Information Management Resource Kit

Module on Digitization and Digital Libraries

UNIT 5. CREATION AND SHARING OF DIGITAL LIBRARIES

LESSON 3. CREATING A DIGITAL LIBRARY COLLECTION

NOTE

Please note that this PDF version does not have the interactive features offered through the IMARK courseware such as exercises with feedback, pop-ups, animations etc.

We recommend that you take the lesson using the interactive courseware environment, and use the PDF version for printing the lesson and to use as a reference after you have completed the course.

Objectives

At the end of this lesson, you will be able to:

- analyze an example of digital library collection development.

Introduction

In this lesson we will apply the steps involved in creating a digital library collection to a specific case study.

Let us walk through each step from identification of user needs to setting up the digital library.

The case study presented here deals with the creation of a digital library collection of student dissertations in an Engineering College.

Note that the case study is illustrative and may not exactly fit your situation!

However, you may find the approach useful.

Click here to read an alternative case study for a small library collection

Introduction

As part of their academic programme, students in the final semester of the B.E. (Bachelor of Engineering) course undertake a six months work project and submit a dissertation.

A copy of the approved dissertation is submitted to the Library and to the Academic Section.

New students are encouraged to consult previous dissertations in the library before deciding on a topic for their projects. Students from neighbouring engineering colleges also consult this collection.

The dissertations collection is thus one of the most frequently consulted collections in the library.

Introduction

Paula is a librarian working at the Engineering College.

Since the final semester is coming up, final semester engineering students always want to consult the Dissertations collection…

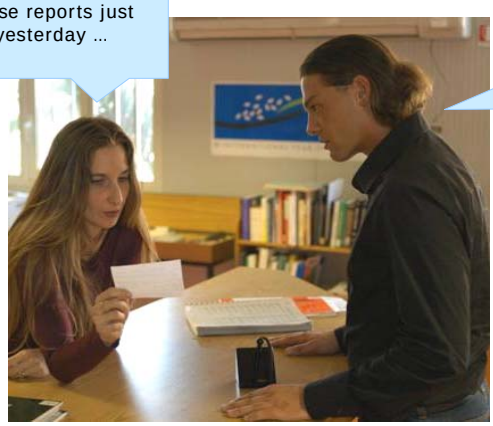
Good morning Paula!

I would like to consult a couple of these dissertations.

Paula, College Librarian

Introduction

I'm sorry but I gave out the last copy of these reports just yesterday …



Oh no! I need them to work on my project topic…

This is a frequent problem.

Sometimes students also complain of torn pages in reports due to frequent consultation.

The Library is unable to handle the situation satisfactorily…

Introduction

Probably, the best way to solve the problem is to set up a digital library of dissertations and provide online access to these on the campus network.

How does she go about setting up the digital library?

Management Approval

Paula discusses the student's requirements in the next meeting of the **Library Committee** and seeks their support.

... so I would suggest setting up a dissertations digital library collection.

Why not!

I think it's a good idea!

We will need to prepare a **detailed Planning document** outlining the tasks involved, resources required and a timeline for implementation.

The Library Committee also suggests setting up a five member Planning Committee consisting of the librarian, a senior staff member from the Academic Section, a faculty member, a student community member, and a senior staff member from the computer centre to prepare the Planning document.

Management Approval

OK, let's ask for Management Approval.

I will write to the Director, enclosing a brief project proposal prepared by Paula, requesting his approval for initiating the project. I will also recommend constituting the Planning Committee.

Michael, Chairperson of the Library Committee

The enclosed **proposal document** outlines the students' requirements, the need for setting up an online version of the dissertations to meet these requirements, the benefits of such a service, a possible strategy for setting up the digital library collection, and a rough indication of required resources.

Management Approval

The Director appreciates the initiative, approves the project in principle, and asks Paula to submit the Planning document for further consideration.

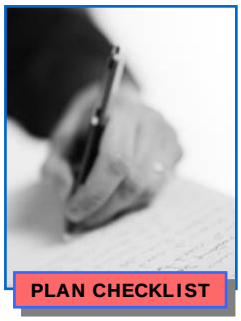
A formal notification is issued from the Director's office constituting the Committee.

Project planning

The Planning committee meets a few times over the next two months in order to develop the Planning document.

Paula co-opts her reference librarian into the committee to help her with researching, pilot testing and information gathering.

