UNITED NATIONS EDUCATIONAL,
SCIENTIFIC AND CULTURAL ORGANIZATION

DIGITAL LIBRARIES IN EDUCATION

Specialized Training Course

STUDY GUIDE

UNESCO INSTITUTE
FOR INFORMATION TECHNOLOGIES IN EDUCATION

UNESCO
UNESCO Institute for Information Technologies in Education (IITE)

WORKING GROUP:

Authors
Ian H. Witten (University of Waikato, New Zealand) – Coordinating editor
David Bainbridge (University of Waikato, New Zealand)
David M. Nichols (University of Waikato, New Zealand)

in conjunction with:
Wayne Mackintosh (University of Auckland, New Zealand)
Tarikere Basappa Rajashekar (Indian Institute of Science, India)

IITE course coordinator
Azat Khannanov (IITE)

Digital Libraries in Education. Specialized training course. Study Guide

The IITE specialized training course *Digital Libraries in Education* has been developed in the frame of UNESCO cross-cutting theme project *Methodologies for Digital Libraries*. The project aims to give an overview of current and future technologies and applications for digital libraries (DL) including ethical, social, pedagogical, organizational, and economic aspects as well as their impact on learning, cultural, and scientific activities.

The course is about the use of DL in education, including emerging areas of application and current and future technologies for creating and distributing DL. It shows educators how to build their own digital library collections for use in the courses they teach. It touches on large-scale national and international DL for education, but is more strongly oriented towards low-budget methods of building and maintaining DL by creative individuals and by self-organized communities of educators, ranging from the personal to institutional levels.

The IITE specialized training course *Digital Libraries in Education* composed of three parts – this Study Guide and two CD-ROMs with course readings and auxiliary materials.

FOR FURTHER INFORMATION PLEASE CONTACT:
UNESCO Institute for Information Technologies in Education
8 Kedrova St. Bld. 3, Moscow, 117292, Russian Federation
Tel.: 7 495 129 2990
Fax: 7 495 129 1225
E-mail: info@iite.ru
Web: www.iite.ru

TABLE OF CONTENTS

PREFACE INTRODUCTION TO THE COURSE

Welcome to the course *Digital Libraries in Education*.

Digital libraries are large, organized collections of information objects. Well-designed digital library software has the potential to enable non-specialist people to conceive, assemble, build, and disseminate new information collections. This has great social impact because it democratizes the dissemination of information. In particular, it will revolutionize the way in which education is conducted and educational materials are prepared.

The emergence of the World Wide Web is changing society's view of information by making unprecedented volumes of information freely available. Of course, it is an unreliable source of enlightenment, and undiscriminating use is dangerous – and, unfortunately, widespread. Nevertheless, the web abounds with accessible, high-quality information. Many educational establishments, international organizations, social groups, non-profit societies and charities make it their business to create sites on which they collect and organize information.

Viewed as an educational resource, however, the web exhibits serious deficiencies: uneven and erratic coverage, transience and unpredictability (will this piece of information still be there tomorrow?), and manifest dangers (will my students encounter inappropriate information?). But a far greater tragedy is that whole segments of society become disenfranchised – for while most family homes in rich countries have some degree of access to the Internet, only a tiny minority of citizens in the developing world can tap this wealth of information. Digital libraries address these problems by providing reliable sources of appropriate material. They empower educators to create collections specifically for their students, collections that mix information from different sources. They permit alternative means of distribution (e.g. CD-ROM/DVD, a very practical format in developing countries).

This course is about the use of digital libraries in education, including emerging areas of application and current and future technologies for creating and distributing digital libraries. It shows educators how to build their own digital library collections for use in the courses they teach. It touches on large-scale national and international digital libraries for education, but is more strongly oriented towards low-budget methods of building and maintaining digital libraries by creative individuals and by self-organized communities of educators, at levels ranging from the personal to the individual institution.

Free, open-source software is a key component of this strategy, and high-quality open-source digital library software is already available. This course includes the Greenstone digital library software, and in it you, the educator, will learn how to use Greenstone to create your own information collections from your own material, incorporating material from other digital libraries and from the web if you so desire, and distribute it to your students in the form of a web site or a self-installing CD-ROM – or both.

Widely adopted international standards are another key component. They are vital for ensuring that you can incorporate documents in different forms and from different sources. They enable individual libraries to share information by communicating with each other, and provide the basis for coordinated regional, national, and international strategies for creating and disseminating educational material.

This course is based on two external resources, apart from this Study Guide:

1. The textbook *How to build a digital library*, by Ian H. Witten and David Bainbridge, published by Morgan Kaufmann, San Francisco, California, in 2003: students of this course must acquire or borrow this book.

2. An interactive CD-ROM (requires a Windows computer) entitled *Digitisation and Digital Libraries*, which is a module of the Information Management Resource Kit produced by the Food and Agriculture Organization of the UN, Rome, in 2005: this is distributed with the course.

3. Course CD-ROM entitled *Digital Libraries in Education* (distributed with the course), which is set of additional readings, example collections, and Greenstone digital library software that is used for the practical component of the course runs on Windows, Linux, or Macintosh OS/X.

This is the context for the course of study that you are embarking on. Before beginning, here is a summary of what the course is about, who it is for, how it is organized, and the assumptions we are making about how you will study it.

What can you expect to learn?

This course will tell you:

- what digital libraries are;
- how they are being used in education – and how they might be used;
- what metadata is and how it helps in organizing digital libraries;
- the different formats in which electronic documents are represented;
- how multimedia can be used in digital libraries;
- how to build and manage digital library collections;
- what standards exist for digital libraries and educational metadata;
- what systems are available for constructing institutional document repositories.

Who is the course for?

The target audience consists of two main groups:

1. Educators:

- teachers;
- teacher trainers;
- tutors.

2. Information professionals in the field of education:

- librarians;
- digital library developers;
- information system managers;
- educational authorities.

The course also has a secondary audience comprising:

3. Students.

4. Researchers.

The focus of the course is on education at the secondary level and higher level.

What previous knowledge is assumed?

This course has been designed on the assumption that you have some experience in organizing course material for conventional classroom teaching environments, and that you have some practical hands-on experience of using a computer for tasks, such as word processing. We so assume that you are a library user.

How much study time is needed?

We are assuming that you will devote approximately 50 hours of study time to this course, including answering all the assignments – and up to 70 hours if your native language is not English or you are slow at working with computers. This includes 10–15 hours of practical exercises building digital library collections of various kinds; it also includes 3 hours of optional "enrichment" exercises.

How is the course organized?

The course is organized in five modules, each of which has two or three study units. Your study time will *not* be evenly distributed between the modules. We expect it to be approximately as follows:

Module	Number of units	Total study time
1	3	7–10 hours
2	2	7–10 hours
3	3	15–20 hours
4	3	10–15 hours
5	3	10–15 hours
total		**50–70 hours**

What resources will you need?

Each trainee will require:

- This Study Guide: IITE course on *Digital Libraries in Education*.
- Accompanying Course CD-ROM entitled *Digital Libraries in Education* containing auxiliary material for *Digital Libraries in Education*. This contains:

  – the Greenstone digital library software;
  – example collections, all manuals, etc.;
  – sample files for the exercises;
  – a digital library collection of course readings, including this Study Guide;
  – a collection based on the textbook *How to build a digital library*;
  – various other sample collections.

- Textbook: *How to build a digital library*, by Ian H. Witten and David Bainbridge. Morgan Kaufmann, CA, 2003
- Interactive CD-ROM: Information Management Resource Kit (IMARK), Module *Digitisation and Digital Libraries*, Food and Agriculture Organization of the UN, Rome, 2004.
- Access to a computer (Windows 98 or above).

Internet access is not required for the course, but some aspects may be enhanced by it.

Evaluation procedure

Each module comprises several assignments, each of which includes a series of questions. The primary intent of the assignments is to lead you through the course material. They also provide an opportunity for you to test yourself. If you are taking this course in an educational workshop, your trainers may use the assignment questions to evaluate your performance. If you are an independent distance learner, you may be asked to submit your answers by mail or e-mail to the institution that has organized your course.

Acknowledgements

This course was commissioned by the UNESCO Institute for Information Technologies in Education (IITE) and prepared by a New Zealand-based team of specialists in the study and practice of digital libraries and education. Members of the team were David Bainbridge (Scotland and New Zealand), Dave Nichols (England and New Zealand), and Ian H. Witten (Canada and New Zealand). We are grateful to Wayne Macintosh (South Africa and New

Zealand) and Dr T.B. Rajashekar (India) for their help in the early stages of this project. We thank all members of the New Zealand Digital Library Project at the University of Waikato, particularly Michael Dewsnip who produced the course CD-ROM and Shaoqun Wu for her careful testing of the material. The course is closely modelled on the specialized IITE training course *Information and Communication Technologies in Distance Education*, and we are most grateful to Michael G. Moore of Pennsylvania State University who coordinated the production of that course.

Ian H. Witten
Coordinating editor

MODULE 1 THE CONCEPT OF DIGITAL LIBRARIES AND THEIR ROLE IN EDUCATION

Goal

To understand the concept of digital libraries, perceive the opportunities – both actual and potential – that they present in education, and experience using one.

Objectives

Upon completion of Module 1, you will be able to:

- compose your own definition of a digital library;
- persuade your library director not to embark on wholesale digitisation of the entire library;
- distinguish digital libraries from the World Wide Web;
- explain why digital libraries are especially important in developing countries;
- list examples of how digital libraries are being used in education, and explain several ways in which they are being used;
- write a brief proposal for how digital library software could enhance students' educational experience in your own local setting;
- explain the difference between open source and commercial software;
- install a Greenstone collection on your computer;
- use all features of the Greenstone reader's interface to find information;
- in the context of this software, explain what is meant by a *document*, a *collection*, and a *library*;
- distinguish three different ways of building collections in Greenstone.

Introduction to Module 1

The first module explains what is meant by a digital library, shows you how they are used in education, and lets you experience the use of a digital library collection for finding information.

The module is divided into three units. The first analyses the concept of digital libraries and shows examples in an eclectic range of areas, with an emphasis on cultural, historical, and humanitarian applications, as well as technical ones. The second introduces the ways in which digital libraries are being used in education and explores potential uses, too. The third introduces the Greenstone digital library software, which is used throughout this course.

This module, like the others in the course, contains activities designed to help you think about digital libraries and their role in education. The activities can also be used for assessment (and grading by your instructor).

Module 1 Readings

Unit 1.1. Introducing digital libraries

Reading 1 The textbook *How to build a digital library*: Preface and all sections of Chapter 1 except Section 1.4.

Purpose This opening material explains the concept of a *digital library* and sets it in the historical context of library evolution over the ages. It gives several illustrations. It discusses the role of digital libraries in developing countries and reviews issues of copyright and ethics. As you read this chapter, think about how these definitions and scenarios might be applied in your own local library context.

Reading 2 CD–ROM IMARK: Unit 1 – Conceptual overview (4 lessons).

Purpose This gives a wealth of further information on digital libraries, including many examples, and a detailed account of practical copyright issues. It is recommended as enrichment material that will increase your understanding and experience of digital libraries and their application.

Unit 1.2. Digital libraries in education

Reading 1 Course CD–ROM, Readings: "The roles of digital libraries in teaching and learning," Gary Marchionini and Hermann Maurer. *Communications of the ACM* (1995), Vol. 38, No. 4, pp. 67–75. http://www.ils.unc.edu/~march/cacm95/cacm.html (also available in the folder "Readings" of the accompanying Course CD–ROM entitled *Digital Libraries in Education*).

Purpose This article describes how digital libraries are evolving to meet the needs of teaching and learning and identifies issues for continued development. As you read it, think about the kinds of teaching resources that you currently use and what you might be able to do if a vastly increased wealth of resources were easily available.

Reading 2 Course CD–ROM, Readings: "Digital libraries and education: trends and opportunities," Hans Roes. *D–Lib Magazine*, July/August 2001, Vol. 7, No. 7/8. http://www.dlib.org/dlib/july01/roes/07roes.html (also available in the folder "Readings" of the accompanying Course CD–ROM entitled *Digital Libraries in Education*).

Purpose This article will acquaint you with a wide variety of strategic, policy, and organizational issues that confront librarians and educators who actively address the increased use of information and communication technologies in higher education.

Unit 1.3. The Greenstone digital library software

Reading 1 The textbook *How to build a digital library*: Section 1.4.

Purpose This briefly introduces the Greenstone digital library software. As you read, recall the different areas that you identified in Unit 1 for applying digital libraries in an educational setting and consider the extent to which Greenstone may satisfy them.

Reading 2 Course CD–ROM *Digital Libraries in Education*, Readings: *Digital Libraries in Education* collection.

Purpose This collection contains all the readings for the course, including this Study Guide, in a form that is fully searchable, and also illustrates some other features of Greenstone collections. It serves a dual purpose: to give you some experience in browsing around a digital library and as a source for the readings.

Reading 3 Course CD–ROM *Digital Libraries in Education: How to Build a Digital Library* collection.

Purpose	This collection indexes the textbook and also contains colour versions of all the figures. It illustrates an unusual kind of Greenstone collection and simultaneously provides a comprehensive index to this resource for you to consult during the course.
Reading 4	CD–ROM IMARK: Lesson 6.1. – Greenstone tutorial.
Purpose	This self–study lesson introduces many aspects of Greenstone. Study it to learn about installation options, interface features, and collection–building approaches.

UNIT 1.1 Introducing digital libraries

Before reading further, take some time to reflect on what a digital library is, in particular, how it might differ on the one hand from a conventional bricks-and-mortar library and from the World Wide Web – on the other.

Assignment 1

Think about your local university library or any other library that you are familiar with.

- What major challenges would be encountered when making a digital library with the same contents?
- How would library users benefit, and in what ways would they suffer?
- Given that only limited resources were available, what kinds of material would you choose to digitise first?
- What might an electronic analogue of the present library catalogue look like?
- What additional searching and browsing possibilities might the digital library offer?
- In what ways would your digital library differ from the World Wide Web?
- Why do people with access to the web need digital libraries?

If you're unsure about your answers to some of these questions because you don't yet have any experience of digital libraries, don't worry! – we will revisit them at the end of the unit. Meanwhile, just carry on.

Educational digital libraries, like any other digital library, can exist at many different scales. The examples seen today are merely the first tentative and hesitant steps, for digital library technology is only just beginning to be accessible to people and institutions at varying levels. How things develop in the future will depend on visionary and creative individuals, as well as a lot of hard teamwork. You, as a student of this course, will help mould the development of digital libraries in education: you too are a potential visionary.

Assignment 2

Read the Preface and the beginning of Chapter 1 (pp. 1–5) of the textbook and write a paragraph (one for each point below) describing your vision of:

- a large-scale educational digital library, at the national level;
- a medium-scale educational digital library, created by a university department;
- a small-scale educational digital library, created during a course by an individual student.

There are many different definitions of the term "digital library": a selection is given below. As you read them, think about the respects in which they resemble each other and how they differ from each other and from traditional libraries.

The first two definitions suggest that a digital library is the same as a traditional library, or a traditional information retrieval system, except that the material is represented digitally.

- *A library that encodes journals, books, and information into a digital format.*
- *A collection of texts, images, etc., encoded so as to be stored, retrieved, and read by computer.*

Other definitions emphasize traditional library functions but go further by stipulating that new services can be offered. The second definition below envisages a rather comprehensive set of services, which includes preservation of information (under "protection") – a traditional library virtue that, surprisingly, is not mentioned by any of the other definitions.

- *A collection of digital representations of information content, along with hardware, software, and personnel to support the functions of a traditional library plus knowledge worker operations like searching, browsing, and navigation.*
- *An integrated set of services for capturing, cataloguing, storing, searching, protecting, and retrieving information.*

Some definitions emphasize technical features: size, scale, distributed structure, network access. The second one below notes the possibility of capturing dynamically changing information, something a conventional library cannot do.

- *A collection of a very large number of digital objects, comprising all types of material and media, that are stored in distributed information repositories and accessed through national computer networks.*
- *A large collection of information that has been stored in digital form. A digital library can include documents, images, sounds, and information gathered from ongoing events (e.g. continuous pictures from a weather satellite).*

Others reflect a strong relationship with the web. Is the World Wide Web a digital library, or not? – we revisit this question after the next assignment.

- *Digital libraries can include reference material or resources accessible through the World Wide Web. Digitised portions of a library's collection or original material produced for the web can also be included in a digital library.*

A related term is "virtual library". While this might be considered synonymous with "digital library" in that both emphasize the intangible nature of the material stored, it is more often used to denote a portal to information that is available electronically elsewhere.

Assignment 3

Read Sections 1.1 and 1.2 (pp. 5–20) of the textbook and write an essay describing your views on the extent to which digital libraries are likely to compete, co-exist with, or complement conventional bricks-and-mortar libraries in the field of education over the foreseeable future – say 20 years (suggested length: 500 words).

Enrichment exercise 1

The IMARK Module gives a wealth of further information on digital libraries. At this point, we recommend that you work through Lessons 1.1 and 1.2. The former defines the concept of digital libraries, discusses their benefits, and sketches the processes involved in their creation and use. The latter describes many examples of digital libraries that exist today. Working through this material will increase your understanding and experience of digital libraries.

Most computers automatically start to install the software required to run the IMARK Module. (Technically, they start automatically if AutoPlay is set for your computer: usually it is.) If this is not the case, that is, if nothing happens when you insert the disk, open My Computer, open the CD-ROM drive (normally D:), and execute the program IMARK_VI0Een *by double-clicking it.*

Once the installer starts it will ask you a series of questions for initialization purposes. For a quick start, perform the following:

- *Click <**START COURSE**>*
- *Click <**Start Learning**>*
- *Accept the terms and conditions*
- *Fill out the requested details in the learner profile*
- *Click <**Save**>*
- *Register, if on line (recommended)*
- *Choose <**Start course**>*
- *Enter lesson 1.1 ...*

The question was raised above whether the World Wide Web is a digital library. We think not. As the Assignment 3 reading explains, the web lacks two crucial elements: organization and selectivity. A library is an organized collection, organized according to some overall principles and not as a tangled web of links. And the information in a library has been selected according to some criteria, whether explicitly articulated or not. You might think of libraries as "curated" collections, literally ones that are "taken care of" by some guiding body, such as an administrative director. No one takes care of the web. But since the early days of the web, people have tried to bring order to it and make it easier to find things by developing "search engines." We now briefly introduce a popular web search engine.

Case study: The Google search engine, http://www.google.com

During 7 years since its inception in 1998 until the time of writing (early 2005), the Google search engine was one of the fastest–growing cultural phenomena the world has ever seen – and one of its most economically profitable enterprises.

Search engines begin by making a copy of the entire web or at least as much of it as they can. (Google is thought to have several copies distributed across servers in different parts of the U.S.) They keep the copy up to date by periodically re–freshing it from the web, perhaps every few weeks. From it they build an index that records, for every word that appears in the entire text, the numbers of the web pages in which it appears. This data structure, which is very large, makes it pos–sible to find the numbers of all the pages that contain a particular word or set of words. When users present a "query", which is a set of words, the system returns a list of "search results", a list of the pages that contain all those words.

For most queries, the list of matching pages is huge. Google's success stems from two major innovations. One is the way in which it sorts the results for presentation to the user. It first analyses the structure of links between web pages in order to determine a quantity that represents the importance, or "rank", of each page. The rank is determined by the number of pages that point to it – its popularity, in other words – and by the ranks of those pages. The idea is that if several high–ranking pages point to your page, its rank will also be high. The apparent circularity of this definition does not, in fact, invalidate it, though it does make page ranks tricky to compute.

Google's second key innovation is based on the observation that a page is often best described by the pages that point to it. Just as your own opinion about yourself carries less weight than other people's opinion of you, there is a world of difference between a claim that a web page makes about itself – for example, that it has "good" coverage of a certain topic – and similar statements about the page made by other pages that point to it. Google uses the text associated with a link, called "anchor text", as an important clue to the content of the page the link points to.

Almost overnight, Google made the web far more useful, particularly for non–specialist users, many of whom regard it as the Internet's front door. It provides a way to find information in an otherwise anarchic system.

Although full-text searching, typified by web search engines, can be a helpful finding aid, it is less sophisticated and generally less useful than the kind of organization that is standard in libraries – author indexes, lists of titles, subject indexes, keywords, summaries, and so on. Moreover, the web lacks selectivity, that other key library virtue. However, although the web is not in itself a digital library, it certainly contains many examples of digital libraries. Moreover, focused subcollections of the web, formed by choosing pages according to sound principles of selection and organizing them according to sound principles of organization, certainly can be a digital library. You will learn how to create such collections in this course.

Digital libraries are especially critical in developing countries. One reason for this is that traditional sources of information – for example, books – are often hard to obtain there, and digital libraries make it possible for large numbers of people to access them at a potentially low replication cost. Another is that web access in developing countries is typically low, widening the knowledge gap between the developed and developing world. Digital library technology can ameliorate this, because – despite many people's assumptions to the contrary – it does not have to depend on the Internet for distribution.

Developing countries are eager consumers of digital libraries. But it is crucial for sustained development that these countries are not relegated to becoming "read-only cultures" in the digital revolution. One way to ensure this is by participating in producing information collections. If you live in a developing country, this course will help by showing

you how to create your own educational digital library collections. You will learn how to make these collections accessible internationally over the web, or put them on to CD-ROM or DVD, thereby creating a self-contained information collection that can be viewed even on modest computer equipment.

Assignment 4

Read Section 1.3 (pp. 20–24) of the textbook. From your own personal experience, explain how the lack of access to information negatively impacts education and identify several different application areas for digital libraries in the realm of education that will differentially benefit developing countries.

Copyright is usually a central issue for digital libraries. Copyright law varies from one country to another, and if you have practical questions about particular situations you need to consult a copyright lawyer.

Enrichment exercise 2

IMARK Lessons 1.3 Copyright: basics and legal framework *and* 1.4 Copyright issues and procedures concerning libraries *give a more detailed account of copyright issues. At this point, we recommend that you work through Lesson 1.3. Lesson 1.4 is not part of the present course, but if you have time you should work through it, too. This material will enrich your understanding of practical issues concerning copyright.*

Assignment 5

Read Section 1.5 (pp. 28–35) of the textbook.

The Google search engine (see box above) makes a private copy of all web pages it encounters to serve as a basis for indexing. When a user clicks on a link in the search results, Google transfers the user to that page on the web – not to the private copy. However, the user has the option of clicking on a secondary link to see the copy in Google's cache – a particularly useful option if the original page no longer exists. When Google shows the copy, it precedes it with a clear disclaimer that states that they are not the original web page and gives the date of download. For example, here is Google's disclaimer for the page www.greenstone.org:

> This is G o o g l e's cache of http://www.greenstone.org/ as retrieved on 4 Oct 2006 05:10:20 GMT.
> G o o g l e's cache is the snapshot that we took of the page as we crawled the web.
> The page may have changed since that time. Click here for the current page without highlighting.
> This cached page may reference images which are no longer available. Click here for the cached text only.
> To link to or bookmark this page, use the following url:
> http://www.google.com/search?q=cache:vs5EsdeaNiAJ:www.greenstone.org/+greenstone+digital+library&hl=en&gl=ru&ct=clnk&cd=1
>
> *Google is neither affiliated with the authors of this page nor responsible for its content.*

Comment on Google's policy from an (a) ethical, (b) legal point of view.

Unit closing

Upon reaching this point you have learned a great deal of background information on digital libraries. The reading for Assignment 2 included four scenarios intended to dispel the myth that digital libraries are no more than a routine development of traditional libraries with bytes instead of books. The reading for Assignment 3 discussed the concept of *digital library* and placed it in the historical context of library evolution over the ages. This course explicitly envisages applications in developing countries and the next reading (Assignment 4) examines the use of various digital libraries in this context. The final reading was concerned with the difficult question of copyright and ethical issues surrounding the reuse of material.

Let's revisit the questions that were raised at the beginning, in Assignment 1, and give you a brief commentary on them from our standpoint. Do not worry if reading this unit has changed your views – that's what education is all about! We asked you to think about your local university librar, or any other library that you are familiar with, and posed these questions.

- *What major challenges would be encountered when making a digital library with the same contents?*

To digitise an entire library would be a formidable undertaking. One challenge would be the sheer magnitude of the task of converting all the library's contents into digital form by scanning them into the computer. Scanning books is much easier if they can be taken apart by removing their spines, but in this application the job would probably have to be done non-destructively. Dealing with old and fragile material would be particularly time-consuming. But an over-riding problem would be the legality of the whole enterprise. Typically only a small percentage (if any) of a library's content is out of copyright and it would be impossible for the library to obtain permission to digitise from all copyright holders. Overall, we hope that studying this unit has convinced you that digital libraries are unlikely to be simply digitised versions of existing physical libraries.

- *How would library users benefit, and in what ways would they suffer?*

Making the digital library available over the web would open up readership to a host of potential new users. A major advantage for existing library users would be convenience of remote access – from workplace or home, if they had an Internet connection, or perhaps from branch libraries. Another would be the possibility of new ways of finding information in the library – for example, full-text search. However, most people do not enjoy reading books online, and if the collection had been destroyed during the digitisation process (see above) library users would suffer immensely. Moreover, the cost of the operation would inevitably mean that resources were diverted from other projects that might well be more worthwhile. Again, it is highly unlikely that the benefits would outweigh the tremendous cost.

- *Given that only limited resources were available, what kinds of material would you choose to digitise first?*

This is a question that is not really addressed by the readings in this unit. The answer depends greatly on local considerations. Priority should be given to locally produced material for which copyright permission can be obtained that is in high demand by readers – examples might include course notes or other high-volume teaching material. Digitisation can be used to promote special strengths of the library, such as unique collections of primary source material. It can also reduce handling of fragile originals, especially if they are heavily used. Other considerations include resource-sharing partnerships with other libraries and available funding opportunities.

