 1 Identification of journal titles in Google Scholar

<table>
<thead>
<tr>
<th>Lists</th>
<th>Titles selected</th>
<th>Clearly identified</th>
<th>Not clearly identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Directory of Open Access Journals</td>
<td>35</td>
<td>13 (22.41)</td>
<td>22 (7.43)</td>
</tr>
<tr>
<td>BioMed Central</td>
<td>37</td>
<td>31 (53.44)</td>
<td>6 (2.02)</td>
</tr>
<tr>
<td>Highwire</td>
<td>282</td>
<td>14 (24.13)</td>
<td>268 (90.54)</td>
</tr>
<tr>
<td>Total</td>
<td>354</td>
<td>58 (16.38)</td>
<td>296 (83.61)</td>
</tr>
</tbody>
</table>

**Note** Figures in parentheses indicate percentage

Distribution of document types

The hits retrieved on first five pages of Google Scholar were analysed in terms of the document types categorized as follows: links, citations, and full texts (the document type 'Link' is just a literature reference with abstract; 'Citations,' an offline reference to a particular work with bibliographic information, are denoted by a prefix 'CITATION,' while the records prefixed by PDF or an arrow indicates direct access to 'Full Text'). Every record was only counted as a hit when the exact title was found.

A total of 2751 records or journal articles—1456 from BioMed Central, 698 from Highwire, and 597 from DOAJ —were analysed as Google Scholar hits. From the analysed data, the records with full text availability were 65.50%. A comparatively low ratio was found for links (32.71%). However, the results for the records with citations were negligible reaching only 1.78%, as shown in Table 2.

As clearly indicated in the table above, BioMed Central showed the best coverage of the records (52.92%) delivered from Google Scholar, while a relatively low percentage was found for Highwire (25.37%) and DOAJ (21.70%). Furthermore, Google Scholar generated most of the records as citations (85.71%), full text (56.77%), and links (43.44%) from BioMed Central. However, DOAJ and Highwire provided a comparatively low ratio of records as compared to BioMed Central. Maximum number of records was retrieved as full text, followed by links and then citations.

Distribution of web servers

Google Scholar in collaboration with the largest and most well-known scholarly publishers

Table 2 Distribution of document types

<table>
<thead>
<tr>
<th>Lists</th>
<th>Records (%)</th>
<th>Links (%)</th>
<th>Citations (%)</th>
<th>Full text (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highwire</td>
<td>698 (25.37)</td>
<td>304 (33.77)</td>
<td>4 (8.16)</td>
<td>390 (21.64)</td>
</tr>
<tr>
<td>Directory of Open Access Journals</td>
<td>597 (21.70)</td>
<td>205 (22.77)</td>
<td>3 (6.12)</td>
<td>389 (21.58)</td>
</tr>
<tr>
<td>BioMed Central</td>
<td>1456 (52.92)</td>
<td>391 (43.44)</td>
<td>42 (85.71)</td>
<td>1023 (56.77)</td>
</tr>
<tr>
<td>Total</td>
<td>2751</td>
<td>900 (32.71)</td>
<td>49 (1.78)</td>
<td>1802 (65.50)</td>
</tr>
</tbody>
</table>

**Note** Figures in parentheses indicate percentage
and university presses, both commercial and open source, makes its contents visible through web servers. The distribution of web servers is visible with their host names, description, and frequency, as shown in Table 3.

In Table 3, 25 frequently occurring web servers have been included out of the 141 hosts collaborating with Google Scholar. The description column categorizes various types of web servers, open-access (OA) publishers, and commercial publishers (CP). The frequency of the web servers at the top of the list, that is, BioMed Central, shows noteworthy cooperation with Google Scholar with an occurrence of 717 times, while as DOAJ they occurred 392 times out of the top 25 contributors to Google Scholar. The interpretation of the data shows only two data providers as commercial and the rest as OA.

**Table 3** Distribution of the top 25 web servers

<table>
<thead>
<tr>
<th>Sl no.</th>
<th>Web servers</th>
<th>Host names</th>
<th>Description</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>biomedcentral.com</td>
<td>BioMed Central</td>
<td>OA</td>
<td>717</td>
</tr>
<tr>
<td>2</td>
<td>doaj.org</td>
<td>Directory of Open Access Journals</td>
<td>OA</td>
<td>392</td>
</tr>
<tr>
<td>3</td>
<td>am soc microbial</td>
<td>American Medical Association</td>
<td>OA</td>
<td>369</td>
</tr>
<tr>
<td>4</td>
<td>pubmedcentral.nih.gov</td>
<td>National Institutes of Health</td>
<td>OA</td>
<td>258</td>
</tr>
<tr>
<td>5</td>
<td>microbialcell factories.com</td>
<td>Microbial Cell Factories</td>
<td>OA</td>
<td>66</td>
</tr>
<tr>
<td>6</td>
<td>archive.biomedcentral.com</td>
<td>BioMed Central</td>
<td>OA</td>
<td>61</td>
</tr>
<tr>
<td>7</td>
<td>cat.inist.fr</td>
<td>Institut de l'Information Scientifique et Technique</td>
<td>CP</td>
<td>57</td>
</tr>
<tr>
<td>8</td>
<td>ijpmonline.org</td>
<td>Indian Journal Of Pathology and Microbiology</td>
<td>OA</td>
<td>49</td>
</tr>
<tr>
<td>9</td>
<td>bioline.org</td>
<td>Bioline International</td>
<td>OA</td>
<td>43</td>
</tr>
<tr>
<td>10</td>
<td>soc leukocyte biology</td>
<td>The Society For Leukocyte Biology</td>
<td>OA</td>
<td>41</td>
</tr>
<tr>
<td>11</td>
<td>pdconnect.com</td>
<td>Paritonial Dialysis International</td>
<td>OA</td>
<td>35</td>
</tr>
<tr>
<td>12</td>
<td>scipub.org</td>
<td>Science Publications</td>
<td>CP</td>
<td>35</td>
</tr>
<tr>
<td>13</td>
<td>japr.highwire.org</td>
<td>The Journal Of Applied Poultry Research</td>
<td>OA</td>
<td>26</td>
</tr>
</tbody>
</table>

**Top 10 web servers per journal list**

Table 4 represents the top 10 most commonly occurring web servers analysed from each of the journal lists being studied. The web servers at the top show notable collaboration with Google Scholar.

It is observed that titles selected and included by Highwire are also placed in their databases by the host names included in column 1.
Harvesting capability of a Google scholar: a comparative study of three major journal lists

<table>
<thead>
<tr>
<th>Sl no.</th>
<th>Web servers</th>
<th>Host names</th>
<th>Description</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>aidstherapy.com</td>
<td>Aids Research and Therapy</td>
<td>OA</td>
<td>25</td>
</tr>
<tr>
<td>15</td>
<td>japr.fass.org</td>
<td>The Journal Of Applied Poultry Research</td>
<td>OA</td>
<td>24</td>
</tr>
<tr>
<td>16</td>
<td>archopht.highwire.org</td>
<td>Archives of Ophthalmology</td>
<td>OA</td>
<td>22</td>
</tr>
<tr>
<td>17</td>
<td>etc.online.com</td>
<td>Emerging Themes in Epidemiology</td>
<td>OA</td>
<td>22</td>
</tr>
<tr>
<td>18</td>
<td>mayoclinicproceedings.com</td>
<td>Mayo Clinic Research</td>
<td>OA</td>
<td>21</td>
</tr>
<tr>
<td>19</td>
<td>archsurg.highwire</td>
<td>Archives of Surgery</td>
<td>OA</td>
<td>18</td>
</tr>
<tr>
<td>20</td>
<td>internationalbreastfeedingjournal.com</td>
<td>International Breast Feeding Journal</td>
<td>OA</td>
<td>18</td>
</tr>
<tr>
<td>21</td>
<td>virologyj.com</td>
<td>Virology Journal</td>
<td>OA</td>
<td>18</td>
</tr>
<tr>
<td>22</td>
<td>malariajournal.com</td>
<td>Malaria Journal</td>
<td>OA</td>
<td>17</td>
</tr>
<tr>
<td>23</td>
<td>ojrd.com</td>
<td>Orphanet Journal of Rare Diseases</td>
<td>OA</td>
<td>17</td>
</tr>
<tr>
<td>24</td>
<td>peh-med.com</td>
<td>Philosophy, Ethics and Humanities in Medicine</td>
<td>OA</td>
<td>17</td>
</tr>
<tr>
<td>25</td>
<td>Pophealthmetrics.com</td>
<td>Population Health Program</td>
<td>OA</td>
<td>16</td>
</tr>
</tbody>
</table>

OA – open access; CP – commercial publisher

Table 4 Distribution of top 10 web servers per journal list

<table>
<thead>
<tr>
<th>Sl no.</th>
<th>Highwire</th>
<th>DOAJ</th>
<th>BioMed Central</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>am soc microbial</td>
<td>doaj.org</td>
<td>biomedcentral.com</td>
</tr>
<tr>
<td>2</td>
<td>cat.inist.fr</td>
<td>biomedcentral.com</td>
<td>doaj.org</td>
</tr>
<tr>
<td>3</td>
<td>soc leukocyte biology</td>
<td>pubmedcentral.nih.gov</td>
<td>pubmedcentral.nih.gov</td>
</tr>
<tr>
<td>4</td>
<td>pdiconnect.com</td>
<td>archive.biomedcentral.com</td>
<td>microbialcellfactories.com</td>
</tr>
<tr>
<td>5</td>
<td>japr.highwire.org</td>
<td>ijpmonline.org</td>
<td>aidstherapy.com</td>
</tr>
<tr>
<td>6</td>
<td>japr.fass.org</td>
<td>bioline.org.br</td>
<td>malettiemetaboliche.it</td>
</tr>
<tr>
<td>7</td>
<td>archopht.highwire.org</td>
<td>scipub.org</td>
<td>etc.online.com</td>
</tr>
<tr>
<td>8</td>
<td>mayoclinicproceedings.com</td>
<td>microbialcellfactories.com</td>
<td>ojrd.com</td>
</tr>
<tr>
<td>9</td>
<td>archsurg.highwire.org</td>
<td>immunome-res.com</td>
<td>peh-med.com</td>
</tr>
<tr>
<td>10</td>
<td>pdi.highwire</td>
<td>pathogens.plosjournals.org</td>
<td>malariajournal.com</td>
</tr>
</tbody>
</table>
pattern is observed in the case of DOAJ and BioMed Central.