Project planning



Click on the button above to view and print the checklist

Now let's have a look at how Paula and her staff are facing the project planning stage.

The Planning document they are developing is based on the Plan Checklist, so we suggest you download and print this document in order to better follow the process.

Once in a while you will be asked to provide your suggestion on what decision should be made…

Project planning

Who has expressed the need for creating a digital library collection?

- Management
- Users
- Library staff

Please click on the answer of your choice

Project planning

Now, we should define the main purpose of the digital library collection…

What is the main purpose of this collection?

- improving preservation of some rare or delicate material
- providing improved access and visibility to certain material
- facilitating reuse of documents

Please click on the answer of your choice

Project planning

The committee assesses the **potential user community** likely to use the online collection.

They are predominantly students (both undergraduate and graduate) and faculty of the college.

This is the user population profile:

• Students: Undergraduate (3,000) and Graduate (200). External undergraduate students visiting the library in the course of the year to consult dissertations (about 1,000).

• Faculty: 250

• Other possible users: Staff in the Academic Section and Registrar's office.

Project planning

The committee examines the **volume and scope** of dissertations to be covered.

The volume of source material is about 250 dissertations per year, 50 dissertations per discipline for 5 engineering disciplines offered by the college. Each dissertation is about 100 A4 size pages.

Regarding coverage:

• retrospective coverage is limited to the last 4 years only, as it has been observed that student demand is usually for the previous four years' papers;

• in the future, the Committee suggests that the students be asked to submit an electronic version of the dissertation to the library, in addition to the print version.

Project planning

The selection and analysis of source material ends with the examination of the **copyright issues** associated with making them available on the Web.

Paula consults the legal office. In fact…

It is not clear to me who has the copyright for the dissertations – the students or the college? What precautions should we take?

I see… the student is the creator of the dissertation and thus has the copyright. We need to get the student's permission for digitization and hosting on the web.

The committee recommends that the Academic Section first obtains permission from students. Furthermore, due to the limited Internet bandwidth available to the college, the committee decides to limit full text access to the campus network; Only metadata access is allowed on the Internet.

Project planning

Now the committee has to consider some **key features** of the dissertations **digital library collection**.

Which of the following features would you select for the collection?

- The collection is static in nature (new documents are not added)
- The collection is dynamic in nature
- Usage is restricted to a limited group of users
- There are no user restrictions
- The collection is delivered on CD Rom
- The collection is delivered online

Please select the options of your choice (2 or more) and press "Check Answer".

Project planning

Document format for storage and delivery is a very important issue. The committee has identified the following criteria for format selection:

Many dissertations contain diagrams, formulas, non-Roman scripts, etc.

It is important to retain the look and feel of the original dissertations, and reproduce the original content accurately.

The format should preferably be standards-based and non-proprietary.

The solution must be cost-efficient!

Based on the committee's considerations, which of the following is the more appropriate format for storage and delivery?

- PDF
- Microsoft Word
- XML

Please click on the answer of your choice

Project planning

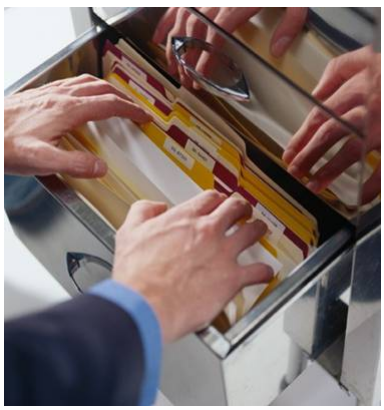
The estimate of storage size is approximately 5-7 MB per dissertation at an average of 100 pages per dissertation and about 70-100 KB per page in searchable image PDF format. This comes to about 7,000 MB (7 GB) for document files alone, for four years of dissertations.

All these decisions regarding the file format are made based on pilot experiments conducted in the library using an A4 sheet feeder scanner, and selected portions from a few dissertations.

The Committee notes that for efficient scanning the spine of the dissertations needs to be opened. This problem will not arise in future since students will be required to submit an electronic version of their dissertations!

More information about searchable image PDF format

Project planning



Although the storage and delivery format has been chosen, it is important to define a format for archiving and preservation.

What would you suggest?

- JPG
- TIFF
- PDF

Please click on the answer of your choice

Project planning

A few students and faculty members were consulted to define required access points (fields) for browsing and searching the dissertations collection…

Search / retrieval and presentation

The following access (search) points are defined: Course name, Course code, Department name, Exam year and month, Full text keyword search.