- *What might an electronic analogue of the present library catalogue look like?*

Again, this unit has not examined user interface issues, though you will see many examples in subsequent units. Note that in many libraries today – the overwhelming majority, in developed countries – the catalogue is already electronic. These provide similar facilities to card catalogues, including searching and browsing by title, author, and topic. Once library users have learned how to work with them, most prefer electronic catalogues. A big disadvantage of an entirely digital library, although not really the fault of the catalogue, is the loss of serendipity through not being able to browse books on the shelves. In a physical library, readers use the catalogue to find a book and then often find great value in browsing neighbouring books. In a digital library this is not so easy to arrange.

- *What additional searching and browsing possibilities might the digital library offer?*

This unit has drawn attention to full-text search, which is the foundation of web search engines like Google, and many digital libraries offer this facility. There are numerous other possibilities that have not been mentioned here, such as automatic extraction of hierarchies of phrases from the documents in the library or automatically produced lists of acronyms and their definitions, again extracted from the documents themselves. You will meet both these examples again in Unit 1.3.

- *In what ways would your digital library differ from the World Wide Web?*

You have learned that digital libraries are focused collections of information that has been carefully selected and organized, in contrast to the web in which anyone can add information and organization is haphazard. Digital libraries

have boundaries; the web does not. As was mentioned above, this difference is sometimes characterized by the word "curated", which applies to digital libraries but certainly not to the web. And digital libraries can bring information to people who lack access to the web.

- *Why do people with access to the web need digital libraries?*

It may well be that the information in a digital library is already on the web – most of the above-mentioned collections are. However, although full-text search is a powerful means of locating information, it is often hard to find what you want on the web and to be sure that what you have found is authoritative. As focused collections of selected material, digital libraries are usually better and more reliable sources of information than the web at large.

We end with a final caveat to this assignment: there may be more valuable ways of deploying the resources that would be required for the monumental task of digitising an entire library. The readings introduced many examples of worthwhile digital libraries, such as the Humanity Development Library in the Kataayi organization in rural Uganda, the online physics archives for researchers, the efforts to preserve the language and traditions of the Zia Pueblo in New Mexico, and libraries of popular music. Collections for disaster relief, preserving indigenous culture, and locally produced information are also mentioned in more general terms. This unit and the associated readings have stressed that digital libraries are about increasing access to information: worthwhile opportunities exist wherever there is a need for this.

The first unit of Module 1 has taken you through the first chapter of the textbook. To summarize, you have:

- reflected on what a digital library is and how it might differ from a conventional library and even from a digitised version of a conventional library;
- articulated some visions of digital libraries in education;
- thought about how you might define a digital library and reviewed several different definitions;
- differentiated digital libraries from the World Wide Web;
- learned about web search engines and thought about their power and their weaknesses in finding information;
- considered roles for digital libraries in developing countries;
- reviewed the ethical and legal basis of copyright.

Now you are ready to take the first step towards learning how digital libraries can be and are being used in education. Please move on to Unit 1.2.

UNIT 1.2 Digital libraries in education

Please begin this unit by reflecting on how you use documents in your own learning or teaching.

Assignment 1

Think about a particular course you have taught or are familiar with from a student's point of view:

- What information resources do students require?
- How could each student's learning experience be enriched if they had access to selected relevant information on the web?
- How could the learning experience be enriched if students had access to relevant information resources that you yourself possess?
- Would students benefit from being able to build their own collection of resources relevant to the course, and if so, how?
- Will students find it useful to consult material they have learned in the course in (say) 5 years' time? How will they do so?
- What might a unified information resource for the course look like?

There are three distinct roles that digital libraries can play in education:

- as an environment for learning (student experience);
- as an authoring space (again, in support of student experience);
- as a resource for teaching (course development).

Assignment 1 asked you concentrate on the student experience, which corresponds to the first and second roles. As the third role implies, teachers are also library users. The tasks that they perform when developing courses are quite different from students' tasks when taking them.

All libraries – physical and digital – are service organizations, based upon the fundamental requirement to serve their users. When examining the applications of digital libraries in education we need to consider the needs of various users and stakeholders:

- students;
- teachers;
- educational authorities, including governments and teaching standards organizations.

Each group has different requirements for the content and organization of educational digital libraries. The present unit adopts the perspective of students and teachers; Units 3.2 and 5.3 examine how metadata standards can support the needs of learners and educational authorities.

What is it about digital libraries that differentiate them from traditional physical libraries? Here are several distinguishing characteristics:

- current content: it is easier to update a digital library with new material;
- content from primary sources: e.g. NASA distributes authentic data that students and researchers can use;
- comprehensive content: both breadth and depth are easier to achieve;
- content can be easily presented in many different formats (images, maps, audio, video, data series, etc.);
- content is more easily accessible: it is easier to distribute new data over computer networks than to update paper copies;
- content can be published: students can easily create their own content and publish it in a digital library.

To this list can be added two further advantages:

- the possibility of resource reuse: teachers can share resources in ways that are not practical with paper-based materials;
- ease of integration: as more student work is digital it becomes easier for students to incorporate authentic content into their own work.

Ease of integration is both an opportunity, enabling richer coursework, and a threat, making plagiarism easier for students. In Module 3 Unit 3.3 we will see that issues of copyright have been anticipated when describing educational content.

Assignment 2

Read "The roles of digital libraries in teaching and learning" by Gary Marchionini and Hermann Maurer. (You can find this article, like the others that you will study in this course, in the accompanying course CD-ROM. If you would prefer to read it electronically, skip forward to the Practical Exercises in Unit 1.3 to learn how to do so, then return to this Assignment.)

From the perspective of a teacher creating lessons:

- What types of resources would be most useful?
- How should usage of digital resources be monitored? Is there more danger of plagiarism when content can be easily copied?
- How do teachers currently locate material for use in lessons?

Material on the Internet can be used in educational applications, but will not be presented in the same controlled environment as material in a digital library. Teachers may appreciate the limits that curated collections set for students because they can:

- specify resources to focus student activity;
- limit access to some resources: for reasons of student age or content type;
- monitor the usage of resources.

Assignment 3

Read "Digital libraries and education: trends and opportunities" by Hans Roes (again, it's in the accompanying course CD-ROM).

- How should content for teachers be organized?
- Would you prefer small lessons of specific content or large subject-oriented collections?

The organization of content for teachers in educational digital libraries depends on the metadata of the elements of the collection. Educational metadata is organised around concepts that are useful for teachers, whereas the metadata of many libraries tries to cater for a wide target audience, such as the general public. A categorisation that supports access by one group of users might not be useful for another group, such as teachers.

Case study: DLESE, http://www.dlese.org/

The Digital Library for Earth Science Education (DLESE) provides:

- access to high–quality collections of educational resources;
- access to **Earth data sets and imagery**, including the tools and interfaces that enable their effective use in educational settings;
- **support services** to help educators and learners effectively create, use, and share educational resources.

Communication networks to facilitate interactions and collaborations across all dimensions of Earth system education.

DLESE resources include electronic materials for both teachers and learners, such as lesson plans, maps, images, data sets, visualizations, assessment activities, curriculum, and online courses.

Material is organized in several ways, e.g.:

- Grade/Level
- Resource Type
- Activity (e.g. Field Trip, computer activity, tutorial)
- Media Type:

  Visual (e.g. map, video, photograph)
  Text (e.g. abstract, bibliography, glossary)
  Audio (e.g. music, oral history)

Here are screenshots showing the various organisational schemes it provides.

Classroom and visual media categories:

Grade level categories:

Text categories:

Audio categories:

Service categories, for communication and collaboration:

Assignment 4

Think about a particular course you have taught or are familiar with. How would you describe:

- the different types of activities that students undertake?
- the different types of media they use?

Now study the categories used in the Digital Library for Earth System Education (DLESE). If you have access to a networked web browser, the categories can be found online at http://www.dlese.org/; otherwise use the screenshots in the accompanying DLESE case study.

Now compare your categories with those used in DLESE.

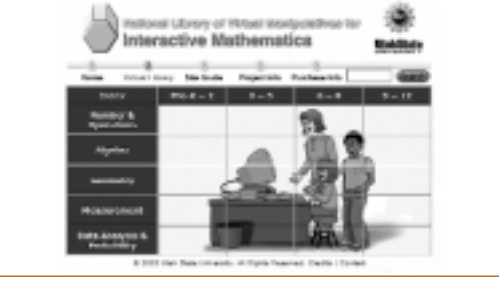
- How many of your categories overlap with those of DLESE?
- Are the DLESE categories specific to Earth Science or are they general enough to be used for any subject?
- Can you think of resources or activities that cannot be described using the categories?

Case study: NLVM, http://matti.usu.edu/nlvm

Digital libraries can contain more than static materials, such as text or images. Interactive software can also be organised and accessed just like any other type of information. The National Library of Virtual Manipulatives for Interactive Mathematics (NLVM) at Utah State University is a collection of software, such as Java applets, that can be used interactively within a web browser to illustrate mathematical concepts, such as fractals, number bases, probability, fractions, and Venn diagrams.

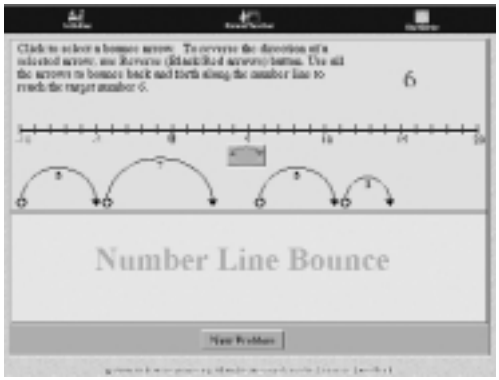
"This is a three–year NSF supported project to develop a library of uniquely interactive, web–based virtual manipulatives or concept tutorials, mostly in the form of Java applets, for mathematics instruction (K–8 emphasis). Ultimately we will make all materials available at several sources on the Internet, creating a national library from which teachers may freely draw to enrich their mathematics classrooms. The materials will also be of importance for the mathematical training of both in–service and pre–service elementary teachers."
(from http://matti.usu.edu/nlvm/nav/projinfo.html)

This screenshot shows the main categorisation of the NLVM:



Beneath these main categories the concepts are simply linked to one instance of virtual manipulative about that concept:

An example of a virtual manipulative on the concept of the number line:



NLVM links the activities to standards documents indicating what is expected of children of different ages. For example, the above “number line” module links to the National Council of Teachers of Mathematics document shown below (at http://standards.nctm.org/document/appendix/alg.htm):

As this example illustrates, the NLVM links the information it contains to particular educational objectives specified by educational authorities. These authorities clearly have an interest – indeed, a responsibility – in ensuring effective educational outcomes for teachers and students alike.

Assignment 5

Compare the DLESE and NLVM systems described in the accompanying case studies.

- DLESE's categorization is much more detailed than NLVM's. Is this necessary?
- Could NLVM benefit from a more detailed categorization? What other access mechanisms might be useful?
- In your own educational context, is it important to link digital libraries of educational material to educational standards? Should these standards be used to organise the library?
- To what extent can global digital libraries be supported by national standards?

Unit closing

Digital libraries present educators with many exciting new opportunities. They can incorporate a huge variety of different content types. While one naturally thinks first of textual documents, upon reflection digital libraries can contain any media type – for example, images, maps, audio, video, even virtual manipulatives. So can conventional libraries, of course, but it is more difficult for them because of physical packaging and different viewing requirements for library users. Digital libraries are viewed on general-purpose computers, which can present all kinds of media in a relatively uniform way. Moreover, they are not restricted to conventional media: digital libraries can include raw data and even interactive software modules, as we saw in the example of the NLVM. Another opportunity lies in the possibility of global access: now teaching material can be shared on a scale that transcends the wildest dreams of a generation ago.

This unit has also shown that digital libraries have different kinds of users and stakeholders. While regional educational authorities are – or should be – interested in conventional school libraries, they are unlikely to use them as a resource. Conventional school libraries are designed primarily for kids, secondarily for teachers, and not at all for educational authorities. But a single digital library can serve the purposes of all three. This broadening user base is reflected in new kinds of structures and classifications, designed to ensure that the same resources can be used appropriately for different purposes.

Of course, all technologies have drawbacks and it is important to recognize them. One that has come to plague many educational situations today is plagiarism. Computer networks make it easy for students to find information relevant to their assignments and incorporate it into their answers. While it is not difficult to include proper attribution, it is even easier, and (students may think) more impressive, to simply leave it out. Essay banks can be organized on a wide, even international, scale, and there are commercial web sites that, for a fee, provide students with plagiarized answers.

Perhaps a more deeply rooted downside is the potential that digital libraries give for increased educational control at the regional, national, and even global level. Most students learn in a local context and relate better when their learning material has a localized flavour that makes it relevant to their own particular situation. Most teachers enjoy the challenge of designing their own courses and preparing their own teaching material. A tendency towards excessive control by central authorities may lead to lower job satisfaction among teachers and a kind of down-skilling of the teaching profession, which would be greatly to the detriment of students, who will always learn best by having at their side a bright, well-motivated teacher.

Up to this point the course has introduced you to the idea of digital libraries and how they can be used in education. Now it is time for some practical experience in the use of one. Please move on to Unit 1.3, where you will meet the Greenstone digital library software.

Sources

We have found three principal sources useful in preparing this unit. The different roles that digital libraries can play in education that we identified at the beginning of the unit are discussed by Masullo and Mack (1996). The characteristics that distinguish digital libraries from traditional ones are identified and discussed by Wallace, Krajcik, and Soloway (1996); the additional possibility of resource reuse is discussed by Mendel (1999).

UNIT 1.3 The Greenstone digital library software

This unit introduces a particular suite of software for digital libraries. Please begin by considering who the users of such software might be and what their various requirements are.

Assignment 1

The people who inhabit a conventional library have two principal roles: reader and librarian. Imagine that (contrary to the caveat at the end of Unit 1.1) your local university library or any other library that you are familiar with has been entirely replaced by a digital library. From your experience of libraries, write a paragraph describing the requirements of the digital library software from the point of view of

- the reader,
- the librarian

in order to allow them to continue the work that they do at present.

Write another paragraph explaining how the role of the librarian might change in this new digital world.

The practical work in this course uses the Greenstone digital library software, which provides a convenient way of organizing information and making it available over the Internet – or on removable media, such as CD-ROM or DVD.

Assignment 2

Read Section 1.4 (pp. 24–28) of the textbook and then answer the following questions in the specific context of the Greenstone digital library software:

- What is a "document"?
- What is a "collection"?
- What is a "library"?
- What software license is Greenstone issued under?

Greenstone is a rapidly evolving system, and since the book was published there have been many additions to it. For example, at that time the interface was available in 12 languages: Arabic, Chinese, Dutch, English, French, German, Hebrew, Italian, Māori, Portuguese, Russian, and Spanish. Today this has been extended to 32, adding Armenian, Bosnian, Catalan, Croatian, Czech, Farsi, Finnish, Galician, Georgian, Hindi, Indonesian, Kannada, Kazakh, Japanese, Latvian, Serbian, Thai, Turkish, Ukrainian, and Vietnamese. This is a powerful indicator of international interest in the software.

Assignment 3

Greenstone is issued under the GNU General Public License,[1] which is officially certified as "Open source software". Read the definition of open source software below and answer the following questions.

Are you allowed:

- to redistribute the Greenstone software freely?
- to sell Greenstone?
- to modify Greenstone in any way?
- to make collections with Greenstone and distribute them freely?
- to make collections with Greenstone and sell access to them?
- to make CD-ROM/DVD collections with Greenstone and sell them?

[1] You will have an opportunity to read the text of this License when you install the Greenstone software in Module 2, Unit 2.2. If you have access to the web you can find it online http:// www.gnu.org.

Open source software

Open source doesn't just mean access to the source code. The distribution terms of open–source software must comply with the following criteria (from http://www.opensource.org):

Free redistribution

The license shall not restrict any party from selling or giving away the software as a component of an aggregate software distribution containing programs from several different sources. The license shall not require a royalty or other fee for such sale.

Source code

The program must include source code and must allow distribution in source code as well as compiled form. Where some form of a product is not distributed with source code, there must be a well–publicized means of obtaining the source code for no more than a reasonable reproduction cost preferably, downloading via the Internet without charge. The source code must be the preferred form in which a programmer would modify the program. Deliberately obfuscated source code is not allowed. Intermediate forms, such as the output of a pre–processor or translator, are not allowed.

Derived works

The license must allow modifications and derived works and must allow them to be distributed under the same terms as the license of the original software.

Integrity of the author's source code

The license may restrict source–code from being distributed in modified form only if the license allows the distribution of "patch files" with the source code for the purpose of modifying the program at build time. The license must explicitly permit distribution of software built from modified source code. The license may require derived works to carry a different name or version number from the original software.

No discrimination against persons or groups

The license must not discriminate against any person or group of persons.

No discrimination against fields of endeavour

The license must not restrict anyone from making use of the program in a specific field of endeavour. For example, it may not restrict the program from being used in a business or from being used for genetic research.

Distribution of license

The rights attached to the program must apply to all to whom the program is redistributed without the need for execution of an additional license by those parties.

License must not be specific to a product

The rights attached to the program must not depend on the program's being part of a particular software distribution. If the program is extracted from that distribution and used or distributed within the terms of the program's license, all parties to whom the program is redistributed should have the same rights as those that are granted in conjunction with the original software distribution.

License must not restrict other software

The license must not place restrictions on other software that is distributed along with the licensed software. For example, the license must not insist that all other programs distributed on the same medium must be open–source software.

License must be technology–neutral

No provision of the license may be predicated on any individual technology or style of interface.

Now let's take a look at a digital library collection built using Greenstone. To do this you use the Greenstone reader's interface, accessed through a web browser. Later, in Module 2, you will build your own collections with Greenstone and for that you will use the Greenstone Librarian Interface (GLI).

First, take out your CD-ROM for the course *Digital Libraries in Education*. Note: this is not the same as the IMARK Information Management Resource Kit CD-ROM, which is also supplied with this course. It contains three pre-packaged Greenstone collections, which can be viewed on any Windows computer (3.1/3.11, 95/98/ME, NT/2000, and XP). Greenstone itself runs on Linux, Unix, and Macintosh OS/10 as well, but it produces CD-ROMs with pre-packaged collections that are designed for Windows only. These collections use only the Greenstone reader's interface, which is why they can be viewed even on primitive Windows computers (3.1/3.11 and 95/98/ME). The Librarian Interface that is used in subsequent modules is restricted to modern versions of Windows (NT/2000 and XP). The reason for this is that this interface, which postdates the original Greenstone system, is written in the Java programming language.

The practical exercise that follows installs these collections on to your computer. Most modules in this course contain several practical exercises. Practical exercises differ from assignments in that they do not require you to answer questions, but take you step by step through a practical procedure. They are a cental part of the course: do *no*t skip them.

Practical exercise: Installing collections

1. Insert your CD–ROM for the course *Digital Libraries in Education* into a Windows computer. If the installation process does not start up straightaway (because the AutoPlay feature is disabled on your computer), navigate to your CD–ROM/DVD drive (normally D:), open the folder *prebuilt*, and double click on *Setup.exe*.

2. During installation you are offered a choice of folder to install in. We recommend the default, which is C:\GSDL.

3. You are also presented with the option to run Greenstone from the CD–ROM or to copy the entire CD–ROM. We recommend the latter: please check the box that says *Install all collection files*. It will take at least a couple of minutes to copy the files across.

4. Finally, the installer offers to install the Netscape browser for you. Do *not* request this except in the unlikely event that you do not already have a web browser on your computer.

This installation procedure is far more complicated to describe than to actually do! It should only take a few minutes.

5. To run *Greenstone*, open the Windows Start menu, Programs, and select Greenstone, then the submenu item *Digital Libraries in Education*, then <***Enter Library***>.

In this unit we study two of the three collections on the CD-ROM: *Digital Libraries in Education* and *How to Build a Digital Library*. We look at the third collection, called *IEEE LOM*, in Unit 3.3 (LOM stands for "Learning object metadata").

Practical exercise: The course collection

First examine the Digital Libraries in Education *collection*.

1. Click the collection's icon. This takes you to the collection's home page, often called the "about" page.

The home page contains an access bar with buttons called search, contents, authors a–z, modules, and acronyms. *This access bar is the key to finding information in any Greenstone collection*.

2. *Click <**authors a–z**>. A list of bookshelf icons appears. Click the one called* Marchionini, G. *to see the two course readings by Gary Marchionini*.

3. *One of these items is a PDF file and the other is an HTML file. Click them both in turn to open up the documents*.

4. *Click the <**contents**> button in the access bar. This shows two bookshelves, one for this Study Guide and the other for the Course Readings. Choose one and look at what it contains*.

5. *Clicking a bookshelf that is open closes it. Close the bookshelf you have just opened and then choose the other one and examine its contents*.

6. *Click <**acronyms**> in the access bar and find the meaning of the acronym "LOM"*.

7. *Click <**search**> and search for the word "LOM". Check out the difference between searching text and searching titles (use the pull–down box on the search page)*.

8. *Click the collection icon* **Digital Libraries in Education** *at the top left. This takes you back to the collection's about page*.

Beneath the access bar on the collection's about page is a search box (just the same as the one that appears on the search page), a description of the collection under the heading "About this collection", and instructions on how to find information in this collection.

Above the access bar is the collection's icon, saying Digital Libraries in Education*, and, on the right, information box* saying about*, above which are three buttons,* home, help*, and* preferences.

9. *Click <**home**>. This returns you to the Greenstone home page*.

10. *Return to the collection (by clicking its icon) and click <**help**>. This gives more information about how to access the collection*.

11. *Click <**preferences**>. This takes you to a page where you can change some of the settings*.

12. *Now explore the collection by navigating freely around it. Click liberally: all images that appear on the screen are clickable. If you hold the mouse stationary over an image, most browsers will soon pop up a brief "mouse–over" message that tells you what will happen if you click. Experiment! Choose common words like "the" or "and" to search for – that should evoke some response, and nothing will break. (Note: unlike many search systems, Greenstone indexes all words, including these ones.)*

Assignment 4

Read the Help page for the Digital Libraries in Education collection and then answer these questions.

- What does this collection contain?
- Name five ways to navigate to a target document in this collection.

- How many documents in the collection are written by Erik Duval?
- Compare the number of times the words "he" and "she" appear in the collection.
- How many times does the word "metadata" appear in titles? In the text itself?
- What's the difference between a some and an all search?
- What does "MODS" stand for?
- How do you switch the interface from English to Russian? Does it stay in Russian when you go to the Greenstone home page?
- Find a search term that yields different results depending on whether you have ignore word endings or whole word must match set on the Preferences page.
- What's the difference between Graphical and Textual interface format (on the Preferences page)?

The second pre-packaged collection on the course CD-ROM that we look at here is called *How to build a digital library*. As the name implies, this is based on the book that you are studying. However, it does not present the contents in readable format – for the book is published by a commercial publisher who would not permit free electronic redistribution. Instead, the collection contains an index of the book's *sentences*. You can find and read individual sentences, but not paragraphs, sections, or chapters. This is accomplished by treating each individual sentence as a Greenstone "document" in its own right. Documents are independent entities, so are all these sentences. This is rather an unusual application that illustrates the flexibility of Greenstone.

When you reach a sentence (i.e. document) in this collection, it is accompanied by its page number in the published version of the textbook. The search function in the collection acts as a rather sophisticated index to the book, an interactive index that accommodates full-text search. As well as this index, the collection includes:

- online Appendices for the book;
- colour versions of all figures;
- textual versions of the XML figures;
- front matter and Chapter 1 samplers;
- a hierarchical phrase browser (accessed by a button called *phrases*);
- a list of acronyms.

The last two are Greenstone facilities that are easy to provide in any collection. The system analyses a specified set of text, in this case all the figure captions, and extracts a hierarchy of phrases that the phrase browser presents using a Java applet. Try, for example, typing *XML* into the hierarchical phrase browser to see the four contexts in which that word occurs in captions. The system also analyses the full text of all documents and extracts any acronyms, with their definitions.

Assignment 5

Use the *How to Build a Digital Library Collection* to answer these questions.

- How many sentences contain the word *education*?
- What story from the *School Journal* collection is featured in the book?
- How many acronyms used in the book begin with the word *standard*?
- What does *tapu* mean?
- How many times does the word *library* appear? The word *libraries*?
- How many times does *Library* appear with an initial capital letter?
- How many times does some derivative of the word *form* appear?
- Name an English poem that was probably written in about 1000 A.D.
- Who is Alan Kay?
- On what page is the first mention of some aspect of Chinese culture?

Most of these questions would be rather difficult to answer from the printed book.

The Information Management Resource Kit Digitisation and Digital Libraries contains several practical lessons on the Greenstone digital library software.

Assignment 6

Work through *Lesson 6.1 Greenstone tutorial* of the *Information Management Resource Kit Digitisation and Digital Libraries*. This will teach you to:

- recognize the key features of the software;
- identify the installation requirements and options;
- identify which interface features can be provided to end users;
- identify the three different collection-building approaches it provides.

Answer the questions at the end of the lesson.

Unit closing

You began this unit by thinking about what would be needed to support the two principal kinds of people that inhabit conventional libraries: readers and librarians. Digital library software must of necessity be usable by readers with practically no training – digital libraries would be a complete failure if you had to study a book in order to be able to learn how to use them. It should be universally accessible: available freely; capable of running on any computer, including primitive ones; usable without necessarily being connected to the Internet; able to present material in any language; have interfaces in many different languages; and so on. Few digital library systems satisfy all these requirements. In particular, most assume that the Internet is ubiquitous. Dispensing with this assumption has severe technological implications because now the digital library *server* – the search system and so on – must run on primitive computers in order to be universally accessible.