**Types of web servers**

Table 5 shows the types of the web servers categorized as OA, commercial, organizational, foreign language, and unidentified with their numbers, respectively. The web servers found in languages other than English are shown under ‘foreign languages,’ while the web servers that could not be identified are classified as ‘Unidentified.’

The information obtained from Table 5 elucidates the highest percentage for the OA-web servers (49.64) followed by the unidentified ones (26.95). Only 10.63% of the web servers were found to be present in foreign languages. The organizational contribution was only 9.92%. However, a negligible ratio (2.12) was found for the commercial data providers.

<table>
<thead>
<tr>
<th>Types of web servers</th>
<th>Number (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open access</td>
<td>70 (49.64)</td>
</tr>
<tr>
<td>Commercial</td>
<td>4 (2.12)</td>
</tr>
<tr>
<td>Unidentified</td>
<td>38 (26.95)</td>
</tr>
<tr>
<td>Foreign language</td>
<td>15 (10.63)</td>
</tr>
<tr>
<td>Organization</td>
<td>14 (9.92)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>141</td>
</tr>
</tbody>
</table>

**Distribution of web servers in foreign language**

All the web servers listed in Table 6 are in foreign languages. Thus, the details could not be identified by the author. As evident from Table 6, 15 such web servers were found in the study, with malettiemetaboliche.it occurring 23 times.

<table>
<thead>
<tr>
<th>Sl. no.</th>
<th>Web servers in foreign languages</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>malettiemetaboliche.it</td>
<td>23</td>
</tr>
<tr>
<td>2</td>
<td>agopuntura.org</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>dld.go.th</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>faz.net</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>gami-mpl.ird.fr</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>medicacentar.info</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>paduaresearch.cab.unipd.it</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>proaders.cict.fiocruz.br</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>prof-tt-publichealth.com</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>rue89.com</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>sld.cu</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>tau.ac.il</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>unife.it</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>vliz.be</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>webdoc.sub.gwgd.dc</td>
<td>1</td>
</tr>
</tbody>
</table>
Distribution of unidentified web servers
Table 7 presents the web servers along with their frequencies, which could not be identified during the analysis of the data. There are 38 unidentified web servers. The website ncbi.nlm.nih.gov has the maximum frequency of 5, while the rest of the web servers range between 1 and 4.

<table>
<thead>
<tr>
<th>Sl. no.</th>
<th>Web servers</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ncbi.nlm.nih.gov</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>grande.nal.usda.gov</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>pubs.drdc.gc.ca</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>bioethicsforum.info</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>case.edu</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>hawaii.edu</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>psic.info</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>tspace.library.utoronto.ca</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>works.bepress.com</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>zevep.com</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>agbios.com</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>aidstawania.files.wordpress.com</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>array.bioengr.uic.edu</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>assets.O.pubget.com</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>barebeginnings.ca</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>carbon.structbio.vanderbilt.edu</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>cu.edu</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>donna.metzlabs.net</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>educatingharper.com</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>fstadirect.com</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>genome.imb-jena.da</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>hal.insern.fr</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>harm.live.radical.desighs.org</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>ids-health.com</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td>ihp.edu</td>
<td>1</td>
</tr>
<tr>
<td>26</td>
<td>indexmedicus.afro.who.int</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>m.ehime-u.ac.jp</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>micaplint.com</td>
<td>1</td>
</tr>
<tr>
<td>29</td>
<td>michaelsclander.com</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>obgyn.net</td>
<td>1</td>
</tr>
<tr>
<td>31</td>
<td>ocha.gwapps1.unog.ch</td>
<td>1</td>
</tr>
<tr>
<td>32</td>
<td>pdc.csusb.edu</td>
<td>1</td>
</tr>
<tr>
<td>33</td>
<td>people.stfx.ca</td>
<td>1</td>
</tr>
<tr>
<td>34</td>
<td>resourse.library.tme.edu</td>
<td>1</td>
</tr>
<tr>
<td>35</td>
<td>thuddle.net</td>
<td>1</td>
</tr>
<tr>
<td>36</td>
<td>uni-bielefeld.de</td>
<td>1</td>
</tr>
<tr>
<td>37</td>
<td>webra.cas.sc.edu</td>
<td>1</td>
</tr>
<tr>
<td>38</td>
<td>nyc.gov.sg</td>
<td>1</td>
</tr>
</tbody>
</table>

Organizational collaboration with Google Scholar
The organizational contributors to Google Scholar are clearly mentioned in Table 8. The results obtained from the table clearly indicate that the frequency of the organizational contributors lies between 1 and 4. The website
nbimcc.org and savethechildren.org occurred 4 and 3 times, while the rest occurred only once.

Findings and conclusions

All results and conclusions in this study are current and based on sample tests (100 hits per query) and are valid as of May 2009. Like the widely used, familiar search service Google Web Search, Google Scholar offers fast searching with a simple, user-friendly interface. The pros of this are that the search is free of charge and is done across interdisciplinary full-text collections. The Google Scholar approach offers some potential for literature retrieval, for example, automatic citation analysis and the ranking built up from this, and oftentimes direct downloading of full text, which is sometimes also described as a subversive feature. However, the service cannot be seen as a substitute for the use of special indexing and abstracting services and professional database as a very low percentage (16.38%) of titles were identified out of the coverage corpus selected. The various findings are listed below.

- Of the total titles (354) submitted to Google Scholar, only 16.38% were clearly identified.
- Out of the ‘clearly identified’ titles (58), Google Scholar provided most of the records as full text (65.50%), followed by links (32.71%) and a negligible ratio of records as citations (1.78%).
- Google Scholar showed the best coverage of the records (52.92%) from BioMed Central, while a relatively low percentage was found for Highwire (25.37%) and DOAJ (21.70%).
- Google Scholar generated most of the records as citations (85.71%), full text (56.77%) and links (43.44%) from BioMed Central. However, Google Scholar generates almost an equal percentage as links and full text from Highwire and DOAJ.
- Out of 141 hosts collaborating with Google Scholar, the frequency of the web servers at the top of the list, that is, BioMed Central shows noteworthy cooperation with the Google Scholar with an occurrence of 717 times, whereas the other two directories, that is, DOAJ and Highwire occurred 392 times and 66 times respectively out of the top 25 contributors to Google Scholar.
- Only two data providers were identified as commercial and the rest as OA among the top 25 web servers.
- It is observed that titles included by directories under observation are also placed in other databases, thus leading to the duplication of results in Google Scholar.
- The highest percentage of results is from OA-web servers (49.64%) followed by the unidentified ones (26.95).
- 10.63% of web servers were found to be present in foreign languages. Organizational contribution was only 9.92%.
Harvesting capability of a Google scholar: a comparative study of three major journal lists

However, a negligible ratio (2.12%) was found for the commercial data providers.

- The number of the unidentified web servers was 38.

The results show that the expanding sector of OA-journals (DOAJ list) is under-represented among the servers. Google Scholar claims to provide 'scholarly articles across the web,' the claim is true according to the study as the ratio of articles from OA open-journals or full-text journals is comparatively high. In comparison with many abstracting and indexing databases, Google Scholar does not offer the transparency and completeness that is expected from a scientific information resource. Google Scholar can be helpful as a supplement to retrieval in abstracting and indexing databases mainly because of its coverage of freely accessible materials.

References


A new framework to preserve Tagore songs

Chandan Misra  
Department of Computer Science and Engineering Civil Engineering, IIT Kharagpur  
Kharagpur – 721 302, India  
chandan.misra1@gmail.com

Baidurya Bhattacharya  
Department of Civil Engineering, IIT Kharagpur  
Kharagpur – 721 302, India  
baidurya@civil.iitkgp.ernet.in

Anupam Basu  
Department of Computer Science and Engineering, IIT Kharagpur  
Kharagpur – 721 302, India  
anupam@cse.iitkgp.ernet.in

Abstract
This paper deals with a new framework for archiving and annotating Rabindra Sangeet ('Rabindra Sangeet', n.d) (known as Tagore songs in English) written and composed by Rabindranath Tagore. This framework has a layered structure with each layer having a specific defined functionality. XML ('XML Tutorial', n.d) is used to construct the framework to store different synchronized layers of information in a simple manner. These layers contain notational, audio/MIDI ('Learn About MIDI', n.d), catalogue information including structure of musicsheet. MusicXML ('MusicXML Definition', n.d) and IEEE P1599 ('IEEE Standard P1599-2008', 12 June 2008) standard are two prominent structured formats for defining western music. The inadequate nature of these two formats for representing Indian music system (including Rabindra Sangeet) makes this new framework relevant. This paper gives a brief description of the new framework and a simple application for archiving Tagore songs in MIDI and notational format. There are two ultimate purposes of this work. The first one is to generate the MIDI files of all the songs with various instruments. The second one is to store the characters of the musicsheet in the XML file so that data can be retrieved and presented in a desktop and web environment.

Keywords
Rabindra Sangeet, layered framework, musicsheet, MIDI, API, Unicode
**Introduction**

Rabindra Sangeet, also known as Tagore songs in English, is a form of music composed by Rabindranath Tagore, who added a new dimension to the musical concept of India in general and province Bengal in specific. The grammar of Rabindra Sangeet closely follows North Indian classical music and inherits all the basic features, including *taal*, *maatra*, *raaga*, octaves, microtones, and so on, as discussed in (‘Hindustani Classical Music’, n.d). The musicsheet is differently structured than that in western, and scores and lyrics are written in Bengali—a regional language. This necessitated a new structured format that would fully describe Tagore songs. Songs mean not only MIDI or sound but also the representation of the musicsheet, which is a container of scores and lyrics. More precisely, songs are represented here in MIDI, as well as in a human-understandable, non-audio form like musicsheet.