A few students were consulted to understand how they would prefer to search and retrieve from the dissertations digital library collection. The following access points (fields) were defined as useful for browsing the collection: Project title, Department name, Keywords, Submission year, and Student and Guide names. It is also useful to support searching on following fields: Project title, abstract, keywords and full text of the project report.

On submission of the search query, the system would first respond with a display of metadata with a link to a PDF file for each matching dissertation. Users would then click on the full text link to view the full exam paper.

Project planning

Metadata requirement

For each dissertation, the following descriptive metadata will be assigned: Project title, Department name, Submission year, and Student and Guide names, Abstract, Keywords, and number of pages.

The following administrative metadata was identified for addition to each exam paper: Date when dissertation was added to the collection .

Project planning

Digital library software solution

The committee decides to opt for in-house development.

The committee considers various options including in-house application development, free open source digital library software and the purchase of commercial software. The use of open source free software is a very attractive proposition, but the committee decides to opt for in-house development as the Computer centre has the necessary expertise in this kind of application development; Minimal development effort is required in adapting these to the dissertation collection. In-house development is also considered desirable for customizing the application for access management (e.g. user authentication), future enhancement and the possible use of the application for other similar collections.

The database-driven approach has been chosen for developing the application on a Linux platform.

Metadata would be stored in a SQL (MySQL) database. A 2-level hierarchical folder structure of year and course name was designed within which respective PDF files will be stored.

Required applications for search and retrieval would be developed using PHP programming language. For supporting full text search, PDF files will be automatically converted into text format using freely available conversion software and these text files will be used for supporting full text searching.

It is also deemed necessary to develop a web-based content management interface to help library staff to add metadata and to edit any errors.

Project planning

Digital library collection hosting and administration

It has also been decided to host the dissertation database and website on an existing web server in the Computer Centre.

The university already has an excellent intranet infrastructure. Moreover, the collection will be jointly managed by the Library, Academic Section and the Computer Centre. The Library will manage the digitization workflow and provision of services and metadata content management and the Computer Centre will maintain the collection web server, website and the application software. Academic Section will provide support for authentication of student related information.

Project planning

The key decision to be made is whether to carry out digitization in-house or to contract it to an external vendor..

The committee spent considerable time assessing **digitization requirements and workflow**.

It has decided to out-source digitization, based on cost-related considerations.

A digitization requirements specification is prepared, including both digitization and metadata assignment, and vendors are invited to submit digitization output for a few sample dissertations along with their quote for undertaking the full job. After a careful evaluation of submissions, a vendor is selected.

Project planning

Reasons for outsourcing

Pilot tests are conducted to calculate the staff time required and costs involved in producing the required PDF version and for metadata assignment.

Costs are obtained from external vendors for the required quantity and quality of work (e.g. opening and rebinding of spines of dissertations; need for creating searchable image PDF and TIFF versions of dissertations, metadata assignment; quality checking and delivery of PDF files with metadata).

The large number of pages to be digitized requires that the library has to hire more staff, train them and also buy 2 or 3 sheet feed scanners – this would cost much more money. Additional PCs would also be required for metadata assignment and quality assessment.

Project planning

Digitization workflow

The following are steps in the digitization workflow:
Preparation of print dissertations for digitization (in Academic Section), including a unique identifier assignment, and authentication.
Collection of print dissertations by the vendor.
Digitization at the vendor's site, including a QA (quality assessment) test done by the vendor as per the university requirements
Entry of metadata on Excel sheets at the vendor's site.
Delivery of digital material (PDF and TIFF files), metadata (on CD-ROM) and original exam papers by the vendor
QA testing in the library to verify the quality and completion of digitization and metadata
Loading of metadata and PDF documents to the digital library collection server, at the computer center.
Due to frequent use the print dissertations in the library were found not good enough for scanning. Fortunately the Academic Section had maintained well the copies in its collection. It was decided to use these for digitization.

File naming

A coding scheme will be developed to uniquely identify each dissertation as a file name identifier to be used during digitization, QA (quality assessment) tests and later in the database for linking metadata with the dissertation.

Project planning

It is now time to discuss the **resources required** for implementing the project.

Let's focus on the **IT infrastructure**. Which of the following resources should the committee include in the project?