This unit has introduced the Greenstone software and discussed the extent to which it meets the requirements of universal accessibility. The GNU General Public License under which it is issued allows great freedom for both commercial and non-commercial use. It is important that you understand your freedoms. Just to make it completely clear, the answer to all the questions in Assignment 3 above are *yes*. You don't need to ask permission to redistribute Greenstone, sell it, modify it in any way, make collections with it, distribute them freely or sell access to them, make collections on removable media and give them away or sell them. In fact, the only restriction that the license places on you is that if you do modify the software, you are obliged to make the modified version freely available, at no charge (you may also sell it if you wish). This means that you can't incorporate components of Greenstone into your own proprietary commercial system.

You have also experienced the reader's interface. This provides a rich means of finding information in a collection by searching and browsing. It is not hard to use: online help is provided (in all the interface languages) and there is no need for any further instructions or documentation.

The specific access mechanisms that are incorporated into any particular Greenstone collection are determined by the person who designed and created the collection. In the next module you will meet the plethora of different document formats that you might want to incorporate into your collection. You will also learn how to build collections using the Greenstone Librarian Interface.

MODULE 2 DOCUMENT REPRESENTATION

Goal

To gain an appreciation for the plethora of different document formats and their strengths and weaknesses, including an understanding of how characters in the world's languages can be represented, and to experience building a digital library collection of documents of various formats (including Word and PDF files).

Objectives

Upon completion of Module 2, you will be able to:

- give examples of challenges that occur when representing various non-European languages on a computer;
- describe how Unicode is used to represent different character sets;
- distinguish between UTF-8, UTF-16, and UTF-32 representations of Unicode;
- distinguish between page description languages and word-processor document formats;
- relate the history of the PostScript page description language and how it evolved into PDF;
- choose appropriate formats for the documents in your digital library;
- discuss the drawbacks of using a proprietary format like Microsoft Word;
- install the Greenstone software and re-install future versions without affecting the collections you have built;
- build collections of Word, RTF, PDF, and PostScript documents;
- add metadata to a Greenstone collection;
- design a Greenstone collection, including search indexes, browsing classifiers, and collection icons;
- explain the difference between extracted metadata and user-assigned metadata;
- locate the collections you have built in your computer's file structure.

Introduction to Module 2

In Module 1 you learned what is meant by a "digital library" and saw how they are used in education. You also used a digital library collection for finding information.

Module 2 gets down to the nitty-gritty details of how digital libraries are constructed. Documents are the digital library's raw material and characters are the raw material from which documents are composed. Although we aspire to build digital libraries whose content glitters and sparkles and reaches lofty intellectual levels, to set them on a solid footing we must start at the level of a young child learning the alphabet. For alphabets are by no means simple – not when they must accommodate the amazing diversity of the world's languages. And the documents that are built from them fulfil such different purposes. For each purpose there is a choice of document format: lowest-common-denominator formats for presentation over the web, page description languages that portray finished documents, word-processor representations designed for convenient and rapid editing, document processing formats that work off-line to give great control over how text appears on the page. Not all documents are textual: you need to learn about image, audio, video, and multimedia formats, too.

The other kind of detail is how to actually build digital library collections. In this module you will install digital library software and learn how to use it to gather together documents, assign descriptive metadata, such as titles and authors to them, design the collection by determining its appearance and the access facilities it will support, build the indexes and browsing structures, preview the newly created collection to check the design, and serve it to users, either over the web or by writing it to CD-ROM or DVD. Again, lofty visions translate into a lot of hard, low-level work.

The module is divided into two units, corresponding roughly to the preceding two paragraphs. The first looks at how documents are represented in digital libraries. The second shows how to use the Greenstone digital library software to build collections of documents. Both units focus on these four document formats: Microsoft Word, RTF (Rich Text Format), PDF (Portable Document Format), and PostScript. You learn about them in Unit 2.1 and build collections from them in Unit 2.2. Unit 2.2 also covers installing the Greenstone software, and performing various maintenance operations.

Module 2 Readings

Unit 2.1. Documents: the raw material

Reading 1 The textbook *How to build a digital library*: Chapter 4 up to and including Section 4.4.

Purpose This reading first describes Unicode, a character encoding standard for all the world's languages, and then goes on to describe page description languages and word–processor formats. As you read this, relate it to your own experience of working with different languages and different document formats. This unit focuses on the Word, RTF, PDF, and PostScript formats.

Reading 2 CD–ROM IMARK: Unit 2 – Electronic documents and formats: Lessons 2.2, 2.6, and 2.7.

Purpose Lesson 2.7 includes valuable practical information on the PDF page description language, which is introduced in Reading 1. Lessons 2.2 and 2.6 are designated as "enrichment items" that are not essential to this unit but provide a different perspective on the material in it; we recommend that you study them if you possibly can.

Unit 2.2. Building a digital library collection

Reading 1 CD–ROM IMARK: Lesson 6.2 – Using GLI: Gathering and enriching documents.

Reading 2 CD–ROM IMARK: Lesson 6.3 – Using GLI: Designing and creating a collection.

Purpose These two self–study units lead you through an extended worked example involving installing the Greenstone software and designing and building a collection using the Greenstone Librarian Interface (GLI) from documents in Word, RTF, PDF, and PostScript formats.

UNIT 2.1 Documents: the raw material

In this unit you will learn how documents, the building blocks of digital libraries, are represented in the computer. There is a host of different kinds of document. Before rolling up our sleeves and getting down to detail, please take some time to reflect on this diversity.

Assignment 1

Think about all the different kinds of material that might go into an educational digital library.

- List as many different formats for computer storage of textual documents as you can.

Consider the individual characters that make up textual documents.

- Apart from the characters found on an English-language typewriter (lower- and upper-case letters, digits, punctuation, and the various additional symbols that occur on a typewriter keyboard), what other character classes are represented in your documents?
- If your educational situation involves languages other than English, write a brief paragraph describing the problems that you have encountered with computer representation of text in this language.

Proceed through this unit by working your way through Chapter 4 of the textbook, which is about different ways of representing documents in machine-readable form.

We first study Unicode, which is a way of representing all the characters used in the world's languages. Because computer technology originated in Western Europe and the United States, all early development was focused on European languages – mostly English. For international use it is necessary to deal with other character sets and systems, which present challenges of surprising complexity – surprising, at least, for Anglophones.

Issues of internationalization are not confined to the representation of individual characters. When computers perform full-text indexing of a document collection, they usually reduce words to their root form or at least offer users an option to do so. This makes queries insensitive to different variants of the words in it, which is often a good idea. Of course, such reduction, called *stemming*, depends on the language in question. Moreover, many oriental languages do not routinely insert spaces between words as European languages do, and this seriously affects full-text indexing.

If you are dealing with international collections, you need to be aware of these issues. Today's information retrieval systems invariably embody stemming methods for English only, although well-designed systems modularise this function so that programmers can extend them for other languages. And today's systems rarely incorporate word segmentation techniques for oriental languages.

Assignment 2

Read the beginning of Chapter 4 and Sections 4.1 and 4.2 (pp. 131–163) of the textbook and answer these questions:

- What is the problem that Unicode aims to solve?
- What is meant by "round-trip compatibility"?
- What is the difference between a composite character and a combining character?
- Why does this difference complicate searching?
- Give an example of a full-text query and a usage context for it where it is important to stem the terms.
- Given an example of a query and context in which it is important not to stem the terms.
- For a local language that you are familiar with, list the shortcomings of using an information retrieval system designed for English only.

Enrichment exercise 1

IMARK Lesson 2.2 gives another perspective on the problems of character encoding: it focuses on how to create multilingual web pages. Digital library software normally takes care of these details for you. Nevertheless, it is useful to understand what your system is doing underneath, and we recommend that you work through this Lesson to enhance your understanding of character coding issues.

Now let's turn our attention to the representation of documents in popular computer formats. We distinguish between page description languages, which are designed for representing finished, typeset documents, and word-processing formats, which are designed for representing works in progress but are often used for completed documents, too. The essential difference is that the former class preserve the document's page format, including the pagination and layout of each page, positions of line breaks, and so on, whereas in the latter these features may change depending on what printer is used, the paper size, default margin settings, and so on.

The textbook gives a detailed account of PostScript, the original page description language, and explains how it was further developed into PDF, the Portable Document Format. In much practical digital library work it is important to know about the PostScript format, because many older documents are still presented in this form. However, if you are making a page version of a document today you will probably use PDF – although on some computer platforms extra software must be purchased to do this, whereas PostScript conversion is provided anyway because PostScript is still the language of choice when computers communicate with printers. IMARK Lesson 2.7 gives a detailed practical account of the PDF page description language and how to use it, which nicely complements the material in your textbook. You should study both of these sources.

Assignment 3

Read Section 4.3 (pp. 163–184) of the textbook and work through IMARK Lesson 2.7. Then answer these questions:

- Are older machine-readable documents more likely to be presented in PostScript or PDF?
- Why is it difficult to scan a PostScript file and extract plain text from it?
- What problems with PostScript led to the development of the PDF format?
- In what respects does PDF go beyond the original notion of a "page description language"?
- Which is better, PostScript or PDF, for (a) online use, (b) international use?

As mentioned above, word-processing formats are designed for representing works in progress, but there is a trend towards using them for communicating finished documents, too. This is unfortunate because despite the widespread adoption of particular word processing systems, they are commercial software and cannot be read on all computing platforms. Despite the convenience for you, if you send a document in Microsoft Word format the recipient may not be able to read it. It is a proprietary secret format, and if you want to read the same document again next year, or next decade, you are implicitly relying on a particular manufacturer to ensure that future versions of the software can read old files. In fact they can't: today's version of Microsoft Word cannot decode many older Word files, and these documents are just as inaccessible as if you encrypted them and threw away the key.

In today's world, we recommend PDF for communicating and storing finished documents. Do not use word-processor formats unless your readers are co-authors who want to edit the document. If you are building a digital library that must include Microsoft Word documents, convert them to PDF (or HTML) first. This gives you better control over how documents look to the reader, provides greater portability, ensures universal access, and improves your chances of being able to recover them in the future.

Nevertheless, people do use word-processor formats and it is helpful if digital library software can deal with such documents, extracting their text for full-text indexing, and converting them to a more universal form for readers without the necessary word-processing software. The textbook introduces Microsoft Word native format (briefly,

since the details are secret), RTF or Rich Text Format, which is designed to allow word-processor documents to be transferred between different applications, and LaTeX, which is often used for mathematical and scientific documents.

Assignment 4

Read Section 4.4 (pp. 184–194) of the textbook and answer these questions:

- Which would you expect to be larger, a Word file or an RTF file for the same document?
- Does either Word or RTF make it possible to add new features to the format without invalidating older documents? How?
- List several ways of including graphical illustrations in RTF documents.
- How easy is it to extract the plain text from documents if they are stored in (a) RTF, (b) Word, and (c) LaTeX?
- Identify three ways to proceed if you have to include Word documents saved with the "fast save" option in a digital library? Discuss their pros and cons.

Enrichment exercise 2

IMARK Lesson 2.6 discusses word-processing software using the example of Microsoft Word. However, it is oriented towards teaching you to use Word's interactive facilities to produce documents that not only look good but are nicely structured internally. The benefit of using well-structured documents in a digital library is that more presentation and browsing features can (potentially) be automated. In order to derive maximum benefit from this module, you should first read Lesson 2.1 for an introduction to what "markup" means and why it is important.

Unit closing

This unit has introduced many low-level issues of character and document representation. In an ideal world people whose goal is to build and use digital libraries would not have to worry about such minutiae. But the world is not ideal.

Some of these issues have their roots in belated internationalization. If the development of computers had been a truly international endeavour, they would have been resolved long ago. In fact, computer technology developed under the assumption of a typographically rather simple language, English, based on a rather simple character set, the Roman alphabet, using rather a simple kind of document, typewritten text. Only gradually is the realization dawning that real, international documents are often far more complex than this.

Other issues have to do with standardization. Although Unicode is an international standard, page description languages are informal "standards" that have not been adopted as accepted international standards. And this is probably a good thing. Since PostScript was first introduced in 1985, there has been more or less continual development: PostScript Level 1, 2, 3; then PDF 1.0, announced in 1992 and overlapping with later PostScript enhancements, developing through to PDF 1.5, released in mid 2003. As ideas evolve about what page description languages should do, so does the software. An annoying problem when dealing with these documents is the plethora of different versions and extensions. In PostScript's early days, documents could always be processed – even ones from faraway lands and foreign computers. Unfortunately, as PostScript evolved it became more and more common to encounter documents that were useless because local software could not interpret them. The introduction of PDF was a relief. – but the same problem has already begun to occur with it, albeit more gradually.

Still others have to do with commercialization. PostScript and PDF are open standards: complete information is publicly available on how they represent documents. Yet they are commercial and cost has been a bar to the widespread adoption of PDF because companies charge for software that creates PDF files. (In the early days even PDF readers were not free.) The native Word format is not open and no one – not even Microsoft software – can reliably decode legacy Word documents. The underlying reason is easy to understand: profitability depends on repeat sales of ever more

enhanced versions of the software, whereas backwards compatibility, though technically challenging, pays nothing. Details of the RTF and LaTeX formats are publicly available, and a great deal of open source software is available for processing LaTeX.

Our intention in this Unit has been to prepare you for some of the thorny technical snags that may occur when you build digital libraries in practice. Although solving the problems may require specialized technical expertise, even programming skills, we believe that a better understanding of what might go awry will serve you in good stead when locating problems and finding ways around them, or communicating with technical experts about potential solutions.

You can sometimes get around problems by resourceful use of existing software. For example, Microsoft programs have a better chance of being able to deal with difficult Word documents than open-source digital library software, whose authors are not privy to the internal details of the document format. If you encounter difficulties with a document and have the Word program, you should try reading it into Word and writing it out in a different format – perhaps PDF. If you find that the PDF version causes problems too, you should try writing it out in HTML instead. Although the document quality will be lower, you will be able to get it into your digital library.

Now you are ready to try building digital library collections from documents in the formats you have learned about here. To do so, proceed to Unit 2.2.

UNIT 2.2 Building a digital library collection

This unit consists entirely of practical work. In Unit 1.3 you installed pre-packaged collections and gained experience with the Greenstone reader's interface. Now you will install the full Greenstone software – which is just as simple to do but takes a little longer to copy files across – and learn how to use the Librarian Interface to build collections. Whereas the reader's interface is easy to use and needs no instruction, the Librarian Interface is more complex because it performs more advanced tasks.

Most of this unit is based on the IMARK Module, which contains an excellent introduction to Greenstone. In this unit you will work through IMARK Lessons 6.2 and 6.3 on the use of GLI, the Greenstone Librarian Interface.

Assignment 1

Work through IMARK Lesson 6.2 and perform all the associated practical work. This involves:

- installing Greenstone from the IMARK CD-ROM;
- starting the Librarian Interface;
- gathering documents together;
- enriching documents by assigning metadata.

You will find Greenstone on the IMARK CD-ROM in the folder software_tools → Greenstone. Double click on setup.exe to start the installation process. There are more files to install this time and it will take about twice as long as before to install the necessary files.

Note: We assume Greenstone has not been installed on your computer before. If it has, it is important to completely remove any old version before installing a new one. (However, you do not need to remove the pre-packaged collections that you installed in Unit 1.3.) Assignment 4 below describes how to do this.

IMARK Lesson 6.2 describes the installation process. However, it does not take you through one installation step that is crucial for our purposes. Once Greenstone has been installed, you will be asked whether you want to install ImageMagick: say "Yes." To install this program, you must have Windows "Administrator" privileges.[2] The remaining steps are straightforward and as before we recommend the default settings. Here is what you need to do.

1. "This will install ImageMagick 5.5.7 Q8. Do you wish to continue?" **Yes**.
2. "Welcome to the ImageMagick Setup Wizard". Click <**Next**>.
3. "Information: Please read the following ..." Click <**Next**>.
4. "Select Destination Directory ..." Leave at default and click <**Next**>.
5. "Select Start Menu Folder ..." Leave at default and click <**Next**>.
6. "Select Additional Tasks ..." Leave at default and click <**Next**>.
7. "Ready to Install". Click <**Install**>.

Files are copied across.

8. "You have now installed ..." Click <**Next**>.
9. "Setup has finished ...". Deselect "View index.html" and click <**Finish**>.

Note: the version of Greenstone that you have just installed is accessed through Greenstone Digital Library under the Start → Programs menu. The Digital Libraries in Education collections that you installed in Unit 1.3 are on the same menu under Greenstone.

[2] If you do not have Windows Administrator privileges, the ImageMagick installer will give a cryptic error complaining that it failed to set a particular Windows registry value. If this happens you can continue your work with Greenstone, but you will not be able to build collections of images.

Now answer these questions:

- You now have two separate Greenstone installations on your computer, this one and the one that you installed in Unit 1.3. How do you find each one from the Windows Start menu?
- What five steps are involved in building collections using the Librarian Interface?
- How many documents are there in the collection that you are building?
- Which document formats described in the previous unit are represented in the collection?
- Which of these documents opens up when you double-click it in the Librarian Interface?
- What metadata set is used for the collection, and how many metadata elements are in it?
- Explore the Librarian Interface Help system using the button at the top right. How many subsections are there in the section entitled "Enriching the collection with metadata"?

Assignment 1 has taken you halfway through the process of building a digital library collection. Assignment 2 completes the operation.

Assignment 2

Work through IMARK Lesson 6.3 and perform all the associated practical work. This involves:

- deciding what searching and browsing facilities the collection should offer the reader;
- designing the collection;
- building it;
- previewing it;
- publishing it on CD-ROM.

Now answer these questions:

- Which component of Greenstone determines the document formats that are included in a collection?
- Which part of the Design panel do you use to set up an image for the collection?
- What are the two different search types that Greenstone offers (when Advanced Search is enabled in the GLI)?
- How many search indexes are there in your collection? Where do they appear in the reader's interface? What are their names?
- How many browsing classifiers are there, and where do they appear in the reader's interface?
- How long does it take your computer to build this collection? A rough measure of speed is time per Mbyte of source material: calculate this.
- How much space is used on the CD-ROM containing this collection?
- How many documents appear in the *titles a–z* list?
- How many documents are returned if you search for a common word like *computer*?
- For how many papers in the collection is Sally Jo Cunningham an author or co-author?
- Does the collection contain any duplicate documents with different formats?

This collection contains a mixture of Word, RTF, PDF, and PostScript documents. In Assignment 2, you might have had difficulty figuring out what happened to some of them. You dragged a total of 12 documents into the collection and, if all went well, they all appear in the list of search results for a common query like *computer*. (You can tell if all went well by reading the output of the *Build* phase, which states how many documents were included in the collection.) But how many documents appear in the *titles a–z* list? It may be fewer, because since the list is based on *dc.Title* (that is, Dublin Core *Title* metadata) only documents that have a value for that metadata element will appear. Perhaps you did not assign metadata to all 12 documents. For example, in the *Enrich* phase you may not have been able to view the contents of all the file types – this depends on the software on your computer and how it is set up.

Greenstone has converted all the documents to HTML, and in this collection both the original source document (Word, RTF, PDF, or PostScript) *and* the HTML version are shown. Choose a few documents and compare the two versions. You will see that the HTML version is of lower quality – for example, the formatting is slightly awry and there are no longer any page breaks – but it does contain the correct textual content. It is this text that is indexed for full-text search.

If you were unable to fill in all the metadata because your computer couldn't show, say, the PostScript files, you can do so now by searching your collection for a common word (like *computer*) and finding the title, author, etc. in the HTML version of the file.

Next, let's see how Greenstone can build a simpler kind of collection from these same documents. Greenstone itself can extract metadata from document files, though of course it does not do such a good job as a person. Using this extracted metadata, it is easy to build a rudimentary collection very quickly. Greenstone incorporates sensible defaults to ensure that something sensible results even if no attention at all is paid to the collection design process. Before you repeat the collection-building procedure, you will first learn how to set up a shortcut in the Greenstone Librarian Interface to the source files, for convenience.

Assignment 3

To set up a shortcut to the source files, return to the Gather panel and navigate to the folder in your local file space that contains the original Word, RTF, PDF, and PostScript files used in Assignment 1. Select this folder and then right-click it. Follow the instructions to set up a shortcut. Close all the folders in the file tree and you will see the shortcut to your source files.

Repeat the procedure process described in IMARK Lessons 6.2 and 6.3, using a different name for the collection. Omit all the steps involving the Enrich and Design panels – in other words, just drag in the same 12 files and go straight to the Create panel. Now,

- How many search indexes are there in this new collection, and what are their names?
- How many browsing classifiers are there, and what are they?
- What appears in place of the image that you set up for the collection in Assignment 2?
- How many documents appear in the *titles a–z* list?

Since you haven't typed any metadata in, what is the *titles a–z* list based on? Where do these titles come from? Having built the collection, choose a file in the Enrich panel and scroll to the end of its metadata list. Here you see metadata values preceded by "ex.", and ex.Title gives the metadata value used for the *titles a–z* classifier. The ex.Source metadata element is always set to the document's filename, so the filenames classifier always shows all the documents in the collection.

Greenstone extracts this metadata from the document files. Apart from Title, this metadata includes the Encoding (e.g. iso_8859_1), file format (e.g. PS for PostScript), file size (in bytes), Greenstone's internal document identifier, the language in which the document is written (e.g. "en" for English), and the name of the plug-in that processed it.

- Automatic metadata extraction is not completely accurate (and never can be). Identify some errors in the extracted titles. What method do you think Greenstone uses to extract the titles?
- The two collections you have built differ in the metadata they use for the browsing classifiers. Where in the Librarian Interface is this information specified?

Greenstone extracts metadata even when you specify the metadata explicitly in the Enrich phase. Reopen the collection you built in Assignments 1 and 2 and find the extracted metadata using the Enrich panel. Although the metadata is extracted, it is not used in this collection.

Assignment 3 illustrates two high-level messages that it is worth stating explicitly. First, manually assigned metadata is (or should be) far superior to automatically extracted metadata. Of course, it is far more expensive to produce. Second, Greenstone incorporates sensible defaults. If you do no more than drag in a few files and click *Build*, Greenstone does its best to produce something acceptable. This policy pervades the whole design of Greenstone. Of course, better results can be achieved by doing more work in the *Enrich* and *Design* phases.

The next exercise is about how to manage your Greenstone installation. As noted in Assignment 1, your computer now has two distinct Greenstone installations: you installed one in Assignment 1 of this unit and the other in Unit 1.3. Assuming that you installed the software in the default place that was suggested by the installer, the Greenstone that you

installed in Assignment 1 will be in *C:\Program Files\gsdl*, and the Greenstone that you installed in Unit 1.3 will be in *C:\GSDL*. We also assume that in Unit 1.3 you followed our recommendation to copy the entire CD-ROM by checking the box that says *Install all collection files*. (If you did not do this, we ask you to reinstall your CD-ROM for the course *Digital Libraries in Education* as instructed in Unit 1.3.)

In this practical exercise you will update the Greenstone software to a more recent version, and combine these two Greenstone installations by moving the prebuilt collections you installed in Unit 1.3 collections in the new installation and deleting the other one. Through these exercises you will learn where the Greenstone software resides on your computer, where it keeps your collections, and how to keep up to date with the latest releases of the software.

Practical exercise: Updating your installation

Before starting this exercise, ensure that your computer is not running GLI or the Greenstone local library server. Normally quitting GLI is enough to also quit the server.

The first step is to remove the Greenstone installation that you installed in Unit 1.3, which is in *C:\GSDL*. First we will move the three collections it contains – the *Digital Libraries in Education* collection, the *How to build a digital library* collection, and the *LOM Demonstration* collection, to the Windows desktop to prevent them from being deleted as well. Greenstone keeps each collection in an individual subfolder of the collect folder. Move the entire folder C:\GSDL\collect to your Windows desktop. Note that it contains three subfolders, one for each collection.

Now, from the Windows *Start* menu, go to the Greenstone menu item and select <**Uninstall**> to remove the Greenstone installation that you installed in Unit 1.3. This is a mini version of Greenstone that contains the Reader's interface along with some prebuilt collections (three, in this case): it is provided on any CD–ROMs you make with Greenstone. You cannot build collections with this mini version of Greenstone.

Next, we ask you to update your Greenstone software. The version that you installed from the *IMARK Information Management Resource Kit* CD–ROM is Version 2.51. Your *Digital Libraries in Education* CD–ROM also contains a separate copy of the complete software, distinct from the mini version that you have just removed, which is more Version 2.60. That is what we will use from now on.

Note: before you install any new version of Greenstone you must completely remove the existing version.

First, make sure you are not running Greenstone. Then remove the old version by going to the Windows Control Panel (from the *Settings* item on the *Start* menu). Click *Add or Remove Programs*, select *Greenstone Digital Library Software*, and *Remove* it. (To do this you may need to have Windows "Administrator" privileges.) At the end of this procedure you will be asked whether you would like all your Greenstone collections to be removed: in this case say *Yes*, because you have no valuable collections. Note that the installer will only remove files that it installed in the first place and will not remove any files that you might have created afterwards. You should clean up your system by checking for the folder *C:\Program Files\gsdl*, which is where Greenstone was installed, and removing it completely if it exists.

Now install the version of Greenstone on your *Digital Libraries in Education* CD–ROM. First, insert the CD into the drive and cancel the auto–installation procedure if it starts automatically. Open the CD–ROM by right–clicking on the appropriate drive (usually D:) of My Computer and selecting <**Open**> from the menu. Of the files that appear, invoke the one called *setup.exe* by double–clicking it. Then follow the installation instructions, choosing the default settings. You do not have to install ImageMagick again.