This paper deals with the design of an XML database where the symbols of the actual musicsheet are stored as characters. Previously, the musicsheets were preserved in books or as scanned images on the web. The symbols are transformed into fonts and Unicode (‘What is Unicode?’, n.d) and stored into an XML file. As this is an archiving work, much other information like metadata of a song, notational information to generate MIDI, and so on, is necessary. Therefore, this framework is structured in different layers, each of which includes different information. A single XML file contains catalogue information of the song, the characters of the musicsheet, MIDI information like volume, pitch, frequency, note number of the song to be played, and so on. The stored data are retrieved to generate sound and to display in the web application. As music-related data are heterogeneous in nature (Haus and Ludovico 2005) with different encoding formats, this XML data should be synchronized. Basically, this framework is built upon the idea of MusicXML and IEEE P1599. Nevertheless, a new framework is needed for several reasons. The prime requirement of the work is to write the musicsheet using Bengali fonts and Unicode. Changes in the existing framework are required to make the musicsheets resemble the original manuscripts archived in books or on the Web (Haus and Ludovico 2005).

**Why is a new framework needed?**

The science of music differs from place to place and culture to culture. Therefore, there are many distinctions between Western and North Indian classical music. A new format was necessary to encapsulate the new features. The major differences are described below.

In Rabindra Sangeet, there are seven *swara-s* or notes. The names of the seven *swara-s* are *Shadja* (*Sa*), *Rishabh* (*Re*), *Gandhãr* (*Ga*), *Pancham* (*Pa*), *Dhaivat* (*Dha*), and *Nishãd* (*Ni*). The group of seven notes is called *Saptak* (*Sapt* = seven).

If we consider the *C* note (the first white key) on a piano or keyboard as the tonic or the *Sa*, then the position of the other notes would be as follows. The seven *swara-s* occupy the seven white keys of the keyboard. This is shown in Figure 1.

It is seen in Figure 1 that in-between the seven *swara-s* we have five intermediate notes. These notes are called *vikrit swara-s* (altered notes). In this context, the original notes are referred to as the *shuddha swara-s* (pure notes).

![Figure 1 The seven white keys of the keyboard representing the seven swara-s](image-url)
The swara-s between the pairs Sa and Re, Re and Ga, Pa and Dha, Dha and Ni are known as komal Re, komal Ga, komal Dha, Komal Ni, respectively. The term komal means soft or flat. The swara between Ma and Pa is called kari Ma; kari means sharp. The swara-s Sa and Pa do not have altered forms and they are known as achal or immovable swara-s. Table 1 shows the 12 swara-s, their western counterparts, and the notes if C is taken as the Sa.

There are eight more notes in Rabindra Sangeet, which are not present in western music. The swara-s between the pairs Sa and komal Re, Re and komal Ga, Pa and komal Dha, Dha and komal Ni, komal Re and Re, komal Ga and Ga, komal Dha, and Dha, komal Ni and Ni are known as atikomal Rishabh, atikomal Gandhar, atikomal Dhaivat, atikomal Nishad, anukomal Rishabh, anukomal Gandhar, anukomal Dhaivat and anukomal Nishad.

It should be noted here that this analogy to the keyboard is not technically correct. In western music, the instruments are tuned to the chromatic or even tempered scale, whereas in north Indian classical music the notes are based on the natural or diatonic scale.

Another thing to note is that, in western music, the frequencies of the notes are fixed. For example, the A of the middle octave is supposed to have a frequency of 440 Hz. This is not the case in Indian music. The relative frequency, and not the actual frequency, is of importance here. In the above example, we have taken C as our tonic (Sa) but we could consider any note as Sa and the pitch of the other notes will be relative to the Sa. The ratios of frequencies for various notes with respect to the Sa are also shown in Table 1.

One of the requirements of preserving Tagore songs is to make the musicsheet same as the original. The symbols are totally different from western music. Here, staff notations are not used. Therefore, fonts are used here instead of images. One reason for that is that the duration of any note is also a symbol. If notes are played in real time, the notes and their associated duration symbol should be animated simultaneously.

<table>
<thead>
<tr>
<th>Rabindra Sangeet name (symbol)</th>
<th>Solfa</th>
<th>Scale of C</th>
<th>Ratio to Sa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shadja (Sa)</td>
<td>Doh</td>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>Komal Rishabh (Re)</td>
<td></td>
<td></td>
<td>256/243</td>
</tr>
<tr>
<td>Shudha Rishabh (Re)</td>
<td>Re</td>
<td>D</td>
<td>9/8</td>
</tr>
<tr>
<td>Komal Gandhãr(Ga)</td>
<td></td>
<td>D#</td>
<td>32/27</td>
</tr>
<tr>
<td>Shuddha Gandhãr(Ga)</td>
<td>Mi</td>
<td>E</td>
<td>5/4</td>
</tr>
<tr>
<td>Shuddha Madhyam(Ma)</td>
<td>Fa</td>
<td>F</td>
<td>4/3</td>
</tr>
<tr>
<td>Kari Madhyam(Ma^)</td>
<td></td>
<td>F#</td>
<td>45/32</td>
</tr>
<tr>
<td>Pancham(Pa)</td>
<td>Sol</td>
<td>G</td>
<td>3/2</td>
</tr>
<tr>
<td>Komal Dhaivat(Dha)</td>
<td></td>
<td>G#</td>
<td>128/81</td>
</tr>
<tr>
<td>Shuddha Dhaivat(Dha)</td>
<td>La</td>
<td>A</td>
<td>5/3</td>
</tr>
<tr>
<td>Komal Nishãd(Ni)</td>
<td></td>
<td>A#</td>
<td>16/9</td>
</tr>
<tr>
<td>Shuddha Nishãd(Ni)</td>
<td>Ti</td>
<td>B</td>
<td>15/8</td>
</tr>
<tr>
<td>Shadja(S’a)</td>
<td>Doh</td>
<td>C’</td>
<td>2</td>
</tr>
</tbody>
</table>

There are certain symbols that have special meaning in Rabindra Sangeet like meed and touching notes.

![Meed symbol connecting two swaras](image)
"Meed" shown in Figure 2 has a special meaning in the audio point of view. It connects two or more notes and the melody of the song slides from one to another. This means that the melody starts from a frequency and slides through intermediate frequencies and stops at the frequency of the destination note. The more the intermediate notes the more melodious is the song.

Touching note is also a new concept. These notes can be associated before or after the main note, as shown in Figure 3 and 4. In Rabindra Sangeet, every note is given a duration. The touching note plays for a smaller fraction of the whole duration and the associated main note is played for the rest of the duration. The touching notes are similar to the main swara-s or notes but are represented in smaller glyphs.

The difference in the science of the audio and the structure of musicsheet (unlike staff notation) between Indian and Western style paved the path to construct a new and simpler framework.

New features in this framework
This framework also has some new features. These features make this format useful for other genres of Indian classical music and languages. It is small in size and robust to work with. The names of the elements in the XML file are given Indian music keywords. Therefore, it is easier for a developer or a newbie in XML programming to code and manage.

As Rabindra Sangeet is almost similar to other genres of Indian music, the scores are encoded with fonts and lyrics in Unicode. Hence, any music score and lyrics can be written with other Indian regional languages. There may be a requirement to search for a song with a keyword or sort some songs alphabetically. There are some wonderful algorithms that accomplish these tasks easily (Akhtaruzzaman 2008 and Rahman and Sattar 2008).

In this framework, the application developer is set free on the decision of creating an API (‘Application programming interface’ n.d.). Application can be built using raw information or using an API of a programming language that supports MIDI. The developer may create his own programming language library or use the JAVA API (David 2008).

Description of the framework architecture
This new framework is built on the existing standard IEEE P1599 (‘IEEE Standard P1599-2008’, 12 June 2008). Unlike the existing framework (Haus and Longari 2005), this structure contains five layers. A brief description of the layers and their functionality is described beneath.

The layers
This work is not limited to MIDI files generation but also transformation of the scanned images to printed musicsheets. Therefore, a layer of characters that define each symbol of the musicsheet is required. Moreover, the MIDI layer is required to generate the audio and also a mapping layer that maps these characters to the audio layer. This means that the layers are of different format and must be synchronized so that data and metadata can easily communicate with each other. Figure 5 describes the layers, their main purposes, and formats.

Information layer
The information layer keeps all the metadata needed to preserve Tagore songs. It contains a
A new framework to preserve Tagore songs

large amount of data including the name of the section of the song or parjay, the genre of the song or vishay, age of Rabindranath when he composed the song, and so on.

Musicsheet layer
The musicsheet layer tells us about the structure of the format of the music sheet. An example of the Rabindra Sangeet musicsheet is given in Figure 6.

The musicsheet is divided into two sections: the score and the lyrics. For ease of coding, the score line is divided into three lines. Each line contains some particular symbols that can appear for that line only. The first line contains the repetition symbol. It tells us what portion of the song should be repeated. The second line contains the actual notes and taal symbols and also the timing information. The last line is provided for a special symbol called meed stated earlier. These three lines are encoded with .TTF or .FON file format defined specifically for this purpose.

The lyrics part contains only one line. It is encoded with Bengali Unicode. Rabindra Sangeet musicsheet has a special format for which it depends upon the number of taal. The taal is analogous to beats in western music. The number of column of the musicsheet is calculated by Equation 1.

Columns= taal * 4+5  (Equation 1)

As shown in Figure 6, the number of taal is 3. The number of columns in the musicsheet of that song is 17. Of these 17 columns, two are for line-end characters and three are for end-of-taal characters.

Swaraliipi layer
The swaraliipi layer constructs the actual musicsheet. It does not tell us about the structure of the musicsheet but tells us about the material

Figure 5 Layers of the architecture and their corresponding formats

Figure 6 Structure of a single line of the musicsheet
that it contains, that is, notes, taal, meed, timings, and, of course, the lyrics. Actually, music or particularly MIDI files is generated by retrieving the data from this layer and processing it through proper logic in the mapping layer.