- Backup online and offline storage for the collection, PDF and TIFF files.
- A Web server for hosting the collection and website.
- Several PCs in the Library for staff managing the digitization workflow and metadata.
- Digital library application software development.
- Network connectivity.
- Additional disk storage space for the collection web server.

Please select the options of your choice (2 or more) and press "Check Answer".

Project planning

Other resources to be included are…

PERSONNEL

The library will require two temporary staff members for six months to manage the digitization workflow, and QA, related to four years of print dissertations. Library will need half FTE (full time employee) to handle the workflow and QA during regular operations (to handle updating of the collection with new dissertations).

The Computer Centre will require a temporary programmer for six months to support the programming staff in developing applications and setting up the collection website.

FINANCIAL REQUIREMENTS

The project will require money for the following…

- PCs in the Library (2).
- Additional disk storage for the collection server.
- Backup online and offline storage.
- Digitization costs payable to the external contractor (for initial digitization).
- Two temporary staff in the Library for 6 months.
- Half FTE in the Library (at the end of 6 months).
- An additional temporary programmer for application development support in the computer centre, for six months.

Project planning

Let's now define the implementation schedule and timeline…

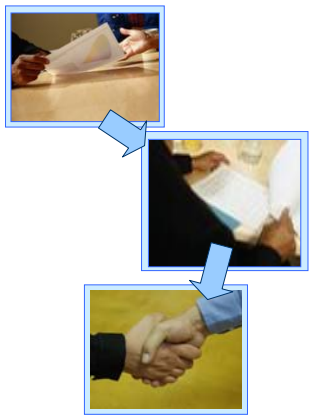
The planning committee estimates that it will take about **six months** to implement the project covering the four year dissertations.

This estimate is based on the estimated time needed for the digitization and development of the digital library application software.

The Planning document is now completed!

Click here to download and print it.

Project implementation



The Planning Committee submits the completed plan document to the Library Committee.

Subsequent to approval by the Library Committee, Paula hands over the plan document, with an executive summary, to the Director for obtaining his approval for project implementation and additional resources required.

Convinced of the usefulness of the initiative, the Director approves it and also asks the Planning Committee to implement the project.

Project implementation

The committee implements the project in the following manner:

- It enters into a formal contract with the external contractor for digitization.
- It hires the required temporary project staff.
- It procures the required additional hardware.
- It initiates the digital library application software development.
- It initiates the digitization workflow.
- It tests the digital library application and user interface with sample data.
- Content received from the contractor is loaded to the digital library, after QA.
- The collection website is set up.

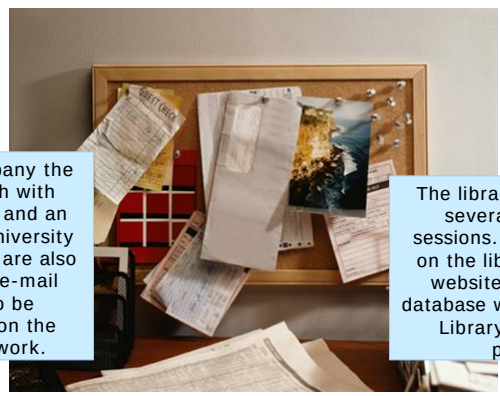
Testing and release

The dissertations digital library in its completed form was released for testing to a few identified groups of people in the library, computer centre, academic section and a few students and faculty members.

Feedback from these groups was incorporated into the final product. It included modifications in the search and retrieval interface and in the help screens.

Promotion and dissemination

Finally, the promotion and dissemination of the new service is managed as follows:



We will accompany the service launch with posters, flyers and an article in the university newsletter. We are also preparing an e-mail message to be broadcasted on the campus network.

The library staff will conduct several demonstration sessions. We will create links on the library and university websites.The dissertation database will be included in the Library user orientation programme.

Summary

Planning and implementing the dissertations digital library collection required the following steps:

Step-1: Obtaining management approval and the constitution of a planning committee consisting of key stake-holders interested and involved in setting up the online collection.

Step-2: Preparation of the plan document consisting of the following: Needs, purpose and benefits of the collection; user community; source material - volume, coverage and attributes; Copyright; digital library collection requirements; digitization requirements and workflow; resource requirements; and implementation schedule and timeline.

Step-3: Obtaining approval from management for the plan, required resources and project implementation.

Step-4: Testing the online collection and subsequent release for use by students.

Step-5: Promotion and dissemination of the online collection.