You will notice some superficial changes in the installation procedure; that is because Version 2.60 uses a differ–ent installer program. There is another important difference that you should be aware of: the new version of Greenstone has been installed in the folder *Program Files\greenstone*, whereas Version 2.51 was placed in the fold–er *Program Files\gsdl* (these are both default locations that you could have changed during installation). This is a new convention that future Greenstone versions will follow too. On this occasion, if you had wanted to save exist–ing collections you would have to explicitly move the contents of your collect folder from the old place to the new one. Future Greenstone versions will always be installed in the new place, *Program Files\greenstone*, so this prob–lem will not happen again.

Now invoke your updated Greenstone installation by going to the *Greenstone Digital Library Software* item on the Windows *Start* menu and selecting *Greenstone Digital Library* – that is, we are starting up the Reader's interface, not the Librarian interface. You will see that at the top of the Greenstone home page there is just one collection, *Greenstone Demo*, in addition to the documented example collections further down the page.

But there are three more collections in the collect folder that you put on your Windows desktop at the beginning of this assignment. Move the three subfolders inside this (they are called *course, howto*, and *lomdemo*) into the collect folder of your Greenstone installation, that is, into C:\Program Files\greenstone\collect

Now quit Greenstone (you can do this by quitting the web browser) and re–invoke it in the same way. These three collections will have appeared on the Greenstone home page.

Finally, invoke the Librarian Interface from the Windows *Start* menu and notice that an additional tab called "Download" appears to the left of the other tabs. You will learn how to use this new facility in Unit 3.1.

The interface to the basic Greenstone system comes in four languages: English, French, Spanish, and Russian. However, many more language interfaces are available. If you wish to operate your Greenstone interface in another language, you should install the Greenstone Language Pack, which is presented on your course CD–ROM. To do this, go to the CD–ROM's *Greenstone Language Pack* folder and double–click on the setup.exe icon.

At this stage, we hope you will agree that it is easy to use Greenstone to build digital library collections of documents of various different formats. But there are pitfalls that were foreshadowed in our discussion of these formats in Unit 2.1. Because of the complexity of the formats – and in particular because of the obscurity of the native Word format – things sometimes don't work.

Greenstone uses third-party open source software to convert documents of different formats to a standard internal representation based on HTML. Sometimes this software isn't capable of converting certain files. For example, as you learned in Assignment 3 (in IMARK Lesson 2.7), a PDF document can be made up of a sequence of scanned images – pictures of text – rather than the text itself. Although it may look like ordinary text to you, to extract it in machine-readable terms requires optical character recognition (OCR), an operation that is more or less complex depending on the quality of the image, the kind of text, and the fonts used – and does not always produce perfect results. The third-party software used by Greenstone's plug-in for PDF documents doesn't attempt this. Thus although the PDF will be visible, the text in it will not be indexed at all. A second example is that text is sometimes extracted incorrectly from Word files which were saved with Word's "fast save" option turned on – although most of the text is there, it may be accompanied by extraneous text that has been deleted but not actually removed from the "fast save" file. A third example is that conversion may fail completely on certain Word files, typically very old ones or ones produced by older Macintosh systems. Even Microsoft Word cannot read certain files that it produced a long time ago.

Such problems are quite rare and we hope you will not encounter them. But you might, and the next assignment prepares you for them. Forewarned is forearmed.

Assignment 4

Build a fresh Greenstone collection from the two files in the folder called sample_files\difficult_documents on your course CD-ROM (the one marked Digital Libraries in Education). These files are called No extractable text.pdf and Weird characters.pdf – their names hint at the problems they will cause. Use the default collection configuration: that is, simply gather the files into a new collection, and build it.

Now preview the collection.

- How many documents appear in the *titles a–z* list?
- How many documents appear in the filenames list?
- What's strange about the document that appears?
- Were any warnings shown while Greenstone was building the collection?
- What kind of PDF file is the document that was omitted from the collection?

The titles and filenames lists show only one of the documents. During the building process this message appeared: "One document was processed and included in the collection; one was rejected."

The Librarian Interface can operate in different modes. So far, you have been using the default mode, called "Librarian". Use the Preferences item on the File menu to switch to Expert mode and then build the collection again. The Create panel looks different in Expert mode because it gives more options: locate the Build Collection button, near the bottom of the window, and click it. Now a message appears saying that the file could not be processed, and why.

- What are Greenstone's other modes, apart from Librarian and Expert?

The Chinese-character anomaly you encountered in Assignment 4 is a mysterious problem that reflects some kind of Windows character-code incompatibility in that particular document. In fact, the problem only appears on Windows computers: on other systems the document behaves perfectly normally. The solution to this particular problem is obvious: define *Title* metadata manually for this file. The document that was rejected in Assignment 4 is a PDF file that comprises a sequence of images: it contains no machine-readable text. Greenstone cannot handle this kind of PDF file. One solution would be to create a machine-readable version of the document by performing optical character recognition (OCR) on it; another is to convert the PDF file into a sequence of image files and process them using an image plug-in (you will build a collection out of a sequence of page images in Unit 5.3).

Other problems occasionally arise with Word files, particularly when produced by old versions of Word. The solution for glitches arising from Word's "fast save" is obvious: turn the option off and resave the file. In other cases, read the document into Word and use its "Save as HTML" feature to produce a file that Greenstone can process. When building digital libraries you sometimes need to be resourceful in finding ways around such difficulties.

Unit closing

In this unit you have installed Greenstone and practiced building digital library collections from documents of different formats.

You have seen five stages in building a collection: gathering the documents together, enriching them with metadata, designing the collection by specifying what searching and browsing facilities it will offer the user, building it, and previewing the result. (In Unit 3.1 you will learn about another possible stage: downloading documents from the web.) Usually, the process is iterative: you cycle through the designing, building, and previewing stages until you are satisfied with the result. You often begin by using low-quality automatically extracted metadata to get a rough prototype that you can show people, then add higher-quality metadata by hand and re-jig the collection design to use it. Also, you often begin with just a handful of documents and once you are satisfied with the collection design, proceed to add all the documents and begin the labour-intensive process of enriching them with metadata.

In our experience, an initial rough collection can usually be created within an hour or two. Then the collection design process may take from one day to one week. Finally, adding all the metadata could take months, even years, depending on the scale of the project.

You have also practiced some maintenance operations on Greenstone. You know where the software is on your computer and where collections are kept. The Greenstone home page automatically shows the collections in the *collect* folder (though there is an additional mechanism for keeping them hidden if desired). You can experiment by moving collections in and out: you may need to restart Greenstone to see them. We have not described the other directories in the Greenstone installation, but you are encouraged to explore them yourself. The whole point of open source software is that it's open for you to investigate.

You have also reinstalled Greenstone. Note that existing collections remain intact. It is a good idea to keep your Greenstone installation up to date by checking for new versions at http://www.greenstone.org, where they appear roughly twice a year. Better still, join the Greenstone discussion group mentioned in Unit 4.3. New versions fix bugs that may have been bothering you, and add new features that may be useful. Backwards compatibility is a high priority for the Greenstone developers: you should not suffer by keeping your software up to date. And it's free!

Now it is time to learn more about metadata and educational metadata. This is covered in Module 3. There is still much more to learn about Greenstone, and future units contain practical exercises to extend your knowledge and gradually turn you into an expert user.

MODULE 3 WORKING WITH METADATA

Goal

To understand the role of metadata in the organization of electronic resources, such as digital libraries and learning management systems, to become acquainted with common metadata standards including ones specifically intended for educational metadata, to appreciate the role and importance of the XML markup language, and to experience defining metadata and using it in various different ways.

Objectives

Upon completion of Module 3, you will be able to:

- explain the meaning of the term "metadata";
- distinguish between internal and external metadata;
- recount the history of the development of HTML, SGML and XML;
- contrast the design goals of HTML, SGML and XML;
- explain the purpose of XHTML and CSS;
- define what is meant by a DTD and explain how it relates to XML documents;
- navigate through the maze of X-standards: XSL, XLink, XPath, XML Schema, XSLT;
- describe the Dublin Core metadata standard;
- discuss the mapping of MARC metadata into Dublin Core;
- compare and contrast the bibliographic formats BibTeX and Refer;
- name popular metadata standards for image metadata and multimedia metadata;
- choose appropriate formats for the metadata in your digital library;
- design and build a rich variety of Greenstone digital library collections;
- explain how the aims of LOM and SCORM differ;
- build Greenstone collections using LOM metadata;
- explain the relationship between LOM and SCORM.

Introduction to Module 3

Documents and metadata are two basic components of any library. Documents provide the content and metadata supplies the organization. In Module 2 you studied the nitty-gritty details of documents and document formats; in the present module you will become acquainted with the nitty-gritty details of metadata and metadata formats. As before, this will be amply illustrated by practical work with the Greenstone digital library software.

Like documents, metadata – often characterized as "data about data" – is a kind of raw material for digital libraries. It provides the basis for organizing both digital and traditional libraries. The related term "markup" historically refers to the process of annotating documents with typesetting information. In an electronic environment this corresponds to annotating documents with formatting commands. More recently, this has been extended to annotating documents with structural information – including metadata – as well. Unit 3.1 covers markup and introduces the markup languages HTML, XML, and related standards. XML is really a scheme for defining markup languages, rather than a single markup language in itself. Unit 3.2 covers metadata. It relates the metadata used in digital libraries to the contents of traditional library catalogues and introduces the idea of extracting metadata from the raw text of the documents themselves.

In Module 2 you learned how to build simple digital library collections using Greenstone. Metadata is the key to organization within a digital library, and Units 3.1 and 3.2 contain a graded series of exercises designed to bring you up to a more advanced understanding of the organizational facilities that Greenstone provides. The specific user-access mechanisms in any individual collection are based on the metadata that is defined within the collection. You will learn how to use the Librarian Interface to design different collections. These practical exercises form a substantial part of the material in these units: you must be prepared to spend a substantial amount of time on them.

Of particular interest to this course is metadata that is specifically educational, which is the topic of Unit 3.3. Associated with most teaching and learning materials are important properties, such as educational level, prerequisites, difficulty, typical learning time, and so on. A popular recent trend is to break educational material down into discrete chunks called "learning objects" and use metadata to relate them to each other and allow them to be glued together into a complete educational experience – a course. Various educational metadata standards have been proposed to allow educators to readily locate, share, and reuse information in educational digital libraries. In Unit 3.3 you will learn about two of the major standards for representing educational content: LOM and SCORM. You will learn how to build Greenstone collections of educational metadata using LOM. LOM uses XML and is a practical example of the concept introduced in Unit 3.1, that XML can be used to define other markup languages.

Module 3 Readings

Unit 3.1. Markup

Reading 1 The textbook *How to build a digital library*: Chapter 5 – Introduction, Sections 5.1, 5.2, and 5.3.

Purpose Introduces the idea of markup and metadata and then describes the Hypertext Markup Language HTML, the Extensible Markup Language XML, and the presentation and formatting of documents marked up in XML.

Reading 2 CD–ROM IMARK: Unit 2 – Electronic documents and formats: Lessons 2.4 and 2.5.

Purpose Lesson 2.4 is not a mandatory part of this course but is recommended as enrichment material to increase your understanding of HTML. Lesson 2.5 introduces XML and provides a complementary perspective to the course text.

Unit 3.2. Metadata

Reading 1 The textbook *How to build a digital library*: Chapter 2, Section 2.2 and Chapter 5, Sections 5.4 and 5.6.

Purpose The Chapter 2 excerpt sketches the fundamental elements of traditional bibliographic organization. Section 5.4 describes two key standards for representing document metadata: MARC and Dublin Core. Section 5.6 discusses how metadata can be extracted automatically from documents.

Reading 2 CD–ROM IMARK: Unit 3 – Metadata standards and subject indexing: Lessons 3.1, 3.2, and 3.3.

Purpose Lesson 3.1 introduces descriptive metadata and explains its purpose. Lesson 3.2 explains the elements of the Dublin Core metadata standard, while Lesson 3.3 describes extensions to the standard.

Unit 3.3. Educational metadata

Reading 1 Course CD–ROM, Readings: "Learning technology standardization: making sense of it all", Erik Duval. *International Journal on Computer Science and Information Systems*, 2004, No. 1, pp. 33–43. http://www.comsis.fon.bg.ac.yu/ComSISpdf/Volume01/InvitedPapers/ErikDuval.pdf

Purpose A review of the state of the art in educational metadata standards.

Reading 2 Course CD–ROM, Readings: "Building educational metadata application profiles", Norm Friesen, Jon Mason and Nigel Ward. *Proceedings of the International Conference on Dublin Core and Metadata for e–Communities*, 2003, pp. 63–69. http://www.bncf.net/dc2002/program/ft/paper7.pdf

Purpose Case studies of the development of two metadata 'application profiles'.

Reading 3 Course CD–ROM, Readings: "An Introduction to ADL and the SCORM", Academic ADL Co–Lab, 2003. http://projects.aadlcolab.org/scourse/latestgreatest/viewer.htm (Is also provided in the folder "SCORM_tutorial" of the Course CD–ROM.)

Purpose	To provide an overview of how learning management systems work and how SCORM allows such systems to exchange content resources.
Reading 4	Course CD–ROM, Readings: "A LOM research agenda", Erik Duval and Wayne Hodgins. *Proceedings of the Twelfth International Conference on World Wide Web*, 2003, edited by Hencsey, G. White, B. Chen, Y. Kovacs L., and Lawrence, S. pp. 1–9. http://www2003.org/cdrom/papers/alternate/P659/p659–duval.html
Purpose	A discussion of the future of educational metadata standards.

UNIT 3.1 Markup

This unit explains what metadata is and how it provides the organizational infrastructure for a digital library. Metadata comes in two flavours: internal metadata or "markup", which is intended to assist readers in navigating within individual documents, and external metadata, which is descriptive information about complete documents. But, as we shall see, the distinction, although it sounds simple, becomes blurred.

In conventional documents, internal metadata for navigation is expressed typographically by using different styles for headings, lists, captions, footnotes, and so on. The information required for navigation becomes inextricably mixed up with the information required for presentation. Typographers use the term "markup" to specify the structure of individual documents and control how they look when presented to the user. Markup is a kind of metadata, the kind that the present unit focuses on. Various different formats for external metadata are the subject of the next unit.

We ask you to begin by thinking about the organization of a conventional physical library and about the organization (or lack of it) of the World Wide Web.

Assignment 1

The World Wide Web is not a digital library, as we discussed in Module 1. One of the crucial elements it lacks is organization (the other is selectivity). The two principal access mechanisms for documents in a physical library are the library catalogue and the shelving conventions that place like books together. The two principal access mechanisms for web pages are hyperlinks from one page to related pages and full-text search using search engines.

Compare and contrast these four access mechanisms [write in note form: suggested length about ten bulleted points].

Suppose you had the power and resources to design and implement a library-style catalogue for the web. What attributes of documents are listed in your local library catalogue? Which of these attributes and what additional ones would you list in a catalogue of the web?

One of the problems we expect you encountered in thinking about a catalogue for the web in Assignment 1 is the question of what granularity to use when assigning metadata – whether to individual web pages, to groups of pages, or to entire web sites. The indexes in conventional libraries normally operate at the level of individual books or series of journals. They rarely reach inside these items to catalogue individual chapters or articles. What is the web analogy of an individual book or journal series? There is no clear answer.

The web as we know it today is based on HTML, the Hypertext Markup Language. And since most digital libraries use the web for distribution – even if it is not the sole distribution mechanism – HTML tends to be the baseline format for interactive viewing of digital library documents, even ones that are available in other formats. As a language for document representation, a discussion of HTML might arguably belong in Module 2. We have chosen to discuss it here because HTML is at heart a markup language. Moreover, it has another important aspect. The word "hypertext" refers to the facility for linking between documents, which is the underlying basis of the web's organization. It also creates a conundrum: what is the appropriate granularity to assign metadata to the web?

Assignment 2

Read the beginning of Chapter 5 and Sections 5.1 and 5.2 (pp. 221–237) of the textbook and answer these questions:

- List three important differences between the aims and underlying philosophies of HTML and XML.
- Where does XHTML fit in?

- How is metadata stored in HTML?
- How is metadata stored in XML?
- What is a DTD?
- How can occasional non-ASCII Unicode characters be represented in HTML (and XML)?
- How can an entire non-ASCII Unicode document be represented in XML (and HTML)?
- How can you extract plain text from an HTML file?
- What are the two kinds of markup?

Enrichment exercise 1

The IMARK Module introduces HTML in Lesson 2.4, which summarizes its features and shows how to produce a simple HTML document. For the present course you do not need to produce HTML documents, although you will use ones that others have produced in the educational digital library collections you build. Nevertheless, it is a relevant and useful skill, and we recommend that you work through this lesson to enhance your understanding of character coding issues.

Markup is a kind of metadata. Structural markup makes the document structure explicit, allowing navigation within the document. Presentational markup is not really metadata, but is often closely intertwined structural markup. Certainly the two were hard to disentangle in the early days of HTML. An important trend over the past decade has been to attempt to distinguish structure from presentation and to provide means for structural markup that can be rendered into different styles of visual representation.

Assignment 3

Read Section 5.3 (pp. 237–253) of the textbook and work through IMARK Lesson 2.5. Then answer these questions:

- Explain how CSS helps to enforce a distinction between structural and presentational markup.
- What are the three components of XSL, and which one relates most closely to CSS?
- What would you use to change an XML document from one format to another?
- Suppose you had a book represented in XML. What language in the XML family enables you to refer to an individual chapter or section?
- What is the purpose of a namespace?

The practical exercise that follows involves creating a Greenstone collection from a moderately large number of HTML files. We have chosen a set of documents relating to the Tudor period in English history, which are on your course CD-ROM (the one marked *Digital Libraries in Education*) in the folder called *sample_files/tudor*. You might imagine that you are a teacher of English history preparing a resource for your students comprising documents that you previously downloaded from the web.

We create a basic collection from these files. Depending on how powerful your computer is, this collection may take some minutes to build, and if you are pressed for time, navigate to one of the folders inside, e.g. *Monarchs*, and use it instead. Then we look at different views of the files in the Gather and Enrich panels, which is useful when working with files of a certain type.

Practical exercise: HTML collection – Tudor

1. Invoke the Greenstone Librarian Interface (from the Windows *Start* menu) and start a new collection called **tudor** (use the *File* menu). Fill out the pop–up dialog with appropriate values and leave **Dublin Core**, which is selected by default, as the metadata set.

2. In the **Gather** panel, open the *tudor* folder in *sample_files*.

3. Drag *englishhistory.net* from the left–hand side to the right to include it in your **tudor** collection.

4. Switch to the **Create** panel and click <**Build Collection**>.

5. When building has finished, **preview** the collection.

6. Note that the browsing facilities in this collection (*titles a–z* and *filenames*) are based entirely on extracted metadata. Return to the Librarian Interface and examine the metadata that has been extracted for some of the files.

You've probably noticed that the collection contains a few stray image files, as well as the HTML documents. This is a mistake. The issue is that many of the HTML documents include images, and although Greenstone attempts to determine which images belong to HTML pages and only considers other images for inclusion in the collection, in this case it hasn't been completely successful. (This is because the web site from which these files were downloaded occasionally departs from the usual convention of hierarchical structuring.)

7. Switch to the **Design** panel and select the **Document Plugins** section. Beside **plugin HTMLPlug** you will see – *smart_block*. This is the option that attempts to identify images in the HTML pages and block them from inclusion – in this case, it's not smart enough. Select the **plugin HTMLPlug** line and click <**Configure Plugin**>. A popup window appears. Scroll down the page to locate the **smart_block** option and switch it off. Click <**OK**>.

8. Switch to the **Create** panel and **build** and **preview** the collection. The collection is exactly as before except that these stray images are suppressed. What is happening is that plug–ins operate as a pipeline: files are passed to each one in turn until one is found that can process it. By default (ie. without *smart_block*) the HTML plug–in blocks *all* images.

Now look at different views of the files:

9. Switch to the **Gather** panel and open in the right–hand side *englishhistory.net* → *tudor*.

10. Change the **Show Files** menu for the right–hand side from **All Files** to **HTM & HTML**. Notice the files displayed above are filtered accordingly, to show only files of this type.

11. Change the **Show Files** menu to **Images**. Again, the files shown above alter.

12. Now return the **Show Files** setting back to **All Files**, otherwise you may get confused later. Remember, if the **Gather** or **Enrich** panels do not seem to be showing all your files, this could be the problem.

One reason why you, as a history teacher, might have made this collection is to put it on CD-ROM to distribute it to students who do not have easy access to the web. In the next exercise we create the files necessary to produce precisely that. When you have these files, you will need to use your system's own software to actually create the CD-ROM. If you want, you can export several collections together. You may wish to defer actually writing the CD-ROM until you have completed this course, and put all the collections you have built on it as a record of your work.

Practical exercise: Exporting to CD–ROM

13. Launch the Greenstone Librarian Interface if it is not already running.

14. Choose **File** → **Write CD/DVD image** and in the popup window select the **tudor** collection as the collection to export. You can optionally name the CD–ROM; otherwise the default "collections" is used. Do so now, entering "Tudor collection" in the field for **CD/DVD name**; then click <**Write CD/DVD image**>.

The necessary files for export are written to:

C:\Program Files\greenstone\tmp\exported_Tudorcollection

You need to use your own computer's software to write these on the CD–ROM. On Windows XP this ability is built into the operating system: assuming you have a CD–ROM or DVD writer insert a blank disk into the drive and drag the contents of *exported_Tudorcollection* into the folder that represents the disk.

The result will be a self–installing Greenstone CD–ROM, which starts the installation process as soon as it is placed in the drive, like the course CD–ROM did when you first used it, back in Unit 1.3.

The documents in the Tudor collection you have built were downloaded from the web when the course CD-ROM was prepared. If your library readers are connected to the web, it might be better to show them the original version of the document rather than the digital library's copy. Then when readers clicked on a document link, the digital library software would give the document's original URL to their browser, which would go directly to that web page. In this way, a networked digital library server could provide searching and browsing facilities for online material. In other situations, it is more appropriate to store source documents within the collection (as your Tudor collection does), so that users can read them when disconnected from the web; the documents remain available even when the website goes down or the URL changes; and the collection can be written to a self-contained CD-ROM.

What is required for an online library that indexes web material is the ability to index the text whilst sending the user back to the original document on the web when a document is requested – essentially running the digital library as though it were a web search engine. Greenstone supports this, and we now alter our Tudor collection accordingly. If you are not connected to the web, you may wish to skip or skim the following exercise.

Practical exercise: A web–based collection

1. Open up your **tudor** collection and in the **Gather** panel inspect the files you dragged into it. The first folder is *englishhistory.net*, which opens up to reveal *tudor*, and so on. The files represent a complete sweep of the pages (and supporting images) that constitute the Tudor section to the *englishhistory.net* web site. They were downloaded from the web in a way that preserved the structure of the original site. This allows any page's original URL to be reconstructed from the folder hierarchy.

2. In the **Design** panel, select the **Document Plugins** section, then select the **plugin HTMLPlug** line and click <**Configure Plugin**>. A popup window appears. Locate the **file_is_url** option (about halfway down the first block of items) and switch it on. Click <**OK**>.

Setting this option to the HTMLPlug means that Greenstone sets an additional piece of metadata for each document called URL, which gives its original URL.

It is important that the files gathered in the collection start with the web domain name (englishhistory.net in this case). The conversion process will not work if you dragged over the tudor folder, because this will set URL metadata to something like:
http://tudor/englishhistory.net/tudor/...
rather than
http://englishhistory.net/tudor/...

If you have copied over the tudor folder previously, delete it and make a fresh copy. Drag the tudor folder in the right–hand side of the ***Gather*** *panel on to the trash can in the lower right corner. Then obtain a fresh copy of the files starting with the* englishhistory.net *folder, by opening tudor on the left–hand side and dragging its contents across.*

3. To make use of the new URL metadata, we must change the icon link to serve up the original URL rather than the copy stored in the digital library. Go to the **Design** panel, select the **Format Features** section, and edit the **VList** format statement by replacing:
   [link][icon][/link]
   with
   [weblink][webicon][/weblink]

Click <**Replace Format**> to commit the change.

4. Switch to the **Create** panel and **build** and **preview** the collection. The collection behaves exactly as before, except that when you click a document icon your web browser retrieves the original document from the web (assuming it is still there by the time you do this exercise.). If you are working offline you will be unable to retrieve the document.

In this exercise we provided users with the web version of each document. Alternatively, depending on the anticipated usage, it might make sense to provide links to both the local copy *and* the web version: this can be done using a more complex format statement. Also, if the online resource changes frequently, the website should be rescanned regularly and the collection rebuilt, otherwise the information that is indexed will be out of date.

The same technique can be applied to multiple web sites. Indeed this is a more likely scenario: a teacher may establish a collection of useful resources on the topic of Tudor England by locating relevant web sites (and parts of sites), downloading them, and making a digital library collection. As a final example we return to the **tudor** collection but re-download it from the web rather than using the canned version supplied on the course CD-ROM. Note: to complete this exercise you must have installed Greenstone from the course CD-ROM marked *Digital Libraries in Education*. The IMARK CD-ROM entitled *Digitisation and Digital Libraries* contains an older version of Greenstone that does not have the downloading capability.

Practical exercise: Downloading from the web

The Greenstone Librarian Interface's ***Download*** *panel allows you to download individual files, parts of websites, and indeed whole websites from the web*.

1. Start a new collection called **webtudor** and base it on the **tudor** collection.

2. In a web browser visit *http://englishhistory.net*, follow the link to *Tudor England*, and click <**enter**>. You should be at the URL:
   http://englishhistory.net/tudor/contents.html

This is where we started the downloading process to obtain the files you have been using for the **tudor** collection.