Mapping layer
The mapping layer implies a layer that maps the characters of the swaralipi layer to some distinct English letters. These letters are the standard notation in Western music. Such as C is the standard notation for generating sound of note number 60. Every character has a definition. The library of any language is built using these standard English letters. The program code that uses this library when executed generates sound of the respective frequency. It is in the hands of the software developer to define the mapping layer. The user also has the independence to map the characters of the swaralipi layer to letters other than the western standard notation. The required information for creating the library is given in the audio layer.

Audio layer
The audio layer consists of data that are used to create library in various languages that support MIDI. The note number, frequency, and pitch; all the necessary information is given here. Suppose C is the note to be played. Therefore, we have to define its properties. The note number given is 60 and the frequency given is 261.

The audio layer also includes the name of the MIDI file and the full path of the file where it is stored. Basically, the main purpose of the layer is to give basic information on MIDI and let the users create their own API [in this work java API (David 2008) is used] of different programming languages.

Description of an application based on this framework (Rabindra Sangeet Notepad)
Rabindra Sangeet Notepad, a music editor is built on the proposed framework to show the layers of the framework and their interdependencies. The primary goal of this application is to generate the MIDI and XML files of all Tagore songs for future use. The interface of the application is shown in the Figures 7 and 8.

The interface is built in Bengali Unicode so that anyone who does not know English can work with it. The arrows in the screenshot indicate various sections of the interface to describe their functionality. The user starts with entering the name of the song in section 3. The name of the song is written in Bengali unicode. This name is searched through an XML file. This XML file contains the names of all the saved songs. If a match is found, then the corresponding information of that song is fetched. The catalogue information of that song can be seen in the corresponding text boxes of section 4. If no match is found, then the song is considered to be a new one and catalogue information can be entered in section 4.

The user can see all the saved songs in the white list box of section 5 and view the information of that song accordingly by clicking on any desired song name.

- Tool Bar
- Swaralipi Editor buttons
- Song name input section
- Catalogue information Section
- List of saved song

The musicsheet of a song can be created by selecting the number of taal and clicking on the button in section 6 of Figure 8. Section 7 contains a musicsheet table, which is divided into the number of columns according to Equation 1. The characters of the scores and lyrics are then put into table cells one by one. The user can create or edit the musicsheet either by typing through the keyboard or by clicking the buttons in section 2 of Figure 7. The whole musicsheet can be saved and printed by clicking the buttons in section 1.

Section 1 also contains some more buttons for playing the current song in MIDI and saving it.
in a suitable location. The MIDI player has play, pause, and stop features. The entire process talks about how MIDI sound is generated from merely a series of characters and how are the layers involved. The procedure is broken into three phases, each of which is accomplished in different encoding formats. The first part is done using fonts and Unicode. The second part is done using java-enabled library and the third part using MIDI. The process starts with the preparation of the musicsheet followed by synthesis of the musical data and symbols, and ultimately generation of the sound is elaborated.

**The preparation of the musicsheet**

Scores and lyrics encoding scheme are related to two layers, the musicsheet and the swaralipi layer. The musicsheet layer defines the structure of the musicsheet and the swaralipi layer defines the scores and lyrics of the song. Scores and lyrics are the series of characters that are scanned during the synthesis of the musical data section. These symbols have their definition in the audio layer. The symbols are scanned one by one and mapped through the user-defined library of the mapping layer to generate MIDI.

The number of taal in a song defines the structure of the musicsheet, that is, it defines the number of columns of the musicsheet of that particular song. As an example, the taal number three looks like Figure 9.

6 Taal Number

7. Musicsheet Section

Score symbols are put into the first three lines and lyrics in the last line, that is, in the fourth line. The application is developed in such a way as to render the upper three lines in TTF format and the last line in Unicode format.

Each cell of the musicsheet contains characters that denote notes, timing information, and other data. Actually, when stored in the XML file, \begin{figure}
\centering
\includegraphics[width=\textwidth]{figure7}
\caption{Swaralipi Editor}
\end{figure}
is divided into three parts as shown in the figure: \(\text{\textBF} \), \(\text{\textBF} \), and \(\text{\textBF} \). \(\text{\textBF} \) is the notehead and is stored as letter \textit{s}, \(\text{\textBF} \) is the timing information stored as letter \textit{a}, and \(\text{\textBF} \) is octave information and is stored as digit 6. The reason for this is
Each cell has a row and column number. Therefore, the characters are stored in XML with their associated row and column number. This is shown in Figure 6. The starting time of each cell is also given to get the real time value. It is required to animate the cell on the real time.

**The synthesis of the musical data**

In the first step, the data were stored in the XML. Subsequently, those data were synthesized using the user-defined library. The library uses the characters that are the outcome of the mapping layer. As an example, is mapped to C, w, and 6. C is for note head and w is for the whole duration. Now the audio layer describes the C character as putting the note ON, giving the time duration of 1 second and frequency of 261.

Now we got a string or specifically a series of intermediate characters. This string is put into the next section.
The generation of the sound
The sound is generated using the MIDI format. The java library accepts the string and creates the MIDI file. The MIDI file contains information like volume, pitch, and instrument of the song. The XML document keeps the information like the name of the file, the instrument is being used, the paths of the files, and so on and so forth. The total time is calculated and stored so that real-time animation is possible.

Possible applications on Indian music
This technology can be used for various applications, such as the one listed below.
- A website, displaying musical scores of a singer.
- An embedded media player that plays the MIDI file of the song and an animated musicsheet that is synchronized with the audio.
- An embedded media player that plays the MIDI file of the song and an animated musicsheet that is synchronized with the audio.
- A Raaga editor that displays graphs of various Indian Raaga-s and evaluates the emotion of the singer.
- Editor for the beginners of Indian classical music where Swara-s and Raaga-s are made available using click, listen, and learn approach, for example, see (Baggi, Baratè, Haus, et al. 2005).
- Editor for composing own song with various instruments.

Conclusion
In this paper, a framework and an application for archiving Tagore songs have been illustrated. This music tool is built to generate all the information related to Rabindra Sangeet: symbolic musicsheet, audio, and structure and character notations.

The existing standards are competent enough for archiving Rabindra Sangeet. But the structures defined there are too complex, while Rabindra Sangeet is archived in a simpler way.

The key feature of this framework is that all the complex information of music piece is bundled into a single file with the power to create printable musicsheet and audible MIDI files. As previously stated, the grammar of Rabindra Sangeet is closely related to that of north Indian classical music; the methods and interfaces described so far can be applied to all the genres of Indian classical music. Of course, to realize other genres of the Indian music, the features of the framework may be changed a little or more, but the core concept of unique framework of heterogeneous data will be preserved.

Acknowledgements
The authors are grateful to the Society for Natural Language Technology Research for funding this research work. They are also obliged to Dr Mandar Mitra, Assistant Professor, Computer Vision and Pattern Recognition Unit, ISI Kolkata for his valuable input.

References


**Bibliography**

MIDI Manufacturers Association
**Learn About MIDI**

Rabindra Sangeet

Recordare: Internet Music Publishing and Software
**MusicXML Definition**

Unicode Consortium
**What is Unicode?**

W3Schools Online Web Tutorials
**XML Tutorial**
http://www.w3schools.com/xml/default.asp (accessed on 12 October 2009)

Wikipedia, the free encyclopedia
**Application programming interface**

Wikipedia, the free encyclopedia
**Hindustani Classical Music**

World Digital Libraries 3(1): 63–72
Abstract
The Digital Curation Centre Curation Lifecycle Model outlived the development of a systematic approach for preservation of images of biological silica collected from the southern part of the United States. This paper opted for an experimental approach of collaborative open-access institutional repository (OA IR) digital collection building and digital preservation strategy for Diatomscapes. Using a small representative sample of images of biological silica as a small corpus of digital data to develop, faculty and digital librarian began to collaborate to develop ways to describe, publish, promote, and preserve Diatomscapes for current and future use. This paper shows how publishing, promoting, and preserving faculty research influences faculty support of and contribution participation in OA IR development and digital preservation exploration implementation. Diatomscapes digital collections in Picasa, Flickr, and DigiTool and digital preservation via Florida Digital Archive (FDA) and MetaArchive were a few outcomes. This paper provides empirical insights on how faculty views, practise or not practise digital curation, and preservation of primary research digital content, for current and future use.

Keywords
Institutional repository, Diatomscapes, Curation lifecycle model
Introduction

The library is a crucial partner in planning and envisioning the future of preserving, using, even creating scholarly resources. So are the technology professionals.

—Johanna Drucker

The purpose of the International Digital Curation Conference is to provide an important platform to help raise the awareness of key digital curation issues, discuss digital curation concepts, principles, and best practices, encourage active participation and feedback from all stakeholder communities, and foster a collaborative environment for the advancement of research and discovery in the lifecycle management of digital content whether ‘born digital’ or ‘digitized’ (Digital Library Forum 2005). The Digital Curation Centre (DCC) Curation Lifecycle Model articulates the lifecycle management of digital content from a high-level purview. The DCC Curation Lifecycle Model was developed by the DCC and introduced to the digital curation community at the Third International Digital Curation Conference (IDCC) in Washington, DC. However, the current DCC Curation Lifecycle Model is a culmination of active participation in digital curation research, development, and advancement from the previous international digital curation conferences.

After reading the article, Size counts for little in the Amazing Lives of Diatoms: small worlds, and further guidance from FSU Biological Scientist, Dr A K S K Prasad, a road to collaboration and partnership led to conference presentations with an international audience, library exhibits, and influenced the introduction, adoption, and progressive development of a digital preservation strategy of faculty digital content for FSU Libraries.

Previous FSU Digital Library Center Diatomscapes digitization image enhancement and poster images print work performed with some of Dr Prasad’s diatom images laid the groundwork to extend the reach and richness of his biological science research via collection building, conference presentations, exhibits, and digital preservation. Even though this project is linear and myopic in scope, specifically to the biological science research discipline, its goal is to use this project to develop a collaborative institutional repository (IR)/digital curation model for cross-discipline dissemination.