3. You could do the same thing by copying this URL from the web browser, pasting it into the **Download** panel, and clicking the <**Download**> button. However, several megabytes will be downloaded, which might strain your network resources – or your patience. For a faster exercise we focus on a smaller section of the site. In the **Download** panel, enter this URL:
   http://englishhistory.net/tudor/monarchs/edward6.html
   into the **Source URL** box. There are several options that govern how the download process proceeds. To copy the *monarchs* section of the website, select **Only mirror files below this URL**. If you don't do this, the downloading process will follow links to other areas of the *englishhistory.net* website and grab those as well.

4. Now click <**Download**>. A progress bar appears in the lower half of the panel that reports on how the downloading process is doing.

More detailed information can be obtained by clicking <**View Log**>. The process can be paused and restarted as needed or stopped altogether by clicking <**Close**>. Downloading can be a lengthy process involving multiple sites, and so Greenstone allows additional downloads to be queued up. When new URLs are pasted into the Source URL box and <**Download**> clicked, a new progress bar is appended to those already present in the lower half of the panel. When the currently active download item completes, the next is started automatically.

5. Downloaded files are stored in a top–level folder called **Downloaded Files** that appears on the left–hand side of the **Gather** panel. You may not need all the downloaded files and you choose which you want by dragging selected files from this folder over into the collection area on the right–hand side, just like we have done be–fore when selecting data from the *sample_files* folder. In this example we will include everything that has been downloaded.

Select the **englishhistory.net** folder within **Downloaded Files** and drag it across into the collection area.

6. Switch to the **Create** panel to **build** and **preview** the collection. It is smaller than the previous collection because we included only the *monarchs* files. However, these now represent the latest version of the documents. Since you based your **webtudor** collection on **tudor** it includes the modified *[weblink] [webicon] [/weblink]* format, so the new collection also links back to the original web documents.

Over the years HTML has accumulated a variety of extensions and add-ons, and is now a messy format to contend with. The Greenstone HTML plug-in reads HTML files and parses their content. This process does not always go smoothly. If web pages use certain features, the documents in the collection may not be faithful renditions of the originals. In extreme cases a document might appear as a blank page – or not at all. In other cases the presentation of the page changes significantly: it might lose a background image or some glitzy functionality.

Here are some known problems.

Frames. Documents that use HTML "frames" are decomposed into components that are treated as individual documents in their own right. HTMLPlug does nothing special with web pages that contain frames.

Comments. HTML applications sometimes use comments <!-- ... --> to convey additional document processing information. HTMLPlug ignores comments.

The <head> tag. HTMLPlug discards the information in a document's <head> tag, which can lead to the loss of style files and JavaScript functionality. (The – *keep_head* switch retains this, but is not enabled by default because it can cause other problems.)

JavaScript is a programming language that can be embedded in HTML to manipulate a document's content and structure – for example, it might load in a new image at run-time. HTMLPlug ignores JavaScript because at build time it must establish all external resources (images and so forth) that are needed to reconstruct the do-cument.

Embedded technologies. HTML pages can embed an ever-expanding set of additional web technologies. There are options to HTMLPlug that help cope with some of these, but using them requires specialist knowledge.

As a general rule, the default settings in HTMLPlug work well if the pages you include in your collection conform to the core HTML syntax – i.e. they avoid JavaScript, embedded web technologies, DOM manipulation or antiquated frames syntax. The web contains countless examples of attractive pages that conform to these constraints, and the "keep it simple" approach also helps ensure that your web pages display as intended on a wide range of browsers and operating systems. Using flashy new features increases the chance that the page will not display properly and sometimes even crashes the user's browser.

Unit closing

This module, Module 3, is about working with metadata. Metadata is the glue that holds the documents in a digital library together. From the reader's perspective, this is what gives the library its value as a collection rather than as a set of independent documents. Yet this unit, Unit 3.1, is about markup. Markup developed historically as one of the processes involved in preparing a book for publishing and was originally concerned with the presentation of a work rather than as a means of expressing metadata for it. From this perspective, the present unit belongs in Module 2 on Document representation. And indeed HTML really is a language for document presentation.

Three factors complicate the issue. One concerns the notion of an individual "work". This is fairly clear in the case of a printed book (although even here there are many questions about whether different editions, versions, translations, audio recordings for the blind, or films based on the book constitute part of the same "work" or not). The notion of a "work" is far murkier in the case of hypertext, however, where web pages are not individual entities in their own right but part of a massive linked structure that we call the web. The second complicating factor is the trend towards separating presentation from structure and encoding each individually, as exemplified by the move from raw, unstructured, presentation-oriented HTML to carefully structured XML with all presentational decisions relegated to separate CSS stylesheets. The third is that, right from the beginning, HTML has explicitly encoded some metadata, namely *Title*, and also provided for other metadata to be placed in a <*meta*> tag. Moreover, even without explicit tagging, it is possible for some metadata to be extracted directly from the text of the document with a fair degree of accuracy – we will learn about this in the next unit (Assignment 3).

As well as learning about HTML, XML, and related standards, this unit has shown you how to build a collection of HTML documents in Greenstone. You learned how to put a Greenstone collection onto CD-ROM and how to use Greenstone to download files from the web to incorporate in your collections.

In practice, a very common application of Greenstone worldwide is to create and organize collections of HTML documents. As you have seen, the Librarian Interface has a *Download* panel that allows you to download individual web pages, parts of web sites, or even whole web sites into a holding area on your computer so that you can then use them in your collections. This provides such a powerful facility we feel obliged to remind you of one of the lessons of Unit 1.1: building and distributing information collections carries responsibilities you should reflect on before you begin. There are legal issues of copyright: being able to access documents doesn't mean you can necessarily give them to others. There are social issues: collections should respect the customs of the community out of which the documents arise. And there are ethical issues: some things simply should not be made available to others. The pen is mightier than the sword – be sensitive to the power of information and use it wisely.

UNIT 3.2 Metadata

This unit continues the exploration of metadata begun in the previous unit. Here we focus on *external* metadata, which is what library catalogues contain. This can be stored externally to the document that the metadata describes. However, many document markup schemes – XML is a good example – often allow "external" metadata to be embedded within the document. The same is true of physical documents: the title, author, publication date, and publisher of a book are usually included within the book itself. External metadata is really metadata that *could* stand alone, separately from the document, and still be useful. In contrast, internal metadata does not really make sense independently of the document itself.

Assignment 1

Read Chapter 2 Section 2.2 (pp. 46–55) and work through IMARK Lesson 3.1. Then answer these questions:

- The textbook states that a bibliographic system supports five different activities: to locate, identify, select, acquire, and navigate information. Review these and discuss them in the context of the tudor collection that you built in Unit 3.1.
- Using a standard classification system (e.g. the one used in your local library, or the Library of Congress Subject Headings) identify some terms and their relationships that apply to the documents in the Tudor collection.

Assignment 2

Read Chapter 5 Section 5.4 (pp. 253–261) of the textbook and work through IMARK Lessons 3.2 and 3.3.

In Greenstone, a file called *marctodc.txt* governs the conversion of MARC metadata to Dublin Core. Locate this file: if you installed the software in the standard place it will be C:\Program Files\greenstone\etc\marctodc.txt. Examine the file and figure out how a typical MARC catalogue record, such as the one in Table 5.2 (p.255) of the textbook, translates into Dublin Core.

The following practical exercise asks you to build an image collection. Image collections rely on external metadata for all searching and browsing operations, whereas in textual collections titles can usually be extracted from documents, and the documents can be searched even if no metadata is available. In this collection you will define your own metadata and provide searching and browsing facilities based on it. You can use your own images or the ones provided in the folder sample_files\images on the course CD-ROM.

Note: you will not be able to build a collection of images unless ImageMagick has been successfully installed on your computer as described in Assignment 1 of Unit 2.2.

Practical exercise: An image collection

1. Start a new collection (File → New) called **backdrop.** Fill out the fields with appropriate information. For **Base this collection on**, pull down the item **Simple image collection (image–e)** from the menu.

You aren't asked to choose a metadata set because the new collection inherits whatever is used by the seed collection.

2. Copy the images provided in the images folder into your newly–formed collection.

3 Change to the **Create** panel and **build** the collection.

4. **Preview** the result.

5. Click <**browse**> in the navigation bar to view a list of the photos ordered by filename and presented as a thumbnail accompanied by some basic data about the image. The structure of this collection is the same as **Simple image collection (image–e)**, but the content is different.

6. Change to the **Enrich** panel and view the extracted metadata for *Ascent.jpg*.

We now add our own metadata and use it to give users a new way to browse the collection. We use the Dublin Core metadata set.

7. The collection (image–e) on which **backdrop** is based uses only extracted metadata. To add a further metadata set, go to the **Design** panel of the Librarian Interface and click <**Metadata Sets**> in the list on the left (the last one). Then click <**Add Metadata Set**> (lower left button).

8. In the window that pops up select **dublin.mds** and click <**Add Metadata Set**>.

9. Now switch to the **Enrich** panel by clicking this tab. The metadata for each file now shows the Dublin core *dc.* fields as well as the extracted *ex.* fields.

10. We work with just the first three files (*Ascent.jpg, Autumn.jpg, and Azul.jpg*) to get a flavour of what is possible. First, set each file's **dc.Title** field to be the same as its filename but without the filename extension.

11. Click on ***Ascent.jpg*** so its metadata fields are available, then click on its **dc.Title** field on the right–hand side. Click on the **Value** text box, enter **Ascent**, and click <**Append**>.

*The **All Previous Values** box will become more useful when more entries have been added.*

12. Repeat the process for **Autumn.jpg** and **Azul.jpg**.

Now we customize the collection's appearance. Building or previewing the collection at this point won't reveal anything new. That's because we haven't changed the design of the collection to take advantage of the new metadata.

13. Go to the **Design** panel (by clicking on its tab) and select **Format Features** from the left–hand list. Leave the **Editing Controls** at their default value, so that **Choose Feature** remains blank and **VList** is selected as the **Affected Component.** In the **HTML Format String**, edit the text as follows:
    Change "_ImageName_:" to "Title:"
    Change "[Image]" to "[dc.Title]"

Metadata names are case–sensitive in Greenstone: it is important that you capitalize "Title" (and don't capitalize "dc").

14. Next click <**Replace Format**>. The first of the above changes alters the fragment of text that appears to the right of the thumbnail image, the second alters the item of metadata that follows it.

15. Go to the **Create** panel and click <**Build Collection**>. Now **preview** the collection. When you click on **browse** in the navigation bar the presentation has changed to "Title: Ascent" and so on.

Because we only assigned metadata to the first three items, after this the title becomes blank because the subsequent items have no dc.Title metadata. To get a full listing, enter all the metadata.

*For some design parameters the collection must be rebuilt before the effect of changes can be seen. However, changes to format statements take place immediately and you can see the result straightaway by clicking **reload** in the web browser. Above, you were asked to build before previewing just to simplify the explanation.*

Next we change the size of the thumbnail image and make it smaller.

16. Thumbnail images are created by the *ImagePlug* plug–in, so we need to access its configuration settings. To do this, switch to the **Design** panel and select **Document Plugins** from the list on the left. Double–click **plugin ImagePlug** to pop up a window that shows its settings. (Alternatively, select **ImagePlug** with a single click and then click <**Configure Plugin**> further down the screen). Currently all options are off, so standard defaults are used. Select **thumbnailsize**, set it to **50**, and click <**OK**>.

17. **Build** and **preview** the collection.

18. Once you have seen the result of the change, return to the **Design** panel, select the configuration options for *ImagePlug*, and switch the thumbnail size option off so that the thumbnail reverts to its normal size when the collection is re–built.

Now add metadata that describes the photos in the collection. Again, for illustration, we focus on the first three images (Ascent.jpg, Autumn.jpg, and Azul.jpg).

19. Switch to the **Enrich** panel and select *Ascent.jpg*. We'll store our description in the **dc.Description** metadata element, so select it now in the right–hand panel.

What description should we enter? To remind yourself of a file's content, the Librarian Interface lets you open files by double–clicking them. It launches the appropriate application based on the filename extension, Word for .doc files, Acrobat for .pdf files, and so on. Double–click Ascent.jpg: the image will normally be displayed by Microsoft's Photo Editor (although this depends on how your computer has been set up).

20. Back in the Librarian Interface enter the text **Moon rising over mountain landscape** as the **dc.Description** field's value and click <**Append**> to have it added.

21. Repeat this process for *Autumn.jpg* and *Azul.jpg*.

22. Build the collection again, to incorporate the new metadata.

23. Now update the format statement to use the new **dc.Description** metadata. Switch to the **Design** panel and enter the **Format Features** section by selecting this from the list of names on the left–hand size and ensure **VList** is selected. In the **HTML Format String**, place your cursor after the text that says:
    [dc.Title]
    and add the following text:
    Description: [dc.Description]
    Then click <**Replace Format**>.

24. **Preview** the result (you don't need to build the collection as was done in step 22 to incorporate the metadata, because changes to format statements take effect immediately). Each image's description should appear beside the thumbnail, following the title.

Now we add a new browsing option based on the descriptions.

25. Switch to the **Design** panel and select **Browsing Classifiers** from the left–hand list. Set the menu item for **Select classifier to add** to **AZList**, then click <**Add Classifier**>.

26. A window pops up to control the classifier's options. Set the menu item for metadata to **dc.Description** and click <**OK**>. Now switch to the **Create** panel, **build** the collection, and **preview** it. Choose the new **descriptions** link that appears in the navigation bar.

Only three items are shown, because only items with the relevant metadata (dc.Description in this case) appear in the list. The original browse list includes all photos in the collection because it is based on ex.Image, extracted metadata that reflects an image's filename, which is set for all images in the collection.

Finally we create a searchable index based on dc.Description metadata.

27. Switch to the **Design** panel and select **Search Indexes** from the left–hand list. Enter the text "descriptions" as the **Index Name**, select **dc.Description**, and click <**Add Index**>.

28. Switch to the **Create** panel, **build** the collection, then **preview** it. As an example, search for the term "autumn".

Greenstone allows metadata to be added to any collection. And that metadata can be used in many different ways, as the basis for full-text indexes, or for browsing facilities, or for display. Greenstone does not impose any particular design on collections. It is not restricted to any particular style of digital library collection; it is rather a framework in which digital librarians, educators, and others can rapidly form the style of collection that suits their needs given the source documents they are working with. The good news is that it's very flexible, the bad news is that to take advantage of this flexibility you have to learn to use it. There are many choices to learn about, and that is why so many practical exercises are included in this course.

In the next exercise we return to the rudimentary Tudor collection that we built in the first practical exercise of Unit 3.1 and enhance it in three different ways. First, we define a small amount of metadata that splits the documents into four categories and add a facility that allows readers to browse the documents in these categories. Then we partition the full-text index into four corresponding pieces. In the exercise we do this on the basis of the metadata value we just defined, although partitioning can be based on any metadata value or on the document filenames. Next we add a hierarchical phrase index. Finally, we look at how the collection-building process can be controlled to help with the development of collections with large numbers of documents.

As you learned in the last assignment of Unit 2.2, the Librarian Interface supports four levels of user. So far we have been operating in Librarian mode, but for this exercise we must switch to Library Systems Specialist mode (or higher). Creating the partitions involves defining a "regular expression". Regular expressions are powerful ways of matching text, but are pernickety and take time to learn; other than describe what must be done to complete this assignment we do not explain them further here. They are the reason why this feature is unavailable in Librarian mode.

Practical exercise: Enhanced Tudor collection

We return to the Tudor collection and add metadata that expresses a subject hierarchy. Then we build a classifier that exploits it by allowing readers to browse the documents about Monarchs, Relatives, Citizens, and Others separately.

1. Open up your **tudor** collection (the original version, not the **webtudor** version), switch to the **Enrich** panel, and select the *monarchs* folder (a subfolder of *tudor*). Set its **dc.Subject and Keywords** metadata to **Tudor period|Mo–narchs.** (For brevity, we refer to this metadata element in future simply as **dc.Subject.**) The vertical bar ("|") is a hierarchy marker. Selecting a *folder* and using the **Append** button to set its metadata has the effect of setting this metadata value for all files contained in this folder, its subfolders, and so on. A popup alerts you to this fact.

2. Repeat for the *relative* and *citizens* folder, setting their **dc.Subject** metadata to **Tudor period|Relatives** and **Tudor period|Citizens** respectively. Note that the hierarchy appears in the **All Previous Values** area.

3. Finally, select all remaining files – the ones that are not in the *monarchs, relative*, and *citizens* folders – by selecting the first and shift–clicking the last. (This is the normal technique of multiple selection.) Set their **dc.Subject** metadata to **Tudor period|Others**: this is done in a single operation (there is a short delay before it completes).

4. Switch to the **Design** panel and select **Browsing Classifiers** from the left–hand list. Set the menu item for **Select classifier** to add to **Hierarchy**, then click <**Add Classifier**>.

5. A window pops up to control the classifier's options. Change the metadata to *dc.Subject* and then click <**OK**>.

6. For tidiness' sake, **remove** the **classifie**r for **Source** metadata (included by default) from the list of currently assigned classifiers, because this adds little to the collection.

7. Now switch to the **Create** panel, **build** the collection, and **preview** it. Choose the new **subjects** link that appears in the navigation bar and click the bookshelves to navigate around the four–entry hierarchy that you have created.

Next we partition the full–text index into four separate pieces. To do this we first define four subcollections obtained by "filtering" the documents according to a criterion based on their ***dc.Subject*** *metadata. Then we assign an index to each subcollection.*

8. Switch to the **Design** panel and click <**Partition Indexes**>. This feature is disabled because you are operating in *Librarian Mode* (this is indicated in the title bar at the top of the window).

9. Switch to *Library Systems Specialist* mode by going to **Preferences** (on the *File* menu) and clicking <**Mode**>. Read about the other modes, too. Note that the mode appears in the title bar.

10. Return to the **Partition Indexes** section of the **Design** panel. Ensure that the **Define Filters** tab is selected (the default). Define a subcollection filter with name **monarchs** that matches against **dc.Subject and Keywords** and type **Monarchs** as the regular expression to match with. Click <**Add Filter**>. This filter includes any file whose **dc.Subject** metadata contains the word *monarchs*.

11. Define another filter, **relatives**, which matches **dc.Subject** against the word **Relatives**. Define a third and fourth, **citizens** and **others**, which matches it against the words **Citizens** and **Others** respectively.

12. Having defined the subcollections, we partition the index into corresponding parts. Click the <**Assign Partitions**> tab. Select the first subcollection and give it the name **monarchs.** Click <**Add Partition**>. Repeat for the other three subcollections, naming their partitions **relatives, citizens**, and **others. Build** and **preview** the collection.

13. The search page includes a pulldown menu that allows you to select one of these partitions for searching. For example, try searching the *relatives* partition for *mary* and then search the *monarchs* partition for the same thing.

14. Return to *Librarian* mode, using **Preferences** (on the *File* menu).

Next we add an interactive hierarchical phrase index, which is called PHIND in Greenstone.

15. Switch to the **Design** panel and choose the **Browsing Classifiers** item from the left–hand list.

16. Choose **Phind** from the **Select classifier to add** menu. Click <**Add Classifier**>. A window pops asking for configuration options: leave the values at their preset defaults (this will base the phrase index on the full text) and click <**OK**>.

17. **Build** the collection again, **preview** it, and try out the new **phrases** option in the navigation bar. An interesting PHIND search term for this collection is **king**.

Finally we look at how the building process can be controlled. Developing a new collection usually involves numerous cycles of building, previewing, adjusting some enrich and design features, and so on. While prototyping, it is best to temporarily reduce the number of documents in the collection. This can be accomplished through the "maxdocs" parameter to the building process.

18. Switch to the **Create** panel and view the options that are displayed in the top portion of the screen. Select **maxdocs** and set its numeric counter to **3.** Now **build.** In fact, you will find that the collection now contains 5 documents (not 3 as you specified: for technical reasons the number you give to **maxdocs** is an approximate value).

19. Preview the newly rebuilt collection's **titles a–z** page. Previously this listed more than a dozen pages per letter of the alphabet, but now there are just three – the first three files encountered by the building process.

The above exercise has introduced you to several useful features of Greenstone. The Tudor collection is the largest we have built so far, but it is very small in the wider scheme of things. Real collections often contain thousands or hundreds of thousands – maybe millions – of documents. The collection development cycle is greatly accelerated by using just (say) 50 of them for development, as illustrated in the last part of the exercise.

Next we build a collection that is based solely on metadata: in this case a file of MARC records. Greenstone contains other metadata plug-ins, for example for BibTeX and Refer (mentioned in the Assignment 4 reading), and also for CDS/ISIS and PROCite, which are also commonly-used bibliographic formats. You could equally well build collections from documents in these formats. A collection of MARC records from the U.S. Library of Congress about the Beatles rock group is provided on the course CD-ROM in the folder *sample_files/marc*.

Practical exercise: A bibliographic collection

1. Start a new collection called **Beatles Bibliography.** Enter the requested information and base it on "New Collection". There is no need to include any metadata sets because the metadata extracted from the MARC records will appear as extracted metadata.

2. In the **Gather** panel, open the *marc* folder, drag **locbeatles50.marc** into the right–hand pane and drop it there. A popup window asks whether you want to add **MARCPlug** to the collection to process this file. Click <**Add Plugin**>, because this plug–in will be needed to process the MARC records.

3. Remove the plug–ins **TextPlug** to PSPlug **(ZIPPlug, GAPlug** and **MARCPlug** remain). It is not strictly necessary to remove these redundant plug–ins, but it is good practice to include only plug–ins that are needed, to avoid possible unexpected side effects.

4. Now select **Browsing Classifiers** within the **Design** panel and **remove** the default classifier for **Source** metadata. In this collection all records are from the same file, so **Source** metadata, which is set to the filename, is not particularly interesting.

5. Switch to the **Create** panel, **build** the collection, and **preview** it. Browse through the **titles a–z** and view a record or two. Try searching – for example, find items that include **George Martin.**

6. Add an **AZCompactList** classifier for the **Subject** metadata. Select this item from the relevant menu of the **Browsing Classifiers** section of the **Design** panel and click <**Add Classifier**>. In the popup window, select **ex.Subject** as the metadata item, activate the **mingroup** option, and set its field to **1.**

AZCompactList *is like* ***AZList****, except that terms that appear multiple times in the hierarchy are automatically grouped together and a new node, shown as a bookshelf icon, is formed. Setting* ***mingroup*** *to 1 means that the bookshelf appears even when there is just one item and is done here to provide a more uniform display*.

7. **Build** the collection and **preview** the result.

8. Make each bookshelf node show how many entries it contains by appending this to the **Format Features** for **VList** format statement in the **Design** panel:
   {If} {[numleafdocs],<td>([numleafdocs])</td>}

9. Click <**Replace Format**>, switch to the **Create** panel, and click <**Preview Collection**> (no need to build the collection again).

Next add fielded searching.

10. In the **Design** panel select **Search Types** from the left–hand list and activate the **Enable Advanced Searches** options.

11. **Build** the collection once again and **preview** the results. Notice that the collection's home page no longer includes a query box. (This is because the search form is too big to fit here nicely.) To search, you have to click **search** in the navigation bar. Note that the *Preferences* page has changed to control the advanced searching options.

To finish off the collection, brand it with an image that will be used to represent the collection on the Greenstone page and appear at the top of each page of the collection.

12. From the **General** section of the **Design** panel, click the <**Browse**> button next to the label **URL to 'about page' icon** and use the resulting popup file browser to access folder *sample_files\marc* in the course CD–ROM. Select *beatles_logo.jpg* and click <**Open**>.

Greenstone copies the image into your collection area, so the collection will still work when the CD–ROM is removed from the drive.

13. Repeat this process for the **URL to 'home page' icon**, selecting the same image.

14. Now **build** the collection and **preview** it.

Assignment 3

Read Chapter 5 Section 5.6 (pp. 266–280) of the textbook. Greenstone contains facilities for (a) language identification, (b) acronym extraction, (c) keyphrase extraction, and (d) creation of phrase hierarchies. (You met the hierarchical phrase browser in the enhanced Tudor collection, and also when you examined the course CD-ROM's *How to build a digital library* collection in Unit 1.3, Assignment 5.) Choose one of the following projects.

- The language of a document appears as extracted metadata and can be viewed in the Enrich panel. Explore the limitations of this facility by creating some mixed-language documents and importing them into a collection.
- Acronym extraction is switched on for each plug-in individually in its configuration panel. Find some documents that contain acronyms and their definitions and explore the success of the extraction scheme.
- Keyphrase extraction is switched in the same way. Explore the quality of the keyphrases that are extracted from some documents of your choice.

Write a short report on your investigation.

Unit closing

In this unit you have learned a great deal about metadata. You have learned about the role that a bibliographic system plays in supporting the library user's activities and have got an understanding of the purpose of metadata.

You have learned about two particular metadata standards in detail: MARC and Dublin Core (unqualified and qualified). There are countless others, but in many ways these represent polar extremes in the landscape. MARC is traditional and librarian-oriented – critics might dismiss it as complicated and heavyweight, particularly when used to describe resources that are more ephemeral than printed library holdings. Dublin Core is user-oriented and directed at the new world of the web – critics might dismiss it as superficial and amateurish. Qualified Dublin Core adds some further structure. Intermediate standards include MARCXML, which is an XML schema for representing MARC records, and MODS (for "metadata object description schema"), a method recently developed by the U.S. Library of Congress for representing bibliographic information that lies somewhere between MARC and Dublin Core in expressivity. We study MODS in Unit 5.1.

Exercises with Greenstone form a large part of this unit. As before these have a dual aim: to provide a practical perspective on the material in the unit and to extend experience of using Greenstone to build educational digital library

collections. You have encountered some advanced features, such as how to partition the full-text index and how to add a hierarchical classifier that implements a browsing structure based on explicitly assigned metadata. We expect that you will need to practice with your own collections before you become confident in using these facilities.

Metadata has two distinct roles in Greenstone: it can be added interactively to documents in collections, or collections can be built that consist of metadata. In this unit you have experienced both. You built a collection of images that is accessible only via the metadata you entered; users can search that metadata and browse it. Plain image collections have no full text or internal hyperlinking, so metadata is the only access mechanism. You also built a collection of MARC records (they could equally well have been CDS/ISIS, PROCite, BibTeX, or Refer records), and gave it a fielded search interface that resembles many library catalogue systems.

The dual role of metadata in Greenstone reflects the polar extremes mentioned above. The metadata added to documents in the Librarian Interface tends to be lightweight, expressed in a standard, such as Dublin Core or some locally extended version. This Interface is not a full-strength metadata editor. It is intended for full text collections, which have smaller numbers of documents than metadata-only collections, and the metadata you specify is really only for the purpose of helping library users navigate the collection. If you have carefully-prepared, comprehensive, "heavyweight" metadata in a such form as MARC records, you should be using a specialized editor to maintain them, not the Librarian Interface. However, as we have seen you *can* build Greenstone collections of such records that resemble library catalogues.

These two ways of treating metadata in Greenstone operate at different scales. The Librarian Interface is designed for many thousands of documents, which may be very large individually, but it does not have the capability of dealing with hundreds of thousands of individual metadata records. (It has been tested with 10,000 small documents containing individual metadata records; the collection takes about 20 seconds to load into the Librarian Interface on contemporary computers). However, Greenstone can deal with very large collections of metadata (for example, it has been tested on collections of well over 10 million MARC records) – provided they are not broken up into individual metadata records in the Librarian Interface.

UNIT 3.3 Educational metadata

As the previous two units in this module have shown, effective metadata is crucial to the organization of a digital library. It supports the capability of users to find materials through browsing and searching. This unit builds on your knowledge of metadata by examining how to define and use metadata that is specifically educational. We look at how educational metadata is used in real digital libraries and work through practical exercises based on real-world data.

Think back to the two case studies of educational digital libraries we examined in Module 1, Unit 1.2: *Digital Library for Earth System Education* (DLESE) and the *National Library of Virtual Manipulatives for Interactive Mathematics* (NLVM). The designers of these digital libraries organized their resources in particular ways to aid their users. The designs were constrained by the presence, or absence, of appropriate educational metadata.

Providing educational metadata for large libraries of resources used by millions of educators and students is a major organizational challenge. Educational authorities – and governments – have realized that this presents problems that cannot be solved at the level of the individual teacher. Agreed metadata standards are essential for effective national and international deployment of educational digital libraries.

A metadata standard like Dublin Core provides a framework for electronic resources on the Internet in general. What might a metadata standard tailored for educational resources look like? Researchers have worked on this problem for years and several different standards have been proposed.

Today's most important educational metadata standards for learning objects are LOM and SCORM, which we study in detail in this unit. These probably strike you as obscure acronyms – but unfortunately the whole area is awash with confusing acronyms, as you will soon see in Assignment 1.

The LOM standard

The LOM standard provides a means to describe the content of a *Learning Object*, which a combination of resource and pedagogy intended for use in educational settings. LOM provides the metadata foundation for many educational digital libraries.

The standard's full name is *IEEE LTSC LOM*.

IEEE is the Institute of Electrical and Electronics Engineers, a professional body that aids in setting standards in many areas, including computing technologies.
http://ieee.org/

LTSC is the Learning Technology Standards Committee, a group organised by the IEEE to help set standards for educational technology.
http://ltsc.ieee.org/

LOM stands for Learning Object Metadata, a standard developed by the Learning Technology Standards Committee.

The LOM standard will specify the syntax and semantics of Learning Object Metadata, defined as the attributes required to fully/adequately describing a Learning Object. Learning Objects are defined here as any entity, digital or non-digital, which can be used, reused, or referenced during technology supported learning. Examples of technology supported learning include computer-based training systems, interactive learning environments, intelligent computer-aided instruction systems, distance learning systems, and collaborative learning environments.
http://ltsc.ieee.org/wg12/

Assignment 1

Read "Learning technology standardization: making sense of it all" by Erik Duval and then answer the following questions:

- How important are open standards in education?
- Who should be involved in setting educational metadata standards?
- Explain the difference between Learning Objects, Learning Object Metadata (LOM), and Learning Management Systems.
- Must a learning object be in digital form to be accompanied by a LOM record?
- Is it permissible for a learning object to have more than one LOM record?

Discussions of standards can be quite abstract. To make our discussion more concrete, we show a complete LOM record, specified in XML, in the table below. We will use exactly this format of record in the Greenstone example in Assignment 3. After the first few lines, the table places the record into three columns for conciseness, but in actuality it is a single long file. Some items are shown in bold to help you follow through the description below; this boldface is *not* part of the LOM record.

The initial <**lom**> tag includes some XML specifications pertaining to namespaces. Following that, the record contains six top-level sections: *classification, general, lifecycle, educational, rights*, and *technical*; these broadly group the metadata that the record supplies. The LOM standard includes three further top-level categories: *meta-metadata, relation,* and *annotation*. None of the nine categories are compulsory.

The resource that this metadata record describes is entitled *A Teddy Bear Picnic*: this is stated in the <**general**> section as the <**title**> tag. The <**classification**> section gives its subject classification as *Arts*. The LOM standard allows information to be provided in more than one language if desired; in this case the overall language of the resource is set to English through the <**language**> tag. Individual <**langstring**> tags could override this by including an attribute specification, but in this case they do not: they all specify English information. Further down the <**general**> section is a description of the resource: *This film is about a teddy bear trying to enjoy his picnic. Until his sandwich runs away ... now he has to get it back!* *(see p.61)*

The resource that this particular LOM record describes is a film, a fact that is prescribed in machine-readable form in the <**technical**> section. Here we see it is a QuickTime video that is 1 minute 38 seconds long, requiring 2053707 bytes (2 Mbytes approx.) of file space. LOM allows a URL to be specified that points to an online version of the resource. The <**lifecycle**>, <**educational**>, and <**rights**> sections of this example presents details, such as the name of the contributor, *Dalvin Chung Trieu*, the copyright owner, the *Vancouver Film School*, and that the status of the work is *Finished*.

The LOM standard is designed to cover a wide variety of situations. It is possible to use a subset of LOM and still remain compliant with the standard. Such subsets are called *application profiles*. Although LOM itself does not specify whether any elements are compulsory or optional when describing an educational resource, application profiles often stipulate such requirements.

Assignment 2

Read "Building educational metadata application profiles" by Norm Friesen, Jon Mason, and Nigel Ward and then answer these questions:

- In which year was the IEEE LTSC LOM standard approved?
- What are the key differences between the two case studies of application profiles in the paper?
- What are the key similarities between them?
- How do the case studies relate to your organization's needs?

```
<lom xmlns="http://linux2.commons.ucalgary.ca/rob_text/ims1_2_1"
xmlns:xsi="http://www.w3.org/2001/XMLSchema–instance"
xsi:schemaLocation="http://linux2.commons.ucalgary.ca/rob_text/ims1_2_1.xsd">
<classification>
 <keyword>
  <langstring>
   Arts
  </langstring>
 </keyword>
 <purpose>
  <langstring>
   Discipline
  </langstring>
 </purpose>
</classification>

<general>
 <language>
  English
 </language>
 <title>
  <langstring>
   A Teddy Bear Picnic
  </langstring>
 </title>
 <keyword>
  <langstring>
   teddy bear
  </langstring>
 </keyword>
 <keyword>
  <langstring>
   picnic
  </langstring>
 </keyword>
 <keyword>
  <langstring>
   run
  </langstring>
 </keyword>
 <keyword>
  <langstring>
   sandwhich
  </langstring>
 </keyword>
 <keyword>
  <langstring>
   run away
  </langstring>
 </keyword>
 <coverage>
 </coverage>
 <description>
  <langstring>
  This film is about a teddy bear trying
  to enjoy his picnic. Until his sandwich
  runs away ... now he has to get it back!
  </langstring>
 </description>
</general>

<lifecycle>
 <version>
 </version>
 <status>
  <langstring>
   Finished
  </langstring>
 </status>
 <contribute>
  <centity>
   <vcard>
    Dalvin Chung Trieu
   </vcard>
  </centity>
  <date>
   <datetime>
    2002–02–01
   </datetime>
   <description>
    <langstring>
  Born and raised in B.C. Dalvin grew up
  having a knack and interest in art since
  grade 3. He went to study Fine Arts at
  Kwantlan College for two years before
  enrolling and completing the classical
  animation program at VFS. He was one
  of the top students in his class and got
  the certificate of excellence. Then he
  went on to further his studies to complete
  the Maya program at VFS. His future
  endeavor is to get into the video gaming
  industry.
    </langstring>
   </description>
  </date>
  <role>
   <langstring>
    Director/Creator
   </langstring>
  </role>
 </contribute>
</lifecycle>

<educational>
 <context>
  <value>
   <langstring>
    VFSdept
   </langstring>
  </value>
  <source>
   <langstring>
    3D Animation/Maya
   </langstring>
  </source>
 </context>
 <language>
  English
 </language>
</educational>

<rights>
 <description>
  <langstring>
  Copyright 2002 Vancouver Film
  School
  </langstring>
 </description>
 <cost>
 </cost>
 <copyrightandotherrestrictions>
  <langstring>
   yes
  </langstring>
 </copyrightandotherrestrictions>
</rights>

<technical>
 <size>
  2053707
 </size>
 <duration>
  1:38
 </duration>
 <location>
 http://tree.vfs.com:8080/aloha/uploads/
 rbrayer/DalvinT.mov
 </location>
 <format>
  video/quicktime
 </format>
</technical></lom>
```

One of the e-learning projects that formed the context to the CanCore initiative described in the Assignment 2 reading is CAREO (Campus Alberta Repository of Educational Objects) (in fact, it is mentioned in Section 4 of the reading). It is representative of the kind of educational repositories that are becoming available over the web. Its goal is the creation of a searchable, web-based collection of multidisciplinary teaching materials for educators across the province and beyond. From http://www.careo.org

The repository contains over 4,000 items, covering the arts, business, education, engineering, law, and science. The project harnesses the combined expertise of its distributed user base through an online submission procedure. It uses the LOM standard to represent metadata. Anyone can use the index, but only project members can submit records and access the full functionality of the system. (You can register as a member online.)

The CAREO project developed a web interface that supports simple and advanced searching, as well as browsing. It contains a workspace area where members can view and review their personal profile, collate repository items of interest located through searching and browsing, and add new items.

An alternative would be to use standard digital library software and we now examine a small Greenstone collection of LOM records.

This collection is a sample excerpt of educational resources from the University of Calgary's Learning Commons Educational Object Repository. Taken from the subject areas of the arts and science, 40 items from the repository were exported in the IEEE LOM (Learning Object Metadata) format and digested into a Greenstone collection. From *About this collection* on the collection's home page

Learning object repositories allow users to search and browse metadata records. This demonstration goes one step further: Greenstone provides full-text indexing of all the online resources that contain text. Users can browse around the collection's items on any particular subject with titles sorted alphabetically or view the items chronologically. They can search the text or titles of the items in the collection, optionally restricted to arts or science. When an individual item is reached, various views are provided. You start with the LOM record in tabulated form, divided into sections that can be expanded or contracted to reveal more or less information. Tabs at the top change the view of the learning object: one for "Metadata XML" displays the metadata in its original LOM format. If the learning object references a suitable type of online resource, there will be a third tab that displays the source document.

Assignment 3

In a web browser, go to the home page of your Greenstone installation and explore the "LOM Demonstration" collection identified by the IEEE LOM icon using its search and browse features. Then answer these questions:

- How many items are in the collection and in the Arts and Science categories separately?
- Does this mean there are any items in the collection that do not define a title? (General.Title in LOM parlance)
- How many items do not specify date information?

The SCORM standard

Now we examine the other prominent standard for learning objects, the Sharable Content Object Reference Model (SCORM). It was created by the Advanced Distributed Learning Initiative, which is a collaborative effort between U.S. government, industry, and academia to establish a new distributed learning environment that permits the interoperability of learning tools and course content on a global scale.

SCORM aims to foster creation of reusable learning content as "instructional objects" within a common technical framework for computer and web-based learning. From http://www.adlnet.org/index.cfm?fuseaction=scormabt

In essence, SCORM acts as a wrapper to various components specified using existing standards – one of which is a particular profile of the LOM standard.

To help clarify the relationship between LOM and SCORM, consider the following extract, which relates to CanCore. As we learned in Assignment 2, CanCore is an application profile for LOM.

The relationship between CanCore [read: LOM] and the Sharable Content Object Reference Model (SCORM) is not one that can be captured through a simple one-to-one comparison. SCORM, in its own words, "is a reference model that references a set of interrelated technical specifications and guidelines designed to meet the [Department of Defence's] high level requirements for web-based learning content" (SCORM Version 1.1). CanCore, on the other hand, references only one standard – the Learning Object Metadata standard – to meet the needs of educators and educational technologists in Canada and elsewhere. (The additional technical specifications referenced by SCORM include the IMS Content Packaging Specification, and the AICC [Aviation Industry CBT Committee] CMI Data Model, and the AICC CM1001 Interoperability Guidelines.) CanCore and SCORM are related and thus can be compared only in terms of their use of the Learning Object Metadata standard. From http://www.cancore.ca/scorm.html

This explanation is swamped by acronyms. You will learn their meanings in the following Assignment.

Assignment 4

Your CD-ROM for this course contains a tutorial on SCORM. To access it, click on the file SCORM_tutorial\viewer.htm. Work through this and then answer the following questions:

- What are the three components to the runtime system, and what do they do?
- What are the three components to the content aggregation model, and what do they do?
- What are the three elements to the content model, and what do they do?
- What is the purpose of content structure in the content aggregation model?
- What are the categories of the SCORM metadata information model? Comment on their relationship with LOM.
- The SCORM tutorial is an example of an online teaching resource and you have just taken on the role a student to learn about this standard. How useful did you find the tutorial? Was it pitched at the right level for you?
- Using the template below, which is based on the web form used to submit LOM metadata about educational resources to the Careo project's repository, provide metadata that describes the SCORM tutorial. (You may not need to complete every field.) Look back at the Teddy Bear Picnic LOM example to see how some of the terms are used.

Form for submitting LOM metadata to Careo

General	
Title	
Description	
Language	
Coverage	
Keyword	
Aggregation Level	
Structure	
Educational	
Language	
Description	
Typical Learning Time	(description)
	(time)

Difficulty	(value)
	(source)
Typical Age Range	(value)
	(source)
Intended User Role	(value)
	(source)
Semantic Density	(value)
	(source)
Interactivity Level	
Learning Resource Type	(value)
	(source)
Interactivity Type	(value)
	(source)
Context	(value)
	(source)
Classification	
Purpose	
Keyword	
Description	
Lifecycle	
Version	
Status	
Contributor	(role)
	(entity)
	(date)
	(description)
Technical	
Location	
Size	
Format	
Duration	
Requirement	(type)
	(name)
	(source)
	(max. version)
	(min. version)
Other Platform Requirements	
Installation Remarks	
Relation	
Resource	(catalogue entry)
	(description)
	(identifier)

Kind	(source)
	(value)
Rights	
Description	
Copyright (*yes or no*)	
Cost (*yes or no*)	

Not only is the SCORM tutorial an example of an online teaching resource, it is in fact packaged as a Sharable Content Object (SCO) using SCORM (!) and can therefore be imported into any learning management system that supports the standard, such as the standalone Reload SCORM player (www.reload.ac.uk), or a web based learning environment, such as Moodle (www.moodle.org) where users access the resource over the web (both are open source systems). Accessing the tutorial in such an environment provides a richer experience, where the full functionality of SCORM can be exploited. For instance, modules can be constrained to ensure that students complete the necessary prerequisites before moving on to another part of the course or the learning experience can be supported by a teacher who monitors progress online and responds to student requests.

We have been able to view the tutorial without the aid of such software because packaged within the SCO is a set of web pages that can be viewed with a regular browser to provide the basic educational content of the resource. We have extracted these files from the SCO and placed them on the course CD-ROM, giving its home page (*viewer.html*) as the starting point for the above assignment. To be more precise, the tutorial is published as a SCO formed from a zipped up set of files – this is the usual currency in which SCORM learning objects are shared online. On the CD-ROM we have unzipped the *complete* SCO for the tutorial since it is instructive, from the point of view of learning about the SCORM standard, to peruse the files it contains. Of particular interest is the file *imsmanifest.xml*, an abridged version of which is shown below.

```
<manifest identifier="SCOurse–RELEASE–2004–9–24" version="1.5"
 xsi:schemaLocation="http://www.imsproject.org/xsd/imscp_rootv1p1p2
 imscp_rootv1p1p2.xsd
 http://www.imsglobal.org/xsd/imsmd_rootv1p2p1
 imsmd_rootv1p2p1.xsd
 http://www.adlnet.org/xsd/adlcp_rootv1p2 adlcp_rootv1p2.xsd">
 <metadata>
  <schema>ADL SCORM</schema>
  <schemaversion>1.2</schemaversion>
  <lom>
   <general>
    <title>
     <langstring xml:lang="en–US">Academic ADL Co–Lab SCOurse
     </langstring>
    </title>
    <description>
     <langstring xml:lang="en–US">
      The Academic ADL Co–Lab's SCOurse project is a
      Collection of Sharable Content Objects (SCOs) intended
      to help people and institutions learn about ADL's
      Sharable Content Object Reference Model.
     </langstring>
    </description>
    <keyword>
     <langstring xml:lang="en–US">SCORM course</langstring>
    </keyword>
<!–– additional keywords defined ––>
   </general>
   <rights>
    <cost>
```

```
        <source>
          <langstring xml:lang="x–none">LOMv1.0</langstring>
        </source>
        <value>
          <langstring xml:lang="x–none">no</langstring>
        </value>
      </cost>
<!–– … ––>
    </rights>
  </lom>
 </metadata>
 <organizations default="TOC1">
  <organization identifier="TOC1">
   <title>Academic ADL Co–Lab SCOurse</title>
   <item identifier="ITEM1" identifierref="SCO1_1_1">
    <title>
     Overview of the Advanced Distributed Learning Initiative
    </title>
   </item>
   <item identifier="ITEM2" identifierref="SCO1_2_3_a">
    <title>
     Specifications, Standards, and the SCORM®
    </title>
   </item>
        <item identifier="ITEM3" identifierref="SCO1_2_3_b">
    <title>
      Standards Evolution
    </title>
   </item>
<!–– items up to ITEM13 defined … ––>
  </organization>
 </organizations>
 <resources>
  <resource identifier="SCO1_1_1" adlcp:scormtype="sco"
        type="webcontent" href="1_1_1/index.htm">
   <metadata>
    <schema>ADL SCORM</schema>
     <schemaversion>1.2</schemaversion>
     <adlcp:location>1_1_1/1_1_1.xml</adlcp:location>
   </metadata>
   <file href="1_1_1/images/academiccolab.gif"/>
   <file href="1_1_1/images/ADLlogo_med.gif"/>
   <file href="1_1_1/images/colabimage.gif"/>
   <!—– further images ––>
   <file href="1_1_1/LICENSE/index.html"/>
   <file href="1_1_1/media/ADLobjectives.swf"/>
   <file href="1_1_1/scripts/sco_info.js"/>
   <file href="1_1_1/1_1_1_P1.htm"/>
   <file href="1_1_1/1_1_1_P2.htm"/>
   <!–– further html pages ––>
   <file href="1_1_1/index.htm"/>
   <dependency identifierref="common"/>
  </resource>
  <resource identifier="SCO1_2_3_a" adlcp:scormtype="sco"
        type="webcontent" href="1_2_3_a/index.htm">
    <!–– continues in a similar vein ––>
  </resource>
  <!–– … ––>
 </resources>
</manifest>
```

This file is an inventory of all the individual files that constitute the resource, along with how they are pulled together to form the learning object. As discussed in the SCORM tutorial, three keys areas (starting at the bottom of the file and working up) are: *resources*, which are the building blocks out of which the learning object is formed; *items*, which draw upon the resources and stipulate their organizational structure; and object *metadata*, expressed using the IEEE LOM standard. In this particular *imsmanifest.xml* file the relationship between items and resources is essentially flat, making it rather simplistic. Items are usually organized hierarchically and resources can reference other SCOs. It is also possible for the manifest file to include submanifests, although the facility has not been utilized in this example.

Having developed your own LOM metadata for this resource, you might like to compare it with the metadata specified in *imsmanifest.xml* as it is interesting to see what the authors of the resource came up with.

Software tools exist to help in the formation of SCORM content, for instance the aforementioned Reload team (Reload is short for Reusable eLearning Object Authoring and Delivery) also produce a metadata, content packaging, and learning design editor.

If you have been wondering how it is that the SCORM tutorial appears to exhibit SCORM runtime behaviour even though you accessed it outside a learning management system – for example, some quizzes involve dragging and dropping items into their correct placeholders and upon completion a round of applause is played – this is because the designers of the resource have made use of the ShockWave plug-in. (Your browser must include the ShockWave plug-in otherwise such pages will not work as intended.) Embedded in such pages is HTML syntax that activates the plug-in with a ShockWave file that provides the glitzy behaviour. If you reload a quiz page, for example, half way through completion it resets itself, whereas using the run-time capabilities of the SCORM standard it is possible to design a resource such that a learning management system has can track a student's progress through the quiz.

Future directions

The variety and richness of the materials that can be used for educational purposes means that, despite the comprehensive and complex nature of current standards, much more work needs to be done before standards like LOM can be regarded as stable. Educational metadata standards are still evolving quickly compared to more established standards like HTML or email. However, they are changing within a world of versioning and namespaces that should enable software to cope with records from multiple versions of standards. The reading in the next assignment formulates some ideas for research on learning objects and their use in education and training, focusing on issues that relate to metadata: thus it discusses the possible future of educational metadata standards.

Assignment 5

Read "A LOM research agenda" by Erik Duval and Wayne Hodgins and then answer these questions:

- Which of the research issues identified in the paper are most important?
- How important are metadata registries to LOM interoperability?
- Would you need specialized software to work with LOM data in your organization?

Unit closing

This module has shown how metadata formats can be used to structure information. The metadata standards introduced in Unit 3.2 – principally MARC and Dublin Core – are intended to describe general resources. The present unit has introduced the metadata standards LOM and SCORM, which are designed to capture educational aspects of the resources they describe.

Educational content varies across a wide spectrum, from a pre-school picture book to a video explaining quantum Physics. Consequently educational metadata standards are complex compared to Dublin Core. Different stakeholders have collaborated to produce standards that can represent their concerns, and as usual with joint ventures the resulting

structure tries to accommodate everyone. However, because the metadata is represented in the open format XML, it is easy to write software to process metadata records. In Greenstone, for example, the LOM plug-in builds on existing technology for processing XML files.

Some organizations have responded to the complexity of LOM by specifying an application profile that is a subset of LOM. These subsets are consistent with the standard, which is a good sign and portends that developments in this area will remain based on open standards. This should enable any digital library software to be able to import educational metadata and allow users to browse and search meaningfully across educational categories.

Most computer users do not appreciate the significance of standards. As far as they are concerned, the various applications they use just work together. However, programs, such as email clients and web browsers, are totally reliant on the existence of common standards. The same is true of conventional library catalogues and doubly so for digital libraries. The rich structures that systems like DLESE offer to their users are built on a complex infrastructure of standards. Understanding these standards is an important step towards building educational digital libraries that meet users' needs.

MODULE 4 MULTIMEDIA DIGITAL LIBRARIES

Goal

To understand the principles behind common multimedia data formats, to be able to work with multimedia documents, and build digital library collections containing multimedia, to be able to perform advanced customisation of collections, and to have some insight into the mechanisms underlying the Greenstone digital library software.

Objectives

Upon completion of Module 4, you will be able to:

- describe the different ways in which images can be stored digitally;
- explain artefacts that arise when converting between image formats;
- describe the difference between lossy and lossless compression;
- compare and contrast different formats for audio and video;
- recommend image, video, and audio formats that would be suitable for use in digital libraries under various different circumstances;
- describe the nature of the Open Video digital library;
- discuss different ways of browsing video material;
- build collections of heterogeneous multimedia documents in Greenstone;
- customize a Greenstone collection in many different ways;
- explain each of Greenstone's documented example collections;
- describe how Greenstone imports documents and builds a collection;
- navigate around the Greenstone file hierarchy;
- explain the purpose of plug-ins, classifiers, and format statements;
- explain how Greenstone generates web pages;
- identify what is stored in Greenstone's collection information database and in the various log files it maintains.

Introduction to Module 4

Educational digital libraries often contain multimedia documents. This is because teaching material routinely incorporates multimedia elements: textbooks, diagrams and pictures, slide presentations, online presentations, educational videos, and audiovisual recordings of lectures. Indeed, the prevalence of multimedia documents often distinguishes educational libraries from other scholarly libraries.

Module 1 (Unit 1.1) introduced several alternative conceptions of a digital library, including one in the textbook that began "A focused collection of digital objects, including text, video, and audio..." A key challenge for the future is to integrate objects in all kinds of media into digital libraries in such a way that each becomes a first-class citizen. In today's digital libraries textual documents are paramount; other media types are supported in a secondary way as multimedia files associated with textual metadata that provides the main basis for searching and browsing. Tomorrow's digital libraries will meet the multimedia challenge by incorporating ways of searching and indexing documents in all media, techniques for automatically providing multimedia document summaries and classifying multimedia content, and imaginative facilities for browsing through multimedia document collections.