‘Unless we scholars are involved in designing the working environments of our digital future, we will find ourselves in a future that does not work, without the methods and materials essential to our undertakings’ (Drucker 2009).

Development of digital curation and the International Digital Curation Conference

‘Digital curation is maintaining and adding value to a trusted body of digital information over the lifecycle of scholarly and scientific materials, for current and future use’ (DCC 2005). Digitized content is content in which electronic surrogates of the physical objects, images or analog sound have been created through the process of digitization such as scanning, optical character recognition (OCR) or digital imaging to name a few. ‘Born digital’ content is content in which the original source is created in the electronic format and there is no physical or analog equivalent of the electronic content (DPC 2002). Digitized and born digital content are digital information.

Diatomscapes: images of biological silica

Diatomscapes represents a collection of images of biological silica, and includes diatoms ['microscopic, single-celled plants that thrive in freshwater, saltwater, brackish water, and even semi-terrestrial environments’ (Prasad 2005)] and Radiolarians ['amoeboid protozoa that produce intricate mineral skeletons’ (Dictionary 2008)]. Diatomscapes II is another collection...
of images of biological silica. FSU biological scientist Dr A K S K Prasad developed the term Diatomscapes to describe his images of biological silica faculty research. The following passage taken from the Diatomscapes Images of Biological Silica document briefly introduces the definitions of silica and diatoms through the lens of a faculty scientist.

“Silica, the most abundant material in the earth’s crust, is often referred to amorphous, hydrated, and polymerized silicic acid with unknown molecular weight. Considering the abundance of silica in nature, it is remarkable that only a few groups of organisms have utilized this element in their systems. Siliceous structures occur in a variety of organisms but predominate in the more “primitive” forms of life, in which amorphous, hydrated silica forms cell walls, tests, scales, and other skeletal features. Of the 14 divisions of the algae, diatoms are one of the few groups that make any notable use of silicon.

Diatoms are a Johnny-come-lately class of microscopic, single-celled plants that thrive in freshwater, saltwater, brackish water, and even semi-terrestrial environment, having evolved at about the same time as flowering plants, birds and mammals at the end of the Jurassic period. Diatoms, like birds and mammals, have characteristic mineralized skeletons, generally constituting a considerable proportion of their weight. By contrast to our own internal skeletons, which are phosphotic, those of diatoms are siliceous (referred as “frustules”), being initially composed of amorphous oxalate’ (Prasad 2005).

The rise of diatomscapes: promotion

The faculty ‘Explore the World of Diatomscapes’ exhibition was developed and was on display from 3–17 November 2008 at the FSU Robert Strozier, Library Exhibition Room. The exhibits included 38 large format, laminated posters of images of biological silica from Dr Prasad’s Diatomscapes collection, a short Diatomscapes video, Diatomscapes metadata (brief and not Access to Biological Collection Data [ABCD] compliant), and information on electron scanning microscopes used to digitize Diatomscapes I and II images of biological silica (see Figures 1 and 2).

Diatomscapes digital collections: publication

Until April 2009, the Diatomscapes collections existed online with restricted access, due to pending publication of some images of biological silica. Also, the faculty did not want unpublished images of biological silica from Diatomscapes research made online with public access. However, when asked if permission could be granted to build an online digital collection with restricted, non-open-access access, the faculty agreed with the conditions that the

Figure 1 Explore the World of Diatomscapes faculty exhibition: a poster on display

Figure 2 Explore the World of Diatomscapes faculty exhibition: hall view
images will only be released if the faculty or the digital librarian is contacted directly; in either case, the faculty must grant permission. The faculty allowed Diatomscapes digital collections developed in both Picasa and Flickr (Figures 3 and 4) under the provison that access to images be restricted. The faculty was pleasantly surprised to see some of his Diatomscapes images published online, but was even more ecstatic when his sponsor requested permission to view online Diatomscapes. An unexpected outcome was increased collaboration and communication of developing digital future between courtesy scholars and faculty sponsors. However, after considering the restricted published images of biological silica from Diatomscapes as open-access candidates, the

FSU Digital Library received permission from the faculty to include select images of biological silica from Diatomscapes I (Appendix), while excluding images of biological silica seeking pending publication. The permission for Diatomscapes I ‘open-access’ digital collections was granted for select images of biological silica from Diatomscapes I in FSU Libraries Digital Collections (Figure 6) after a year of developing trust, gaining respect, building partnerships, corresponding via email, and face-to-face meetings; and delivering benefits and desirable outcomes, also known as return on investments.

The dialogue between the faculty and the digital librarian or archivist (data curator), according to Level Two Data Curation Model (Lord and Macdonald 2003), is essential in proper description and metadata information for primary research digital objects, which fortifies the digital object’s verifiability, establishes IR credibility, and maintains professional research integrity. These are imperative for faculty contribution participation in IR/digital preservation collaboration partnerships. It is necessary to find out the needs of the faculty in their digital objects IR/digital preservation development and work with scholars to

Figure 3 Diatomscapes online, unlisted
Source Picasa

Figure 4 Diatomscapes online
Source flickr

World Digital Libraries 3(1): 73–84
information: to what extent maintenance of the integrity of content and behaviour is important; and maintaining security, confidentiality, authenticity, access controls and audit trails of use and change’ (The Digital Archiving Consultancy 2006).

**The Digital Curation Centre Curation Lifecycle Model**

The DCC Curation Lifecycle Model provides a graphical high-level overview of the stages required for the successful curation and preservation of data from initial conceptualization or receipt. The model plans activities within an organization or consortium to ensure that all necessary stages follow correct sequencing. The model enables crosswalk granular functionality to define roles and responsibilities and build a framework of standards and technologies of implementation. It can help with the process of identifying additional steps that may be required or actions that are not required by certain situations or disciplines, and ensuring that processes and policies are adequately documented (DCC 2008).

**Mapping the Digital Curation Centre Curation Lifecycle Model to diatomscapes: preservation**

**Data (digital objects or databases)**

**Digital objects**

Simple digital objects are discrete digital items such as textual files or sound files, along with their related identifiers and metadata. *Diatomscapes*. Thirty eight images of biological donated by Dr Prasad were used as part of this project. The images of biological silica represent diatoms and Radiolarian gathered from 2004 to present mainly from the south-eastern most part of the United States. The images were in tagged image file format (TIFF). However, the stipulation was that the images are not to be...
published online until some species names new to science have been published via traditional publishing venues. Permission to share images is only based on personal requests (limited/ restricted access). The researcher’s view of enacting limited access and the institution’s view of what data to preserve are not new issues as the following excerpt from the Research Information Network (RIN) suggests.

‘Researchers, funding agencies, and institutions need to take full account of the different kinds of data that researchers create and collect in the course of their research, and of the significant variations in researchers’ attitudes, behaviour, and needs, and to make clear the categories of data that they wish to preserve and shared with others in each case’ (Griffiths 2008).

Databases
Complex digital objects are discrete objects, made by combining a number of other digital objects.

Full lifecycle actions

Description and representation information
Assign administrative, descriptive, technical, and structural and preservation metadata using appropriate standards to ensure adequate description and control over the long term.

Diatomscapes: Diatomscapes lacked complete metadata and only consisted of basic metadata elements that could be easily mapped to Dublin Core for simple metadata record to be used in an institutional repository such as DigiTool. These basic metadata elements included scientific name, collection site, date of collection, ecological preference, image id (or unique identifier), and microscopic magnification. The paper proposes gathering more technical metadata for the images of biological silica and using ABCD as a metadata content standard and Metadata Encoding Transmission Standard (METS) as a digital content standard. Currently, there are no appropriate standards or adequate description of the images of biological silica and thus ‘…newly created data sets continue to enter the same trajectory of degradation and loss that has been and is now experienced by legacy data set’ (Kintigh 2006, p. 572).

Preservation planning
Plan for preservation throughout the curation lifecycle of digital material. This would include plans for management and administration of all curation lifecycle actions.

Diatomscapes: Diatomscapes was used as a Florida State University demo test collection for preservation using Dark Archive in the Sunshine State (DAITSS) open-source software as part of the Florida Digital Archive (FDA) at Florida Center of Library Automation (FCLA). The preservation planning of Diatomscapes included creation of METS descriptor file packet that included technical metadata and MD5 check for each digital object within a METS package, online preservation statistics reporting, and ingest report. At the time of the project, FSU Libraries did not have an account established with FDA for the preservation of FSU digital content. The preservation planning of Diatomscapes via FDA includes the management, preservation, and online reporting statistics such as file size, MD5/Sha1, # number of files, and file names of preserved content. ‘In the future, a scholar or researcher will want to know that a digital object is trusted – that is authentic and reliable’ (Jantz 2008).

Community watch and participation
Maintain a watch on appropriate community activities, and participate in the development of share standards, tools and suitable software.

Diatomscapes: The researcher has been introduced to Marine Metadata Interoperability (MMI) and ABCD because of this project. The ABCD is a content standard in use with the Global Biodiversity Information Facility
Diatomscapes exposé: how faculty and digital librarians collaborate

(GBIF) and Biological Collection Access Service for Europe (BioCASE). It is maintained by the Biodiversity Information Standards, formerly known as the Taxonomic Database Working Group (TDWG), which ‘focuses on the development of standards for the exchange of biological/biodiversity data’ (TDWG 2007). ABCD has been identified and recognized as a content standard that should be adopted and implemented for Diatomscapes and the remaining images of biological silica.

Create or receive
Create data including administrative, descriptive, structural, and technical metadata. Preservation metadata may also be added at the time of creation. Receive data, in accordance with documented policies, from data creators, other archives, other repositories or data centres, and if required assign appropriate metadata.

Diatomscapes: Some of the technical metadata created for Diatomscapes was generated by a MD5 checksum, JSTOR/Harvard Object Validation Environment (JHOVE), and curation tools software applications.

Curate and preserve
Be aware of and undertake management and administration actions planned to promote curation and preservation throughout the curation lifecycle.