But do not wait until tomorrow: we want you to start building educational digital libraries today! This module's first unit introduces a wealth of practical background information about how to represent pictures, audio, and video in computer form. The second is an extended practical Greenstone exercise: to build a collection that contains multimedia elements of sound and pictures as well as textual and bibliographic documents. The third is not specifically about handling multimedia documents, but introduces several advanced features of Greenstone that will help you build different types of collections and understand what is going on behind the scenes when you use Greenstone.

Module 4 Readings

Unit 4.1. Multimedia formats and standards

Reading 1 CD–ROM IMARK: Unit 2 – Electronic documents and formats, Lesson 2.3: Formats of electronic pictures.

Purpose An introduction to digital images, including their main characteristics, the distinction between bitmap and vector formats, and issues of conversion between different formats.

Reading 2 The textbook *How to build a digital library*: Sections 4.5, 4.6, and 5.5.

Purpose Section 4.5 describes details of the most popular image formats that you encounter when building practical digital libraries. Section 4.6 covers various formats for video and audio recordings. Section 5.5 discusses metadata standards for multimedia.

Reading 3 Course CD–ROM, Readings: "The Open Video digital library", Gary Marchionini and Gary Geisler, *D–Lib Magazine*, December 2002, Vol. 8, No. 12.

http://www.dlib.org/dlib/december02/marchionini/ 12marchionini.html

Purpose The Open Video digital library is a U.S.–based research project on video libraries that makes its entire collection freely available by use for others.

Unit 4.2. Building heterogeneous collections – No readings

Unit 4.3. Getting the most out of Greenstone

Reading 1 CD–ROM IMARK: Unit 6, Lesson 4.

Purpose This describes the "collection configuration file" that underlies all Greenstone collections and introduces some of the ways in which collections can be customized.

Reading 2 The textbook *How to build a digital library*: Sections 6.1 and 6.4–6.7.

Purpose All your interaction with Greenstone so far has been through the Librarian Interface. Underneath this lies the raw functionality of the system, and these sections tell you about some of it, such as the files and directories involved, the importing and building process, and how documents are stored inside.

Reading 3 The textbook *How to build a digital library*: Sections 7.1, 7.2, and 7.4; 7.3 is recommended as an enrichment item.

Purpose More about the internals of Greenstone: in particular, what happens underneath when readers interact with the system.

UNIT 4.1 Multimedia formats and standards

Libraries are about literature. Our dictionary defines literature as "the writings of a society, in prose or verse", and goes on to add, "Broadly speaking, literature includes all types of fiction and nonfiction writing intended for publication". This seems to be firmly dependent upon writing. But in New Zealand, which is where the Greenstone software emanates from, the earliest cultural tradition is that of the Māori, whose literature consisted of history, tales, poems, and myths handed down by oral tradition. It was only when missionaries arrived from Europe that these were written down and became "literature" by the dictionary definition. We have much to learn from early cultures. One trend in modern writing, for example – particularly Australian and North American – is to draw on the oral story-telling traditions of aboriginal cultures.

Assignment 1

Think about a particular course you have taught or are familiar with:

- Does it include any multimedia information resources?
- How could each student's learning experience be enriched if they had access to selected existing multimedia information resources?
- A quiet revolution is taking place with the advent of low-cost consumer technology for image and movie capture. Given appropriate resources for creating multimedia material, how could each student's learning experience be enriched if the instructor was able to prepare special-purpose multimedia information resources?
- Would students benefit from being able to build their own collection of multimedia resources relevant to the course?
- What might a unified multimedia information resource for the course look like?

To work with multimedia documents you need to know something about how multimedia media is represented in the computer and what the common formats for it are. There is less variation in the basic representation of the data than there is for textual documents. But because of the raw, quasi-analogue nature of these media, file size bloats quickly and so compression schemes are often built into the formats. As well as simply making files smaller, compression has side effects that need to be considered when designing digital libraries.

In the next assignment you will learn about the GIF, PNG, and JPEG formats. The first two are suitable for representing artificially produced images, such as text, computer-generated artwork, and logos. JPEG is designed for continuous-tone images, such as photographic portraits and landscapes.

Assignment 2

Work through IMARK Lesson 2.3 and answer the associated questions. Then read Section 4.5 (pp. 194–206) of the textbook and answer the following questions:

- What is the difference between lossy and lossless compression of images?
- In what ways is the PNG standard for lossless image compression an advance over GIF?
- Laying technical considerations aside, which is more appropriate for open source digital libraries: GIF or PNG? Why?
- When you convert a GIF or PNG image to JPEG, it sometimes goes a bit fuzzy. Why? What happens when you convert it back again?
- Many digital libraries of photographs keep their digital source material in two forms: a lossless format and a lossy one. Why? What do you think the purpose of each is? Which version would be shown to a user in web pages? What particular image standard would be suitable for it?

Multimedia encompasses both video and audio formats. The next assignment focuses principally on the MPEG standard, which includes the MP3 scheme that is widely used for music representation. It also mentions Apple's QuickTime and Microsoft's AVI format for multimedia and WAV, AIFF, and AU for audio.

Assignment 3

Read Section 4.6 (pp. 206–216) of the textbook and then answer the following questions:

- Which variant of the MPEG standard was designed for use over the web, MPEG-1, MPEG-2, or MPEG-4?
- Which would you recommend for multimedia information in an open digital library system: MPEG, QuickTime, or AVI?
- What is the principal advantage of RealVideo and RealAudio over earlier multimedia playback schemes?
- What is the relationship between the MP3 and MPEG standards?
- Name two important advantages of MP3 over WAV, AIFF, and AU for audio information.

None of the above-mentioned standards are completely open. Typically, decoders are free but encoders may involve payments. For example, people who distribute music commercially in MP3 format must pay a license fee. In response to this, the open source community is developing alternative multimedia standards, such as the Ogg Vorbis audio format and the Ogg Theora video format, which both share a common metadata wrapper.

Some metadata standards are designed specifically for multimedia material. TIFF is a widely used file structure that accommodates numerous different formats for images and includes a provision for descriptive metadata. MPEG-7 is an emerging standard for describing multimedia documents. You will learn about both in the next assignment.

Assignment 4

Read Section 5.5 (pp. 261–266) of the textbook and then answer the following questions:

- Why is it that a program that reads TIFF images may fail on certain TIFF files?
- Can you use an XML description as the value of a TIFF metadata field? If so, which field or fields?
- Which standard is more complex and comprehensive, TIFF or MPEG-7?
- According to the textbook, which of these is not a potential application area for the MPEG-7 standard: education, journalism, tourist information, cultural services, entertainment, geographical information systems, remote sensing, surveillance, biomedical applications, shopping, architecture, real estate, interior design, film, video and radio archives, dating services?

There are as yet very few digital library systems that are devoted to video. One is the Open Video digital library described in the next assignment. This is of interest both because it identifies challenges – and solutions – that are particular to the video domain, and because it is a potential source of material for the educational digital libraries that you might build.

Assignment 5

Read "The Open Video digital library" by Gary Marchionini and Gary Geisler and then answer the following questions:

- What special challenges does video present for digital libraries?
- How could educational digital libraries benefit from the Open Video project?
- What format or formats is the material in?
- What is keyframe extraction and why is it relevant to video libraries?
- Name three different ways in which summary information about videos can be presented to users.

Unit closing

Multimedia is usually more difficult to work with than text. Generating and editing multimedia is time-consuming and requires special hardware for audio and video capture, and special software for editing. More importantly, creating high-quality multimedia requires special skills for directing, recording, camera work, sound mixing, and so on. Generating metadata and particularly document summaries is more difficult for multimedia than for text. Multimedia information occupies a great deal more disk space than text. Pictures are worth many thousands – perhaps millions – of words in terms of the resources required to store them. Compression is essential, and the most effective compression methods are lossy. This leads to issues of image or sound quality. It is often necessary to maintain separate versions for archival and display purposes – and perhaps a third version to serve as an image thumbnail. Because of the amount of data involved, it can take a lot of computer time to create these different versions.

A different kind of problem, and potentially a far more serious one, is that copyright for multimedia objects on the web tends to be more closely controlled than for textual material. We do not recommend that you create digital library collections from textual documents that you have downloaded indiscriminately from the web without looking carefully into the copyright situation first, but if you do the worst that is likely to happen is that the copyright owner, on discovering it, angrily asks you to take it off the web immediately. But if you create a digital library of video or audio material downloaded from the web, however innocently, you are liable for hefty fines and very aggressive pursuit for "piracy" by, for example, the Motion Picture Association of America. You may even be branded a "terrorist".[3] Fortunately there is a movement towards creating open access multimedia content and making it available, as the Open Video digital library does.

On the other hand, like it or not, contemporary culture is moving away from text and towards multimedia representation of information. As Assignment 1 pointed out, with the advent of low-cost consumer technology for image and movie capture a quiet revolution is taking place in our lives – think camera phones. Now anyone can keep their photograph album on their home computer, or shoot video and use sophisticated editing techniques to produce a professional-quality movie, or make a CD-ROM of their own music (or that of others). Ever since the advent of broadcast television our text-dominated society has gradually become attuned to the more visceral medium of moving images. Book-lovers may deplore the decline of the printed word, laud the sustained argument carefully built up over pages, and praise the power of the written word to conjure up more imaginative and vivid imagery than any TV can. We may lament the decrease in attention span, the reduction of arguments to sound bites. We may wish that our children spent more time reading books, less time playing videogames. But we must live in the world too, and the world is changing.

Multimedia digital libraries will open up information access for all people, regardless of their literacy level – including people from oral cultures. In time, this promises to help reduce the various "digital divides" that cleave our world – the "social divide" between the information rich and the information poor in our own nations, the "democratic divide" between those who do and do not use the panoply of digital resources to engage, mobilize, and participate in public life, as well as the "global divide" that reflects the huge disparity in access to information between people in industrialized and developing societies. Multimedia may be a challenge, but it is certainly a worthy one.

And the good news is that multimedia objects of all types can be incorporated into collections built with currently available digital library software, such as Greenstone. To find out how, move on to Unit 4.2.

[3] "We're fighting our own terrorist war," said Motion Picture Association of America (MPAA) president Jack Valenti, referring to an antipiracy campaign that included raids on college campuses. Amy Harmon, "Black Hawk Download: Moving Beyond Music, Pirates Use New Tools to Turn the Net into an Illicit Video Club", *New York Times*, 17 January 2002.

UNIT 4.2 Building heterogeneous collections

All the digital library collections that we have built so far handle a single document type. In this unit, which consists entirely of an extended practical exercise with the Greenstone software, you will build a complex heterogeneous collection that includes multimedia elements of sound and pictures, as well as textual and bibliographic documents.

Imagine you are a music teacher and would like to provide your students with a comprehensive collection illustrating a particular type of music to use as a resource for their study. You might want to provide audio recordings in MP3 format, some of your own musical examples played into the computer on a MIDI keyboard, discography information in the form of HTML files, auxiliary material, such as reviews in PDF and Word format, library records for relevant items in MARC format, lyrics and guitar tablature as text files, images of album covers in JPEG format, and so on.

To illustrate what we mean, the course CD-ROM contains a small digital library collection containing information centred on the Beatles pop group.

Practical exercise: Look at what can be done!

1. Copy the entire folder

   sample_files→beatles→advbeat_large

   (with all its contents) into your Greenstone *collect* folder. If you have installed Greenstone in the usual place, this is

   My Computer→Local Disk (C:)→
   Program Files→greenstone→collect

   Put *advbeat_large* in there.

2. If the Greenstone Digital Library Local Library Server is already running, restart it by clicking the CD icon on the task bar and then pressing *Restart Library*. If not, start it up by selecting *Greenstone Digital Library* from the Start menu.

3. Explore the Beatles collection. Note how the *browse* button divides the material into seven different types. Within each category, the documents have appropriate icons. Some documents have an audio icon: when you click these you hear the music (assuming your computer is set up with appropriate player software). Others have an image thumbnail: when you click these you see the images.

4. Look at the *titles a–z* browser. Each title has a bookshelf that may include several related items. For example, *Hey Jude* has a cover image, MP3 audio and MIDI versions, lyrics, and a discography item.

5. Observe the low quality of the metadata. For example, the four items under *A HARD DAY'S NIGHT* (under "H" in the *titles a–z* browser) have different variants as their titles. The collection would have been easier to organize had the metadata been cleaned up manually first, but that would be a big job. Only a tiny amount of metadata was added by hand – fewer than ten items. The original metadata was left untouched and Greenstone facilities used to clean it up automatically. (You will find below that this is possible but tricky.)

6. In the Windows file browser, take a look at the files that makes up the collection, in the

   sample_files→beatles→advbeat_large→import folder.

   What a mess! There are over 450 files under seven top–level subfolders. Organization is minimal, reflecting the different times and ways the files were gathered. For example, *html_lyrics* and *discography* are excerpts of web sites, and *cover_images* contains album covers in JPEG format. For each type, drill down through the hierarchy and look at a sample document.

The messy nature of the source information is typical of material gathered from a variety of different sources. Your challenge, as a budding digital librarian, is to take an agglomeration of unorganised information and built it into an attractive, easy to use digital library collection. You could start by providing comprehensive metadata for each item. This would be a big job, and although certainly be worthwhile for archival purposes, it would perhaps be too large an investment of time for a music teacher who just wants to provide students with a useful resource. In fact, a lot can be accomplished with a far smaller amount of effort by exploiting what structure there is in the source information. That is the approach we will take in this unit.

We will proceed to reconstruct from scratch the Beatles collection that you have just looked at. We develop the collection using a small subset of the material, purely to speed up the repeated rebuilding that is involved.

Practical exercise: Building a basic collection

7. Start a new collection (*File→New*) called **small_beatles**, basing it on the default "New Collection". (Basing it on the existing Advanced Beatles collection would make your life far easier, but we want you to learn how to build it from scratch.) Fill out the fields with appropriate information. Use the Dublin Core metadata set (set by default).

8. Copy the files provided in

   sample_files→beatles→advbeat_small

   into your new collection. Do this by opening up *advbeat_small*, selecting the eight items within it (from *cover_images* to *beatles_midi.zip*), and dragging them across. Because some of these files are in MP3 format you will be asked whether to include the **MP3 Plugin** in your collection. Click <**Add Plugin**>.

9. Change to the **Enrich** panel and browse around the files. There is no metadata – yet. Recall that you can double–click files to view them.

   (There are no MIDI files in the collection: these require more advanced customisation because there is no MIDI plug–in. We will deal with them later.)

10. Change to the **Create** panel and **build** the collection.

11. **Preview** the result.

Relying on default settings, as we have done here, works moderately well, but we can improve the organization and presentation of the collection by iterating over the design process. This is the focus of the remaining exercises.

Practical exercise: Hand–correcting metadata

12. You might want to correct some of the metadata – for example, the atrocious misspelling in the titles "MAGICAL MISTERY TOUR". These documents are in the discography section, with filenames that contain the same misspelling. Locate one of them in the **Enrich** panel. Notice that the extracted metadata element **ex.Title** is now filled in and misspelt. You cannot correct this element, for it is extracted from the file and will be re–extracted every time the collection is rebuilt.

13. Instead, add **dc.Title** metadata for these two files: "Magical Mystery Tour". Change to the **Enrich** panel, open the discography folder, and drill down to the individual files. Set the **dc.Title** value for the two offending items.

*Now there's a twist. The **dc.Title** metadata won't appear in titles a–z because the classifier has been instructed to use **ex.Title.** But changing the classifier to use **dc.Title** would miss out all the extracted titles. Fortunately, there's a way of dealing with this by specifying a list of metadata names in the classifier.*

14. Change to the **Design** panel and select the **Browsing Classifiers** section. Double–click the **Title** classifier (the first one) to edit its configuration settings.

    - Type "dc.Title", before the *ex.Title* in the metadata box – i.e. make it read dc.Title, ex.Title.

    Build the collection again and **preview** it.

Extracted metadata is unreliable. But it is very cheap. On the other hand, manually assigned metadata is reliable, but expensive. The exercise above has shown how to aim for the best of both worlds by using extracted metadata but correcting it when it is wrong. While this may not satisfy the professional librarian, it could provide a useful compromise for the music teacher who wants to get their collection together with a minimum of effort.

In the next exercise we tidy up the collection and add some new features.

Practical exercise: Simple customisation

15. First let's remove the **AZList** classifier for filenames, which isn't very useful, and replace it with a browsing structure that groups documents by category (discography, lyrics, audio, etc.). Categories are defined by manually assigned metadata.

    - Change to the **Enrich** panel, select the folder *cover_images* and set its **dc.Format** metadata value to "Images". Setting this value at the folder level means that all files within the folder inherit it.
    - Repeat the process. Assign "Discography" to the *discography* folder, "Lyrics" to *html_lyrics*, "MARC" to *marc*, "Audio" to *mp3*, "Tablature" to *tablature_txt*, and "Supplementary" to *wordpdf*.
    - Switch to the **Design** panel and select the **Browsing Classifiers** section.
    - Delete the **ex.Source** classifier (the second one).
    - Add an **AZCompactList** classifier. Select **dc.Format** as the metadata field and specify "Category" as the **buttonname.**

    Build the collection again and **preview** it.

16. Greenstone has no pre–defined button for "Category", so it appears in the navigation bar as text. It does, however, have a button for *browse* (it's used in the Beatles collection you looked at in Part I).

    - Go back to the **AZCompactList** classifier for **dc.Format.** Click the **sort** checkbox and leave **Title** in the adjacent text box: this will make the classifier display documents in alphabetical order of title. Also, specify "Browse" as the **buttonname.**

    You will need to build the collection for this to take effect.

17. Alongside the Audio files there is an MP3 icon, which plays the audio when you click it, and also a text document that contains some dummy text. This isn't supposed to be seen, but to suppress it you have to fiddle with a format statement.

    - Change to the **Design** panel and select the **Format Features** section.
    - Ensure that **VList** is selected and make the changes that are highlighted below. You need to insert three lines into the first line and delete the second line.

Change:
<td valign=top> [link][icon][/link] </td>
<td valign=top>[srclink]{Or}{[thumbicon],[srcicon]}[/srclink]</td>

```
<td valign=top>[highlight]
{Or}{[dls.Title],[dc.Title],[Title],Untitled}
[/highlight]{If}{[ex.Source],<br><i>([ex.Source])</i>}</td>
```

to this:

```
<td valign=top>
{If}{[dc.Format] eq 'Audio',
  [srclink][srcicon][/srclink],
  [link][icon][/link]} </td>
<td valign=top>[highlight]
{Or}{[dls.Title],[dc.Title],[Title],Untitled}
[/highlight]{If}{[Source],<br><i>([Source])</i>}</td>
```

- Then click <**Replace Format**>.

To make this easier for you we have prepared a plain text file that contains the new text. In WordPad open the following file:

sample_files→beatles→format_tweaks→audio_tweak.txt

(Be sure to use WordPad rather than Notepad, because Notepad does not display the line breaks correctly.) Place it in the copy buffer by highlighting the text in WordPad and selecting Edit→Copy. Now move back to the Librarian Interface, highlight all the text that makes up the current VList format statement, and use Edit→Paste to transform the old statement to the new one. Remember to press <**Replace Format**> when finished.

Preview the result. If you are using the Greenstone Local Library server, change to the **Create** panel and click <**Preview Collection**>, which causes the local library server to rescan the format statements. You do not need to build the collection again because format statements are only used by the runtime system.

However, you may need to click the browser's *Reload* button to force it to reload the page.

18. While we're at it, let's remove the source filename from where it appears after each document.

- In the VList format feature, delete the text that is highlighted below:

```
<td valign=top>
{If}{[dc.Format] eq 'Audio',
  [srclink][srcicon][/srclink],
  [link][icon][/link]}</td>
<td valign=top>[highlight]
{Or}{[dls.Title],[dc.Title],[Title],Untitled}
[/highlight] {If}{[Source],<br><i>([Source])</i>} </td>
```

Don't forget to click <**Replace Format**> after all this work. **Preview** the result (you don't need to build the collection.)

19. There are sometimes several documents with the same title. For example, *All My Loving* appears both as lyrics and tablature (under *ALL MY LOVING*). The *titles a–z* browser might be improved by grouping these together under a bookshelf icon. This can be done with an AZCompactList.
    - Change to the **Design** panel and select the **Browsing Classifiers** section.
    - Remove the **Title** classifier (at the top).
    - Add an **AZCompactList** classifier, and enter **dc.Title,ex.Title** as its metadata.
    - Activate **min_group** and set it to 1. This gives a uniform appearance by creating a bookshelf for every title.
    - Finish by pressing <**OK**>.
    - Move the new classifier to the top of the list (*Move Up* button).

Build the collection again and **preview** it. Both items for *All My Loving* now appear under the same bookshelf. However, many entries haven't been amalgamated because of non–uniform titles: for example *A Hard Day's Night* appears as four different variants. We will learn below how to amalgamate these.

20. Make the bookshelves show how many documents they contain by inserting a line in the VList format statement in the **Design** panel:

```
<td valign=top>
{If}{[dc.Format] eq 'Audio',
  [srclink][srcicon][/srclink],
  [link][icon][/link]}</td>
<td>{If}{[numleafdocs],([numleafdocs])}</td>
<td valign=top>[highlight]
{Or}{[dls.Title],[dc.Title],[Title],Untitled} [/highlight]</td>
```

You will find this text in *format_tweaks→show_num_docs.txt*, which can be copied and pasted in as before. Don't forget to click <**Replace Format**>.

Preview the result (you don't need to build the collection).

21. Now turn to the images. Dummy documents are displayed here, too. First change to the **Enrich** panel, open the folder *cover_images*, and add **dc.Title** metadata, assigning to each of the ten documents the title of the corresponding album. Remember, you can double–click a file to view it.

22. To suppress the dummy documents, change the **VList** format statement in the **Design** panel again by adding the two highlighted lines, and the close curly bracket:

```
<td valign=top>
{If}{[dc.Format] eq 'Audio',
  [srclink][srcicon][/srclink],
  {If}{[dc.Format] eq 'Images',
    [srclink][thumbicon][/srclink],
    [link][icon][/link]}}</td>
<td>{If}{[numleafdocs],([numleafdocs])}</td>
<td valign=top>[highlight]
{Or}{[dls.Title],[dc.Title],[Title],Untitled} [/highlight]</td>
```

23. Add a **Phind** browsing classifier that sources its phrases from Title and text (the default setting).

24. To complete the collection, use the browse button of **URL to 'about page'** icon in the **General** section of the **Design** panel to select the following image: *advbeatles_large→images→flick4.gif*.

Build the collection again and **preview** it.

Note how in the above exercise we assigned *dc.Format* metadata to all documents in the collection with a minimum of labour. We did this by capitalizing on the folder structure of the original information. Even though we complained earlier about how messy this folder structure is, you can still take advantage of it when assigning metadata.

In the next exercise we incorporate the MIDI files. Greenstone has no MIDI plug-in (yet). But that doesn't mean you can't use MIDI files. We also clean up the *titles a–z* browser.

To do this we must put the Librarian Interface into a different mode. The interface supports four levels of user: Library Assistants, who can add documents and metadata to collections and create new ones whose structure mirrors that of existing collections; Librarians, who can, in addition, design new collections, but cannot use specialist IT features (e.g. regular expressions); Library Systems Specialists, who can use all design features, but cannot perform troubleshooting tasks (e.g. interpreting debugging output from Perl programs); and Experts, who can perform all functions.

You have already encountered Expert mode, very briefly, in Unit 2.2, Assignment 5, and Library System Specialist mode in the enhanced Tudor collection exercise of Unit 3.2. Apart from that, up to this point you have been operating in Librarian mode. However, we switch to Library Systems Specialist mode for the next exercise.

Practical exercise: Advanced customisation

To switch modes, click *File→Preferences→Mode* and change to Library Systems Specialist. Note from the description that appears that you need to be able to formulate regular expressions to use this mode fully. That is what we do below.

25. **UnknownPlug** is a useful generic plug–in. It knows nothing about any given format but can be tailored to process particular document types – like MIDI – based on their filename extension, and set basic metadata.

    - Add **UnknownPlug**.
    - Activate its *process_extension* field and set it to *mid* to make it recognize files with extension.mid.
    - Set *file_format* to "MIDI" and *mime_type* to "audio/midi".

    In this collection, all MIDI files are contained in the file *beatles_midi.zip*. **ZIPPlug** (already in the list of default plug–ins) is used to unpack the files and pass them down the list of plug–ins until they reach UnknownPlug.

26. **Build** the collection and **preview** it. Unfortunately the MIDI files don't appear as Audio under the *browse* button. That's because they haven't been assigned **dc.Format** metadata.

    - Back in the **Enrich** panel, click on the file *beatles_midi.zip*, and assign its **dc.Format** value to "Audio" – do this by clicking on "Audio" in the **All Previous Values** list. All files extracted from the Zip file inherit its settings.

27. Next we return to our *titles a–z* browser and clean it up. The aim is to amalgamate variants of titles by stripping away extraneous text. For example, we would like to treat "ANTHOLOGY 1", "ANTHOLOGY 2", and "ANTHOLOGY 3" the same for grouping purposes. To achieve this:

    - Go to the Title **AZCompactList** under **Browsing Classifiers** on the **Design** panel.
    - Activate **removesuffix** and set it to:
      (?i)(\\s+\\d+)|(\\s+[[:punct:]].*)

    Build the collection and **preview** the result. Observe how many more times similar titles have been amalgamated under the same bookshelf. Test your understanding of regular expressions by trying to rationalize the amalgamations. (Note: *[[:punct:]]* stands for any punctuation character.) The icons beside the Word and PDF documents are not the correct ones, but that will be fixed in the next format statement.