Diatomscapes: After consultation with the FDA Director, setup procedure instructions were sent as to how to ingest, curate, and preserve Diatomscapes via FDA. DAITSS METS SIP Profile was created, FDA SIP Specification followed, and Ingest Report generated as part of submitting materials to FDA for preservation. Real-time statistics are available online.

Sequential actions

Conceptualize
Conceive and plan the creation of data, including capturing of methods and storage options.

Diatomscapes: Diatomscapes were created by JSM-840 and FEI Nova Nano–400 electron scanning microscopes in the FSU Biological Science laboratory. Even though basic metadata has been captured, the need for a systematic approach to metadata creation and digital preservation was recommended to the researcher along with recommendation of using The DCC Curation Lifecycle Model as a framework reference tool to explore digital curation of images of biological silica.

Appraise and select
Evaluate data and select for long-term curation and preservation. Adhere to documented guidance, policies or legal requirements.

Diatomscapes: Diatomscapes images were selected and appraised by FSU biological scientists. Diatomscapes II images were selected on the basis of the fact that the images were created with FEI Nova Nano–400 electron-scanning microscope that replaced and provided higher microscopic magnification than the legacy JSM–840 electron-scanning microscope. Images were selected on the basis of researchers’ attitudes and experiences involved in the discovery and naming of some species.

Ingest
Transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements.

Diatomscapes: FTP to the FCLA FDA server according to DAITSS METS SIP Profile and FDA SIP Specification as outlined in FDA Policy and Submitting Materials to the FDA (FDA 2003).
Preservation action
Undertake actions to ensure long-term preservation and retention of authoritative nature of data. Preservation actions should ensure that the data remains authentic, reliable, and usable while maintaining its integrity. Actions include data cleaning, validation, assigning preservation metadata, assigning representation information, and ensuring acceptable data structures or file formats.

Diatomscapes: MD5 checksum, JHOVE, and other curation tools are used to create technical metadata that is then included in the METS SIP Profile that describes the images of biological silica that are Diatomscapes.

Store
Store the data in a secure manner adhering to relevant standards.

Diatomscapes: The images of biological silica of Diatomscapes are stored in a dark archive (non-web accessible) with technical and preservation metadata included in the DAITSS METS SIP that accompanied the digital objects on ingest.

Access, use, and reuse
Ensure that data is accessible to designated users, on a day-to-day basis. This may be in the form of publicly-available published information. Robust access controls and authentication procedures may be applicable.

Diatomscapes: Diatomscapes are in Flickr and Picasa with restrictive access (non-publicly viewable/accessible) until species names new to science are published. Some images of biological silica are also in Morphbank.

Transform
Create new data from the original by migration into a different format or create a subset by selection or query to create newly derived results, perhaps for publication (derivatives).

Diatomscapes: Diatomscapes TIFF images were transformed into JPEGs for ingest in Flickr and Picasa for sharing; into large-format posters for an exhibition display to promote faculty research; into a short MPEG video and uploaded to Facebook. Select images of biological silica were chosen to create Diatomscapes and Diatomscapes II. Some images of biological silica have been published in scholarly publications/journals.

Occasional actions
Dispose
Disposal of data that have not been selected for long-term curation and preservation, in accordance with documented policies, guidance or legal requirements. Typically, data may be transferred to another archive, repository, data centre or other custodian. In some instances, data are destroyed. The data’s nature may, for legal reasons, necessitate secure destruction.

Diatomscapes: Disposition of images of biological silica from Diatomscapes has not yet been performed at the time of writing this paper.

Reappraise
Return data that fails validation procedures for further appraisal and selection.

Diatomscapes: Reappraisal of images of biological silica from Diatomscapes are currently in review by scientists. Previous images of biological silica created with legacy scanning-electron microscope may need to be rescanned with the current electron scanning for higher resolution and microscopic magnification.

Migrate
Migrate data to a different format. This may be done to accord with the storage environment or to ensure the data’s immunity from hardware or software obsolescence.
Diatomscapes: Diatomscapes images are currently preserved in the TIFF format. However, with the advancement in preservation standards, new and emerging standards have been introduced, leading to best practices. As a result, current preservation standards and file formats will be assessed bi-annually for compliance, and to reduce the threat of technical obsolescence.

**Closing remarks**

*Curation as a finite process, with handover to preservation at its end point*

Curation as a whole life process, with evolving objects, and

*Curation as managing a growing, living collection.*

—Clifford Lynch (Rushbridge 2008)

The DCC Curation Lifecycle Model, though not exhaustive and still evolving, is an excellent model for developing processes toward the development of a digital curation profile for the FSU Biological Science Department, particularly for images of biological silica. It can be used as a model across research disciplines to develop digital curation profiles, policies, and procedures that address some of the challenges of preserving digital content for current and future use.

Even though the Diatomscapes Exposé project is neither sustainable nor optimal in developing an OA IR collection-building/digital curation model due to paucity of time of the faculty and the digital librarian, the efforts to produce the faculty exhibition and presentations are an investment in the future, digital future. The primary goal of the faculty exhibition and conference presentations was to promote, publish, and preserve FSU faculty research, thus stimulating future discussions on streamlining IR development/digital curation research collaborations and partnerships and making them more efficient. The Diatomscapes presentations serve as a progressive marketing model for promotion across multiple FSU research disciplines to stimulate faculty collaboration and partnerships. The responses to Diatomscapes from the United States to the United Kingdom have been very favourable and encouraging including increased collaboration between FSU biological scientist courtesy scholar and sponsor, FSU biological scientists, and colleagues. Even if only intensive discussion of working towards developing a more scholarly research/resource-sharing environment is the sole outcome, at least the scholars are in the design process of digital future and not retrofitted like format migration from technical obsolescence.

Developing sustainable digital future includes constant investment of time, energy, knowledge, skills, and abilities from primary (scholars), secondary (libraries), and tertiary (research and scholarly communities) stakeholders. Though this paper focuses on the primary and secondary stakeholders, it is hoped that there will occur greater involvement with the tertiary stakeholders via faculty research roadshows similar to the ‘Explore the World of Diatomscapes’ library exhibition, which drew interest, comments, and favourable appreciation from the FSU faculty, students (undergraduate and graduate), and the research and learning communities.

The following excerpt from Sustaining the Digital Investment by the Blue Ribbon Task Force best summarizes the guiding principles behind the investment of time, enthusiasm, energy, and effort by faculty and digital librarian for the faculty exhibition and conference presentations.

‘Investment is about the future, and is almost always an uncertain proposition. The value of an investment will ebb and flow with the vagaries of shifting circumstances, priorities, and attention. Yet, without investment, society cannot grow: in business, research, and culture, ongoing development and growth require that a portion of current resources be set aside for maintaining, and ideally, expanding, our future productive and creative capacities. Despite the uncertainty, and despite the cost, we must invest now to
secure the opportunity for a future that builds on and surpasses the achievements of the present’ (Sustaining the Digital Investment 2008, pg. 7, para. 1).

Actions springs out of what we fundamentally desire... and the best piece of advice which can be given to would-be persuaders, whether in business, in the home, in the school, in politics, is: First, arouse in the other person an eager want. —Harry A Overstreet from Influencing Human Behaviour

Future research
Monitor the digital curation projects and activities in the United Kingdom, United States, and other countries. Research the digital curation profiles work at Purdue to develop digital curation profiles for multiple research disciplines at FSU starting with biological science and department of oceanography.

‘A Data Curation Profile is a resource for library and information science professionals, archivists, IT professionals, data managers, and others, who want information about the specific data generated and used in research areas and sub-disciplines that may be published, shared, and preserved for reuse. Data Curation Profiles capture requirements for specific data generated by a single scientist or scholar, based on their needs and preferences for these data. Profiles are derived from several kinds of information, including interview and document data, disciplinary materials, and standards documentation. Data Curation Profiles are also intended to enable librarians and others to make informed decisions in working with data of this form, from this research area or sub-discipline’ (Purdue 2009).

Review continually growing literature in digital curation, build on the partnership and trust within the biological science research discipline to develop similar partnerships across other research disciplines, further explore the context and definitions of digital curation across research disciplines and countries, explore digital curation tools and case studies, review data asset framework (DAF) for future implementation, and introduce the Diatomscapes research and digital collection to research, teaching, and learning for students via mapping to Boyer’s Model of scholarship for course content consideration.

Conclusion
Faculty is willing and interested in contributing to OA IR and preservation of some of their research data when the stewardship, resource management, publishing, and preservation of their research data are professional and beneficial to the researcher. Two comments from FSU faculty regarding their online collections are as follows.

‘Thank you so much for all of work on this. It is a great service to the crustacean community to have these available. Please let me know if I can help with anything on the project.’
—Dr Lawrence G Abele, FSU Provost and Professor, Biological Science

‘I am pleased to have our reports on the FSU digital library collection. It is a worthwhile service for our group, as it is a suitable location for our technically-oriented reports. While we also maintain a collection of our reports on our own website, the extra visibility of a central collection should make it easier for potential readers to find it on the web.’
—Dr Peter Lazarevich, Associate In, FSU Oceanography

‘Data curation + digital preservation = digital curation’ (McGovern 2009) and ‘curation is clearly domain-dependent’ (Rushbridge 2008). Researcher (data producer) and digital librarian/digital curation specialist (data curator) must continue to work together to develop work flows and frameworks to manage research data effectively and address the challenges and complexities of digital curation.

Further research and work must be performed to effectively assess, gather, select, and curate...
Diatomscapes exposé: how faculty and digital librarians collaborate

research data for stewardship for current use and preservation for future re-use across research disciplines at FSU. The open archival information system (OAIS) reference model, and DAF are good building blocks for future digital curation work.

References


World Digital Libraries 3(1): 73–84


Nibert M (undated). **Boyer’s model of scholarship.** Details available at <http://www.webs1.uidaho.edu/mkyte/ui_strategic_plan_implementation/resources/BoyermodulePacificCrestrecd209.4.06.pdf>, last accessed on 28 October 2009


Prasad A K S K. 2005. **Diatomscapes images of biological silica.** Last accessed on 12 April 2008 from Dr Prasad via email correspondence


**Book review**

**Digital Information: order or anarchy?**

Hazel Woodward and Lorraine Estelle (Editors)

Reviewed by Shantanu Ganguly*

* Fellow, Library and Information Centre, TERI

---

* Digital Information: order or anarchy?* is a classic publication in today’s age of digital information, which each and every digital librarian and other library and information professionals, who are planning to develop a digital library, should go through, so that the challenges they come across in the development of the library can be answered to some extent. The chapters in the book also discuss future developments in our digital universe and explore their potential impact on libraries, publishing companies, and all the other players in the marketplace.

The chapters by Anderson, Steele, and Russel essentially focus on the process and evolution of scholarly communication in the digital age from their own individual perspectives. Anderson identifies three ‘crunch points’ for libraries and publishers: searching for information, library collections, and pricing. Anderson works through various future scenarios, around these topics, and points to the essential mission change for academic and research libraries.

Collin Steele focuses on e-books and the future of scholarly communication. Technology is creating many opportunities for book publishers in terms of dissemination; witness the development of new e-book services, e-book readers, print-on-demand publications, and Espresso Book Machines.

Technological advances mean that non-traditional players can now create and disseminate digital information. Libraries are now publishers, and companies such as Google and Microsoft are undertaking mass digitization projects. Alastair Dunning’s chapter, ‘Digitizing the Past’, describes how public-sector funding has resulted in the digitization of some collections.

Issues around intellectual property rights challenge not only the traditional publishers, but also the new players. In their chapter, Wilma Mossink and Lorraine Estelle describe the problems, and some of the work that is currently being undertaken to address them. The power of the Internet means that it has never been easier to create and access content and, perhaps more importantly, to reuse and ‘mash-up’ a number of content sources.

The users of digital content should be able to navigate through a well-ordered information environment to find the information they seek, without unreasonable barriers. However, as Graham Stone points out in his chapter on resource discovery, the landscape is messy and
libraries face enormous challenges in the digital marketplace. The users are persuaded not to start the search for information by using library resources.

The chapters in this book focus mainly on the academic perspective. However, the issues tackled are pertinent to the broader information industry, including trade publishers, public libraries, and national libraries. The editors’ opening chapter provides an overview of the digital revolution and its impact on all these sectors. We live in a period of radical change, which provides for many opportunities but also threatens the status quo. According to Charles Darwin: 'It is not the strongest species that survive, nor the most intelligent, but the ones most responsive to change.'
News

Third International Conference on Digital Libraries

(ICDL) 2010
Shaping the information paradigm
23–26 February 2010

Background
The Information and Communication Technology (ICT) revolution and the advent of the Internet have had significant and far-reaching impacts on the knowledge and information sector, and have added a new dimension to information retrieval platforms. A fierce competition, in terms of global information access, has ensued amongst modern libraries, publishers, database providers, and the Internet service providers. This has resulted in the creation of digital libraries (DLs), knowledge management (KM), and e-learning portals. Also, archiving of indigenous culture and heritage has been conceived and implemented in various parts of the country. DLs, today, adopt the latest ICT tools and techniques to bring together nationwide DLs from across the country on a single platform. They have become a critical component of the global information infrastructure, as they facilitate rapid flow of information, thereby bridging the digital divide, and in the process, providing uninterrupted organized information access in multiple languages to users in all sectors.

About the conference
The Third International Conference on Digital Libraries (ICDL) 2010 was held during 23–26 February 2010 at India Habitat Centre, Lodhi Road, New Delhi. The theme of ICDL 2010 was “Shaping the information paradigm”. The conference aimed at addressing some important contemporary issues. Some of them were as follows.

Radical change from print resources to digital resources, and subsequently to various other formats suitable for video, text, multimedia, and so on.

A suitable integration of the academic environment with digitized resources to meet the needs of the faculty fraternity.

A strategic alliance for the development of DLs and e-learning across the world, focusing on the creation, adoption, implementation, and utilization of DL and open distance and online learning (ODOL) system.
The conference was jointly organized by The Energy and Resources Institute (TERI) and Indira Gandhi National Open University (IGNOU), and inaugurated by Shri Kapil Sibal, Minister of Human Resource Development, Government of India, on 23 February 2010. In the inaugural address, the minister highlighted the following points with regard to the digitization programme in India.

- Users’ choice between print vis-à-vis digitized resources.
- The need for flexibility and interactivity in the digitized system.
- Socialization and interaction between libraries and their users.
- Physical libraries can be used as physical spaces for reading, using computer facilities, and socializing.
- New role for librarians, who disseminate information, thus aiding in the development of information literacy among users.

Smt D Purandeswari Devi, Minister of State for Human Resource Development, was the Guest of Honour. Dr R K Pachauri, Director-General, TERI, welcomed the participants and introduced ICDL 2010, which was followed by an address by Dr Deanna Marcum of the Library of Congress, USA. Prof V N Rajasekharan Pillai, Vice Chancellor of IGNOU, highlighted the role of his institution in promoting e-learning and developing DLs, while Prof. N Balakrishnan, Professor and Associate Director, Indian Institute of Sciences (IISC), Bangalore, also addressed the conference. The vote of thanks was proposed by Dr Debal C Kar, Fellow, Library and Information Centre, TERI, and Organizing Secretary of the conference.

The tutorial session was inaugurated by Dr T Ramasamy, Secretary, Ministry of Science and Technology, Government of India, at the IGNOU Convention Center at 9.00 a.m. on 23 February 2010. Ms Anne Caputo, President, Special Libraries Association (SLA), USA, and Dr Vibha Dhawan, Executive Director, TERI, were also present on the occasion. The vote of thanks was proposed by Mr Sudhir K Arora, University Librarian, Library and Information Centre, IGNOU. A total of nine tutorial sessions were conducted on 23 February 2010.

The conference had 40 invited speakers from 17 countries, including India, and 665 registered participants, of which more than 100 participants were from 35 countries of Asia, Africa, Europe, America, and Australia. Some of the leading DL luminaries, who attended the conference were Michael Sadle, P Anandan, R Chidambaram, A Parsuramen, Anne Caputo, Joyce Chao-chen, Jean Marc Comment, Denise Troll Covey, Hiranmay Ghosh, Stefan Gradman, Minna Karvonen, Akira Maeda, Gill Needham, Edie Rassmusen, Anna Maria Di Sciullo, Ming Zhang, K R Srivathsan, Shigeo Sugimoto, and many more.

Three keynote addresses were delivered on the 24, 25, and 26 February 2010. The first keynote address was delivered by Dr P Anandan, Managing Director, Microsoft India, on 24 February 2010. He emphasized that the present day video is an important source of information and highlighted the characteristics of each type of video.

The second keynote address was given by Dr Jens Thorhauge, Director General, Danish Agency for Libraries and Media, Denmark, on 25 February 2010. He stated that information and knowledge societies develop fast, and as a result, are creating a new form of global...
competition, new digital information sources, and new ways for studying and research, apart from emphasizing on the new need for lifelong learning and media literacy, as also creating digital divides and social exclusion. The third keynote address was given by Dr R Chidambaram, Principal Scientific Adviser, Government of India, on 26 February 2010. He opined that today, huge amount of information is available on the Internet, and if one is not familiar with the topic then one has to be careful about the quality of the supply knowledge, and whether the person who has abstracted the material understands the subject.

A total of 260 papers/abstracts were received. These papers were reviewed by members of the national and international programme committee, and a total of 99 papers were shortlisted for presentation during the conference. A Paper Award Committee was also constituted to confer three best paper awards in three categories, namely, India, Asia, and the world. A total of 30 technical sessions were held on 24, 25, and 26 February 2010. Poster presentations were held on 24–25 February 2010. The entries for the poster presentations were reviewed by members of the National and International Programme Committee, and 120 presentations were finally short-listed. A Poster Award Committee was also constituted to confer three best poster awards. Three volumes comprising tutorial proceedings, two pre-conference proceedings, and a programme guide, both in printed and electronic form, were given to all participants at the time of registration. During the conference, conference bulletins, highlighting the day’s proceedings, were made available on the following day, prior to the commencement of the conference.

The first panel discussion was held on ‘DL security’ on 25 February 2010. Prof. K R Srivathsan stressed upon the security of digital content and its interoperability. Prof. Peter Schirmbacher stressed on the need for using open source virtual learning environment such as Moodle. He further spoke at length on technical issues pertaining to server security and document security. He expressed concern on content authenticity in a digital repository. He also talked about DINI Certificate. Prof Stefan Gradmann pointed out that identity management and user authorization are vital issues in DL service delivery.

The second panel discussion was held on ‘DL policy and standards’ on 26 February 2010. Dr Thorhauge opened the discussion on ‘how to digitize materials and make it available and accessible to users in the best way’. Prof. Michael Seadle talked about the copyright policies for digitization and the need for standards for digital preservation of materials. He further added that the policies and standards we are developing should be implementable. In her discussion, Ms Kalpana Das Gupta mentioned the lack of a holistic approach and the absence of a coordinating body, with regard to the digitization of materials at the national level. India, being a multicultural and multilingual nation, requires different policies and standards catering to different material needs as well as cultural heritage of the country. Prof. Manohar Lal traced the historic evolution of digitization way back to 600 BC, the age of Panini, whom he described as the first digital librarian.

Three morning breakfast sessions were organized before the commencement of the conference on 24, 25, and 26 February 2010. A selected group of experts was invited to brainstorm on the following issues:
- Issues and challenges in DL concept
- Multi-linguality and inter-operability
- Copyright and digital rights management

The conference received tremendous response from across the country and from abroad, with the event attracting partnerships in the form of associates, co-associates, sponsors, and so on. The Ministry of Culture, Commonwealth Education Media Centre for Asia (CEMCA), Council of Scientific and Industrial Research
(CSIR), and Special Libraries Association (SLA) were some of the key associates for the conference. The co-associates were American Center, Austrian Cultural Forum, The British Council, Goethe Institute, and the Royal Embassy of Netherlands. The other institutional sponsors who supported the conference were the Ministry of Communications and Information Technology (MCIT), Defence Research and Development Organisation (DRDO), Indian National Science Academy (INSA), and Raja Rammohun Roy Library Foundation (RRRLF). The conference was also sponsored in different categories by Emerald Group Publishing Ltd, Global Information Systems Technology Pvt. Ltd, Nature Publishing Group, Tausco Subscription Agency Pvt. Ltd, and Taylor & Francis Group. A total of 34 exhibitors, besides TERI and IGNOU, set up stalls during the conference to showcase various digital products and services offered by leading publishers in the world.

Conclusion

The third conference in the ICDL series generated great enthusiasm among the library and information science and technology professionals across the world. It generated awareness on the importance of digitization in the DL community both in India and abroad. The conference, on the whole, facilitated the bridging of knowledge gaps between developing and developed countries; initiated capacity building activities in DLs; provided a forum for facilitating useful interaction amongst information science and technology professionals; and most importantly, facilitated the formulation of recommendations on digitization technologies, Acts, and policies in India. It also helped in bringing together about 650 professionals, experts, digital library workers, and information providers from across the world on a single platform and encouraged the sharing of experiences, concerns, and ideas. The conference also threw light on the present status of the DL and the latest developments and techniques in this field all over the world. To understand the basics of digital preservation tools and techniques, several private and government organizations have already initiated action towards the future preservation of their documents and have approached TERI for inputs and assistance. The Ministry of Culture has started the digitization of the cultural heritage of India.
European parliament backs digital plans
The European Parliament has approved the Europe Union’s Digital Agenda, an action plan designed to boost Europe’s economy through the adoption of digital practices. There was also “strong support” from the European Parliament for the further development of Europeana, Europe’s digital library, with the commission to draft a policy document on its future, including looking at copyright and orphan works.

- The parliament has backed the call for European citizens to have access to modern digital technologies, including high speed Internet, and the essential skills to understand and use them.
- The parliament also stressed that consumers should be able to freely access public services online and content across the European Union, and underlined the need for consumers to know their rights in the digital environment and for a clear legal framework to protect these rights.
- The parliament has now urged the commission to come forward with a proposal for an ambitious digital agenda and action plan, which would enable Europe to progress towards an open and prosperous digital society.
- The parliament also highlighted the need to address a series of copyright-related issues to facilitate the digitization and online accessibility of cultural content, including the issue of orphan works.

Source  http://www.thebookseller.com, last accessed on 11 May 2010

DSpace 1.6 is now available
DuraSpace organization announced the long-awaited release of DSpace 1.6, the popular turnkey open source application for managing and providing access to digital content used to create more than 800 repositories across the world. The community-requested features in the new release include the following.

- Enhanced statistics package, which provides more information on how your repository is being used.
- Embargo facility so that items can be kept dark for a period of time.
- Batch metadata editing tool, which can be used to change, add, find/replace metadata as well as facilitate mass moves, re-order values or add new items in bulk.
- Authority control, which contains integration with the Sherpa Romeo Service for publisher names, as well as the Library of Congress Nameservice

Source  http://Dspace.org, last accessed on 21 April 2010

E-mail preservation
The Arizona State Library, Archives, and Public Records has developed an automated process to preserve official e-mail records produced by Microsoft Outlook. The process also captures metadata needed for the long-term preservation of the information. This work is being implemented as part of the Persistent Digital Archives and Library System project, which is developing a shared curatorial framework for preserving digital public records across multiple states.

Source  http://digitalpreservation.gov, last accessed on 26 April 2010
Forthcoming events

European Conference on Research and Advanced Technology for Digital Libraries - ECDL 2010
6–10 September 2010, Glasgow, United Kingdom
Details available at <http://www.ecdl2010.org/>

International Conference on Dublin Core and Metadata Applications: "Making Metadata Work Harder: celebrating 15 Years of Dublin Core" (DC-2010)
22–24 October 2010, Pittsburgh, Pennsylvania, USA
Details available at <http://dcpapers.dublincore.org/ojs/pubs/announcement/view/4>

International Conference on Knowledge Engineering and Knowledge Management
EKAW 2010
11–15 October 2010, Lisbon, Portugal
Details available at <http://ekaw2010.inesc-id.pt/>

International Conference on Knowledge Discovery and Information Retrieval (KDIR),
25–28 October 2010, Valencia, Spain
Details available at <http://www.kdir.ic3k.org/>

Fifth International Conference for Internet Technology and Secured Transactions (ICITST-2010)
8–11 November 2010, London, UK
Details available at <http://www.icitst.org/>
Eleventh International Conference on Web Information System Engineering (WISE 2010)
12–14 December 2010, Hong Kong, China

International Conference on Digital Libraries and Knowledge Organization
14–16 February 2011, Gurgaon
Details available at <http://www mdi.ac.in/ICDK/Home.html>
Guide to authors

*World Digital Libraries* is an international peer-reviewed biannual journal. The journal seeks quality research papers that present original theoretical approaches. It also seeks experimental case studies related to digital library developments, maintenance, and dissemination of digital information focusing on research and integration of knowledge at the interface of resources and development. The journal will, therefore, keep readers abreast with the current developments and contain articles, reviews, current developments, and case studies, encompassing the following areas.

- Theoretical and methodological issues that relate to the interrelationships among electronic resources management, digital preservation, multiple access, multilinguality, copyright issues, and security aspects.
- Theoretical approaches as well as experimental case studies related to digital library development and maintenance.
- Initiatives towards digitization through lucid case studies.
- Current developments across the globe.
- Dialogues between the scientific community and society at large.

Articles should examine concepts, analyses, and case studies of important issues in the field.

Book reviews should be of recent publications in the field, to be reviewed by an independent reviewer.

Commentaries should discuss critical issues in the field.

**Submissions**

Authors are requested to send a soft copy (in Microsoft Word format) of their contribution to the editor, either in a CD or as an e-mail attachment.

All submissions will be peer-reviewed using the criteria of originality, accuracy, and quality of contribution in these fields.

**Presentation of manuscripts**

Articles must be original, in English, and should not exceed 8000 words. The main text should be double-spaced with headings and subheadings clearly indicated in the text. All tables, figures, and equations should be numbered in Arabic numerals and clearly cited in the text. All measurements should be in metric (SI) units. The manuscript should be arranged in the order given below.

- Short title (10 words is the desired maximum length), subtitle (if desired)
- Author’s name, affiliation, full postal address, and e-mail, telephone, and fax numbers (respective affiliations and addresses for co-authors should be clearly indicated)
- Abstract (not exceeding 200 words)
- Main body of the text, suitably divided under headings
- Acknowledgements, if any
- References
- Appendices (each on a separate sheet)
- Tables (each on a separate sheet)
- Figures (each on a separate sheet)

**Shorter items**

The following shorter items are also welcome and must be typed in the same way as major articles.

- Commentaries (research notes and short communications) and case studies (maximum 5000 words)
- Book reviews (maximum 1200 words)

**In-house style: references**

In the text, the surname of the author(s) followed by the year of publication of the reference should be given, for example, (Hall 1993). In case of several publications by the one author or by a group of author(s) in one year, use notations ‘1993a’, ‘1993b’, and so on. Up to three authors can be mentioned in text references; more than three authors should be limited to the first three authors’ names followed by ‘*et al*’. References must be listed alphabetically at the end of the
paper (double spaced) and should conform to the following style.

For journals
Davis G R. 1990
Energy for planet earth
Scientific American 263(3): 55–62

For books
Carmichael J B and Strzepek K M. 1987
Industrial Water Use and Treatment Practices

For chapters of edited books
Sintak Y. 1992
Models and projections of energy use in the
Soviet Union
In International Energy Economics, pp. 1–53
edited by T Steiner

For grey literature
Togeby M and Jacobsen U. 1996
How conflicting goals concerning environment
and transport influence the policy process?
Paper presented at the Conference on Transport,
Energy and Environment,
3–4 October, Helsingor, Denmark

WBCSD (World Business Council for
Sustainable Development) and UNEP (United
Industry, fresh water, and sustainable
development
Details available at <www.gm-unccd.org/FIELD/
Private/WBCSD/freshwater.pdf>, last accessed
on 9 January 2004

Footnotes
Authors are requested to use as few footnotes
as possible, and keep their length to the
minimum. Footnotes should be indicated in the
text by superior Arabic numerals, which run
consecutively through the paper. They should be
grouped in order of appearance at the bottom of
the concerned page in numerical order and must
be double-spaced.

Accepted manuscripts
On acceptance, contributors are requested
to provide the editor the final version of the
article in soft and hard copy. Please observe the
following instructions.
- Tables, figures, illustrations, should be on
  separate sheets.
- Retain a back-up disc for reference and safety.

Proofs
One set of proofs will be sent to the author before
publication, which should be returned promptly
within 48 hours of receipt. Authors are urged
to check the proofs carefully as late corrections
cannot be accepted.

Offprints
Apart from one free copy of the journal to the
authors, 10 free offprints will be supplied to the
first author. Further offprints and copies of the
journal can be purchased at a reasonable cost,
if ordered when sending the final copy of the
article, or when returning the proofs.

Copyright
The responsibility for the contents of the paper
rests with the authors, not with the editor or
the publisher. Contributions are accepted for
refereeing on the understanding that they have
been submitted only to this journal and not
to any other journal. Only when each author
signs and submits the CTA (Copyright Transfer
Agreement) can TERI Press publish the article.
This CTA enables TERI Press to protect the
copyright material for the authors, but does
not affect the authors’ proprietary rights. The
CTA covers the exclusive rights to reproduce
and distribute the articles, including reprints,
photographic reproductions, or any other
reproductions of similar nature and translations,
and includes the right to adapt the article for
use in conjunction with computer systems
and programmes, including reproduction or
publication in machine-readable form and
incorporation into retrieval systems. Authors are
responsible for obtaining, from the copyright
holder, permission to reproduce any figures for
which copyright exists.