The previous exercise was done in Librarian Systems Specialist mode because it requires the use of regular expressions, something librarians are not normally trained in. Learning how to use regular expressions is not part of this course.

One powerful use of regular expressions in the exercise was to clean up the *titles a–z* browser. Perhaps the best way of doing this would be to have proper title metadata. The metadata extracted from HTML files is messy and inconsistent, and this was reflected in the original *titles a–z* browser. Defining proper title metadata would be simple but rather laborious. Instead, we have opted to use regular expressions in the *AZCompactList* classifier to clean up the title metadata. This is difficult to understand and a bit fiddly to do, but if you can cope with its idiosyncrasies it provides a quick way to clean up the extracted metadata and avoid having to enter a large amount of metadata. If (as we expect) your training is in education or librarianship rather than the computer field it is unlikely that you would be able to do this yourself, but it is useful to have an idea of what can be done so that you can call in specialist help if appropriate.

The next exercise involves expert customisation. It is an optional part of this course and you should only attempt it if you are keen and have time.

Enrichment exercise: Expert customisation

28. To put finishing touches to our collection, we add some decorative features.

    - Using your Windows file browser outside Greenstone, locate the folder sample_files→beatles→advbeat_large

Copy the *images* and *macros* folders located there into your collection's top–level folder. (It's OK to overwrite the existing *images* folder: the image in it is included in the folder being copied.) The *images* folder includes some useful icons and the *macros* folder defines some macro names that use these images. To see the macro definitions, take a look by using a text editor to open the file *extra.dm* in the macros folder.

- Re–Edit your **VList** format statement to be the following

```
<td valign=top>
{If}{[numleafdocs],[link][icon][/link]}
{If}{[dc.Format] eq 'Lyrics',[link]_iconlyrics_[/link]}
{If}{[dc.Format] eq 'Discography',[link]_icondisc_[/link]}
{If}{[dc.Format] eq 'Tablature',[link]_icontab_[/link]}
{If}{[dc.Format] eq 'MARC',[link]_iconmarc_[/link]}
{If}{[dc.Format] eq 'Images',[srclink][thumbicon][/srclink]}
{If}{[dc.Format] eq 'Supplementary',[srclink][srcicon][/srclink]}
{If}{[dc.Format] eq 'Audio',[srclink]{If}{[FileFormat] eq
'MIDI',_iconmidi_,_iconmp3_}[/srclink]}
</td>

<td>
{If}{[numleafdocs],([numleafdocs])}
</td>
<td valign=top>
[highlight]
{Or}{[dc.Title],[Title],Untitled}
[/highlight]
</td>
```

The complete statement is in the file *format_tweaks→multi_icons.txt*.

Preview your collection as before. Now different icons are used for discography, lyrics, tablature, and MARC metadata. Even MP3 and MIDI audio file types are distinguished. If you let the mouse hover over one of these images a "tool tip" appears explaining what file type the icon represents in the current interface language (note: *extra.dm* only defines English and French).

We now change the image used as the background for the collection, this time delving a bit deeper into the files:

29. Open your collection's *macros* folder and locate the *extra.dm* file within it. **Right–click** on it. If prompted, select **WordPad** as the application to **open it with.**

30. The file content is fairly brief, specifying only what needs to be overridden from the default behaviour for this collection. In WordPad, near the top of the file you should see:

    ```
    _httpiconchalk_ {_httpcimages_/beat_margin.gif}
    _widthchalk_ {1800}
    _heightchalk_ {68}.
    ```

    Use copy and paste on these three lines to make this part of the file look like:

```
# Original statements
#_httpiconchalk_ {_httpcimages_/beat_margin.gif}
#_widthchalk_ {1800}
#_heightchalk_ {68}
_httpiconchalk_ {_httpcimages_/tile.jpg}
_widthchalk_ {22}
_heightchalk_ {22}
```

A hash (#) at the start of line signals a comment and Greenstone ignores the following text. We use this to comment out the original three statements and replace them with modified lines. It is useful to retain the original version in case we need to restore the original lines at a later date. These three lines relate to the background image used. The new image *title.gif* was also in the *images* folder that was copied across previously.

31. Within **WordPad**, save *extra.dm*.

32. **Preview** the collection's home page. The page background is now the new graphic.

    Other features can be altered by editing the macro files – for example, the headers and footers used on each page, and the highlighting style used for search terms (specify a different colour, use bold, etc.).

33. If you want to you can reverse the most recent change you made by commenting out the three new lines added (add #) and uncommenting the original three (delete # character). Remember to save the file. To undo all the customized changes made, delete the content of the *macros* and *images* folders.

34. To finish, let's now build a larger version of the collection. To do this:

    - Close the current collection.
    - Start a new collection called *advbeat_large*.
    - Base this new collection on *small_beatles*.
    - Copy the content of *sample_files→beatles→advbeat_large→import* into this newly formed collection. Since there are considerably more files in this set of documents the copy will take longer.
    - **Build** the collection and preview the result. (If you want the collection to have an icon, you will have to add it from the **Design** panel.)

Whether or not you made it through the previous enrichment exercise, which was intentionally very challenging, we want you to end on a high note by making a final enhancement to the collection that adds a useful new browsing facilities – and is easy to do. The Beatles collection is rich in photographic and album artwork, which we will exploit by providing an attractive and serendipitous form of visual browsing called "collaging". Successive images chosen at random from the collection will be placed on the screen at random locations. Over time they fade away and new ones appear on top of what is already there, perhaps obscuring what has gone before. This provides a dynamic, eye-catching – perhaps even mesmerizing – screen-saver-like effect which is generated from the images in the collection itself. But it differs from a screen saver because if the user is interested in a particular image and clicks it, the document containing the image is retrieved from the digital library and displayed.

Practical exercise: Fancy customisation

35. Switch to the **Design** panel and select the **Browsing Classifiers** section. Pull down the **select classifier to add** menu and select **Collage.** Click <**Add Classifier**>. There is no need to customize the options, so click <**OK**> at the bottom of the resulting popup.

36. Now change to the **Create** panel and **build** and **preview** the collection.

We have already included the interactive hierarchical phrase index that we met in Unit 3.2 (Enhanced Tudor collection). It is interesting to use this to explore the lyrics to the Beatles songs.

Unit closing

If you managed to successfully work through the practical exercises that make up this unit, congratulations! And if you completed the optional enrichment exercise on expert customisation, double congratulations – you are well on the way to becoming an advanced user of Greenstone. Even if you did not complete all exercises successfully, you will have learned a great deal about how to build digital library collections of heterogeneous material.

Greenstone is a rapidly evolving system. It is far from perfect and improvements are continually being made – that is why, at the end of Unit 2.3, we placed so much stress on making sure you could install new versions of the software and recommended you to keep your installation up to date by joining the Greenstone discussion group (see next unit) where announcements about new releases are made or checking http://www.greenstone.org for new versions. For example, the Librarian Interface is a fairly recent innovation; before that Greenstone was far more difficult to use.

Consequently, if you are having difficulty with the technical detail involved in the practical exercises, do not despair. New innovations will eventually make this kind of work much easier. For example, format statements are arcane and difficult to understand (you will learn more about them in the next unit) – but eventually an interactive editor for format statements will make it as easy to format material in Greenstone as it is in a contemporary word processor. In the last assignment you laboriously altered your collection to use different icons and a different background – but eventually this will be done with an interactive graphical editor.

Meanwhile, the best strategy for building advanced collections is to find an existing collection with the structure you want – or commission an expert to build one – and use that structure for your new collection by employing the Librarian Interface's "Base this collection on" feature that you encountered in the first practical exercise of Unit 3.2.

UNIT 4.3 Getting the most out of Greenstone

In the course so far we have devoted considerable time and energy in learning how to use the Greenstone Librarian Interface. As its name implies, it is an *interface* to the core body of software that is Greenstone. It is this core software that provides the range of functionality we have exploited in the practical exercises, *controlled* through the Librarian Interface. In this unit we look "under the hood" of Greenstone and study this core software, to gain a deeper understanding of how things work.

All the practical exercises you have undertaken were prepared beforehand and carefully scripted so everything went according to plan. When you progress to working with your own documents and designing your own collections, things will not always go so smoothly. You are entering a design process where there will be tough decisions to make and challenging obstacles to overcome. You will have your reward: it is stimulating and satisfying to see your own creation develop and grow. However, you will fare better if you are equipped with knowledge with which you can reason about what is going on underneath.

To continue the "under the hood" analogy, this unit moves you beyond being a car owner (which entails only a passing familiarity with the workings of the engine) and teaches you the skills of a mechanic. As with a faulty vehicle, when a newly built collection doesn't perform as expected the observed symptoms can be combined with an understanding of the mechanics to suggest things that should be checked. What are the usage permissions on this folder, do these particular files exist, and just what exactly is in that file? You will have the confidence to prod particular parts of the system and open up suspect components, just as a mechanic checks the radiator when an engine overheats. As in any faultfinding exercise, you must always bear in mind that problems are often caused by a combination of circumstances rather than by a single fault.

Chapters 6 and 7 of *How to build a digital library* are specific to the Greenstone system and describe the core software. In this unit we work our way through the most important parts of this material.

First we learn how the documented example collections supplied with your Greenstone installation are structured. These are explained from the point of view of the core Greenstone software, not in terms of how the Librarian Interface could be used to achieve this. When working on a collection, all the choices made in the Librarian Interface's design panel result in a *collection configuration file* that is saved to disk and used by the core software when the build button is clicked. This action triggers document importing and building, two phases of the core software, that consult the configuration file for instructions on what structures and indexes must be built to form the target collection.

When studying the documented example collections you will see the raw content of their collection configuration files. You will recognise parts of this from your use of the Librarian Interface. Format statements appear almost verbatim, and the "about this collection text" is signified by this keyword combination:

collectionmeta collectionextra "..."

Assignment 1

Work through IMARK Unit 6, lesson 4 "Documented example Greenstone collections".

- Identify which (if any) documented collections work with source documents in a format that is of interest to you.
- Identify any documented collections that contain features you would like to provide in your own digital library collections.
- Compare the two lists from above. How much do they overlap?

We next read the first of two chapters on Greenstone in *How to build a digital library* (Chapter 6). Given what you have already learned, the first section is largely revision. We skip over the next one, which describes "The Collector", a Greenstone subsystem that gathers details about the collection and the documents that are to be in it, using a sequence

of web-based forms, and then creates, alters, or builds the collection. "The Collector" has been superseded by a Java applet version of the Librarian, which resembles the Librarian Interface except that it uploads files to a designated Greenstone server for building. More details about this can be found in the Greenstone FAQ.

We pick up the thread in Section 6.4, which introduces the two phases involved in building a collection and the structure of the Greenstone files and directories. Section 6.5 details the Greenstone archive format, an XML-based intermediate format used to store documents and metadata between the importing and building phases. Section 6.6 summarizes the structure of collection configuration files, and Section 6.7 provides technical detail about plug-ins, classifiers, and format statements.

Assignment 2

Read Section 6.1 and Sections 6.4–6.7 of the textbook. To follow this material, use your file manager to locate Greenstone (C:\Program Files\greenstone\ on a typical Windows installation) and, like the mechanic, "prod" some of those files. In normal circumstances every file in this area is read/write accessible by you and can be opened with a text editor, such as NotePad, WordPad, or Emacs. Some files are binary and although your editor may open them they will not be comprehensible. If you open such a file, or indeed any other file (unless you are sure about what you are doing), quit the editor without saving to avoid accidentally changing the system.

Now answer the following questions:

- Explain in broad terms the functions of the two stages in building a collection.
- Give a detailed account of what happens in each stage when a collection is built.
- Where do collections reside in the file hierarchy?
- What directories are formed when a new collection is created? Describe the purpose of each one.
- Explain what is meant by a "persistent document identifier".
- In terms of files generated, what is the end result of the import process?
- What is the default indexing tool used by Greenstone?
- In Figure 6.10, what is the top-level title metadata for the document?
- Compile by hand the table of contents for the document shown in Figure 6.10 (Section 1, Section 1.1, and so on), showing the title metadata associated with each section, along with its section number.
- Explain the purpose of plug-ins in Greenstone.
- Explain the purpose of classifiers.
- Explain the purpose of format statements.

Things move quickly in open source projects. Back in 2003 when the textbook was published, the Librarian Interface was nothing more than a fledgling prototype. Near the end of Chapter 6 you might have noticed a brief passage on building collections graphically (Section 6.8). The software tool mentioned there has evolved into the Librarian Interface used in this course. The uptake of the Librarian Interface has been so successful that it is what many users think of when the Greenstone software is mentioned.

Greenstone can be logically divided into two parts: building and runtime. The last reading focused of the building side. Understanding these components and how they operate is a significant part of forming a collection, but not the complete picture. Strictly speaking, format statements are part of the runtime side, which is why changes to them appear immediately, but they are so ubiquitous that they are introduced in Chapter 6. We revisit them shortly. Broader issues of presentation – such as handling collection-specific icons – are handled at runtime, and can also be controlled through the Librarian Interface. Both runtime and build-time information is contained in the same collection configuration file – the respective parts of the system ignore entries that do not relate to them. The building component of Greenstone is written in the Perl programming language; the runtime part is in C++.

Chapter 7 of *How to build a digital library* focuses on the runtime side of Greenstone and explains how it works. It is intended for those who want to step outside the ambit of the Librarian Interface and manipulate elements of the file hierarchy directly – possibly even modifying source code and recompiling it – to achieve radical changes in the software. We study this in the next assignment.

Only modest modifications – "tweaks", if you will – are possible using format statements. They are like changing the colour of the car or adding a go-faster stripe. The technical information in Chapter 7 equips you to modify the engine and reconfigure the bodywork. Armed with it you could change your car into a tractor – it's that powerful. The Kids Digital Library described in Chapter 7 is in this league. It extends Greenstone to include personal profiles for pupils, who must first log in; a special teacher account with additional capabilities; workspace for students to compose and submit their stories; and a bulletin board for discussion. Collections are served from two different sites and seamlessly presented to the user as a single unified resource.

The chapter explains two broad components to the runtime system: client and server. The backbone to the runtime system is a protocol that links the two and enables the kind of behaviours exemplified by the Kids Digital Library. The division separates presentational issues on the client side from delivering the collection's content by accessing database and index files on the server side. Macro files provide a presentation mechanism that is language independent. The text also describes what is stored in the database files and how the "action" mechanism works and discusses log files and site-wide configuration.

Assignment 3

Read the following excerpts from Chapter 7 of *How to build a digital library*: the beginning, Sections 7.1, 7.2 and 7.4. Section 7.3 (on actions) is optional and left as an enrichment activity.

- Explain what the null protocol is in Greenstone.
- Write macro definitions that achieve the following:

  a macro called _poweredby_ that displays the message "powered by Greenstone" by default, "actionné par Greenstone" if the user interface language is French, and "accionado por Greenstone" if Spanish.

- What sort of information is stored in the collection information database?
- What kind of information is stored in the site-wide log file?
- Where should you look for runtime errors in Greenstone?

For many users, format statements are the bane of Greenstone. They contain small excerpts of raw HTML, augmented by special, tricky, Greenstone commands. If so much as a single special character – such as a "/" in a closing HTML tag, or an a stray comma in an {*If*} statement – is misplaced or missing, the excerpt often fails to display properly, and in extreme cases displays nothing at all. Here are some useful format statements to assist you in forming commonly used constructs:

Basic control over *VList* document items:

```
<td>[link][icon][/link]</td><td><i>[Title]</i></td>
```

Adding a reference to original source document to *VList*:

```
<td>[srclink][srcicon][/srclink]</td>
<td>[link][icon][/link]</td><td><i>[Title]</i></td>
```

Adding an image thumbnail to *VList*:

```
<td>[link]<img
        src='_httpcollimg_/[assocfilepath]/[Thumb]'
        width=[ThumbWidth] height=[ThumbHeight]>
        [/link]</td>
<td valign=middle><i>[Title]</i></td>
```

Adding node information for a hierarchy classifier:

```
<td>[link][icon][/link]</td>
<td>{If}{[numleafdocs],[Title] <i>([numleafdocs])</i>,[Title]}
  </td>
```

All of these have been encountered in one form or another in the practical exercises. They are grouped here for your convenience.

The Greenstone community

In Unit 1.3 you learnt that Greenstone is open source software and that anyone can look at Greenstone's source code. As with many open source projects, Greenstone has an international community of developers who help add new features and fix bugs. There is also a community of Greenstone users who help each other solve problems and come up with new ways to use the software. These communities are excellent resources for finding solutions for problems you may encounter with Greenstone. First, however, we recommend that you familiarize yourself with the sources described below.

Program help. The Librarian Interface has a help system, which you access from the *Help* menu, that contains much useful information. Also, when you leave your cursor over most buttons and text fields in the interface, textual tips pop up to give you on-the-spot assistance.

Manuals. There are several Greenstone manuals on the CD-ROM.

- *Installer's Guide*

Describes in detail the Greenstone installation process. Note that the Installer's Guide assumes that Greenstone is being installed from a CD-ROM distribution. The instructions should be adapted in the obvious way when installing from a web download.

- *User's Guide*

General details on using Greenstone collections, the Librarian Interface, and Greenstone's administrative facilities.

- *Developer's Guide*

A more detailed description of Greenstone's collection building process, including building collections from the command line or DOS prompt. Also describes the structure of the Greenstone runtime system.

- *From Paper to Collection*

This document describes the entire process of creating a digital library collection from paper documents, including the scanning and OCR process.

Other documentation

- *Inside Greenstone Collections*

One of the trickier parts of using Greenstone is coming up with a configuration file for your collection. To help learn how to do it, this document presents and explains the configuration files for a few actual Greenstone collections and also gives an example of how Greenstone's appearance can be customized.

- *Documented Example Collections*

This package contains 11 documented example Greenstone collections whose "about" page describes how they are constructed. They are fully documented in English, French, Spanish, and Russian and are an excellent resource for learning how to build common types of collections. Also, by choosing in the Librarian Interface to base a new collection on one of these collections, you inherit the style and formatting of the collection without having to recreate it.

The textbook that accompanies this course contains much useful material on Greenstone:

How to build a digital library, by Ian H. Witten and David Bainbridge. Morgan Kaufmann, San Francisco, California, 2003.

FAQ – Frequently Asked Questions. The Greenstone developers have compiled a list of questions they are often asked by users. And they've provided answers as well. To see the questions and answers, go to http://greenstone.org and click the tab marked *faq*.

Mailing list archives. For several years the Greenstone user community have used a mailing list to communicate problems and solutions. The archives of this list are a good place to see if someone has asked your question before (but not often enough to make it into the FAQ list). The archives (organized as a Greenstone collection, naturally) are online at:

http://www.nzdl.org/gsarchives

Mailing lists. If you haven't found the answer to your problem yet, try sending a question to the mailing lists. There are two lists, one for users and one for developers (if you're not sure, use the former one):

- *Greenstone User's List*

This list is for general Greenstone discussions. To send a message to this list, address it to greenstone-users@list.scms.waikato.ac.nz.

- *Greenstone Developer's List*

This list is for more technical discussions by people developing or modifying Greenstone. To send a message to this list, address it to greenstone-devel@list.scms.waikato.ac.nz. Note: You need to subscribe to this list before you may post to it.

Joining the community. The Greenstone community welcomes new arrivals, whether users or developers, and encourages them to contribute their expertise to help others. Different people contribute to Greenstone according to their background and interests. Some help answer other users' questions; others may share a complex format statement; and still others may write an extension to the software.

Language translation. Greenstone supports interfaces and documents in many languages. While the developer team are good at programming, they have scant knowledge of the many languages used around the world. They rely on the community to provide the translations of the interface. Helping translate Greenstone into new languages doesn't require any technical knowledge and is a great way to contribute to the community.

If you would like to help translate the interface into a language that is not supported, please email the Greenstone team at greenstone@cs.waikato.ac.nz.

Commercial applications. If you want specialized help with Greenstone, including custom collection building, commercial support is available from:

- DL Consulting, Hamilton, New Zealand

http://www.dlconsulting.co.nz/
contact@dlconsulting.co.nz

Unit closing

Although the topic of the module is multimedia, this particular unit has not addressed multimedia specifically. But multimedia collections stretch the limits of the software and force you into a deeper understanding of what is going on underneath.

In this unit we have explored under the hood of Greenstone and learned what makes it tick. The developers have striven hard to make the Librarian Interface easy and intuitive to use and abstract away some of the details beneath. However, Greenstone is a complex system. As with its counterparts in the physical world – automobiles, jumbo jets, power plants – expert knowledge at a microscopic level is sometimes needed to resolve problems. Knowledge gives you extra leverage for achieving your goals. This unit provides some knowledge and a microscope. Thus equipped,

we hope, your experience with Greenstone will be as satisfying as taking a well-tuned car out for a spin on the open road.

The most important feature of any open source software is its users. The existence of a community of users who help and support each other is one of the most powerful aspects of open source software, and one that distinguishes it from commercial software. The international community is an invaluable resource: knowledgeable, experienced, responsive, available around the clock, and free. This unit has introduced you to the Greenstone community and explained how to get involved. We hope you will join, learn, enjoy, and contribute.

Not everyone wants to be a mechanic. While technically oriented people like the Greenstone developers revel in this stuff, we are conscious that you the reader – presumably an educator or librarian – may find the level of detail in this unit excessive and confusing. If so, take heart in the fact that the worst is over. This course goes no deeper into the innards of Greenstone. The next module reaches out to other standards, different systems, and fresh applications.

MODULE 5 OPEN STANDARDS AND CASE STUDIES

Goal

To be able to relate the material in this course to the wider world of digital libraries, including contemporary metadata standards, institutional repositories, and existing educational digital libraries.

Objectives

Upon completion of Module 5, you will be able to:

- distinguish between a *location* and a *name* and relate this distinction to how documents can be identified on the web;
- find the namespaces in an XML document;
- identify uses of XLink in an XML document and explain what they mean;
- distinguish between an XML Schema and an XML DTD;
- describe the Open eBook format;
- explain the purpose of RDF, MODS, and METS;
- identify the different components in a METS specification and say what they mean;
- describe the function of an institutional repository;
- name and describe the protocol used to transfer metadata between institutional repositories;
- explain the purpose of the Z39.50 protocol and sketch how it is used to transport information;
- discuss the key differences between DSpace and Greenstone;
- take documents exported from a DSpace institutional repository and build them into a Greenstone collection;
- combine the metadata and documents from several OAI servers into a single Greenstone collection;
- explain how Greenstone can be used to serve a collection over OAI;
- describe the services provided by the Merlot multimedia repository;
- use Merlot to search several other educational repositories;
- describe the features that can be offered by Greenstone collections of newspaper images;
- use Greenstone to build a collection of scanned images;
- identify and discuss various ways in which transaction logs can help analyse user behaviour in a digital library;
- recount the tragedy of the BBC's Domesday project and compare its longevity with William the Conqueror's original project of the same name.

Introduction to Module 5

This final module broadens out to examine how digital libraries can interoperate. We study institutional repository systems and look at existing educational digital libraries and collections with educational potential – including failed ones.

The module is divided into three units. The first describes various standards for representing metadata, documents with many linked components, and entire collections. To work with these standards we first need to extend our acquaintance with XML, which underlies them all. The second unit examines the requirements of institutional repositories and describes protocols for interoperating between them. In the third, we meet some actual educational digital libraries and learn about the facilities they provide. We also discuss methods for studying user behaviour.

Throughout the module we continue our study of Greenstone and how it can be used as a practical implementation infrastructure to address the needs of different kinds of digital library system.

Module 5 Readings

Unit 5.1. Metadata standards: METS, MODS, RDF

Reading 1 The textbook *How to build a digital library*: all of Chapter 8 except Section 8.4.

Purpose To gain a deeper understanding of XML and related standards.

Reading 2 Course CD–ROM, Readings: "METS: Metadata Encoding and Transmission Standard", Richard Gartner, Oxford University Library Services, JISC, 2002.

http://www.jisc.ac.uk/uploaded_documents/tsw_02–05.pdf

Course CD–ROM, Readings: "MODS: Metadata Object Description Schema", Richard Gartner, Oxford University Services, JISC, 2003.

http://www.jisc.ac.uk/uploaded_documents/tsw_03–06.pdf

Purpose To gain an overview of the METS and MODS standards.

Unit 5.2. Institutional repositories and interoperability

Reading 1 The textbook *How to build a digital library*: selected parts of Chapter 8.

Purpose To learn about interoperability between digital libraries through communication protocols.

Reading 2 Course CD–ROM, Readings: "Institutional repositories: hidden treasures", by Miriam A. Drake. *Searcher*, May 2004, Vol. 12, No. 5.

http://www.infotoday.com/searcher/may04/drake.shtml

Course CD–ROM, Readings: "DAEDALUS: Initial experiences with EPrints and DSpace at the University of Glasgow", by William J. Nixon. *Ariadne*, October 2003, Vol. 37.

http://www.ariadne.ac.uk/issue37/nixon/

Purpose To learn about institutional repositories in general, then focus on a particular project that is building one such example.

Reading 3 Course CD–ROM, Readings: "StoneD: A bridge between Greenstone and DSpace", by Ian H. Witten, David Bainbridge, Robert Tansley, Chi–Yu Huang, and Katherine J. Don. Working Paper 2005/02, Department of Computer Science, University of Waikato, New Zealand.

http://www.dlib.org/dlib/ september05/witten/09witten.html

Purpose To learn about the open source institutional repository software DSpace, its commonalities and differences with Greenstone, and how the two can be linked together.

Unit 5.3. Case studies of educational digital libraries

Reading 1 Course CD–ROM, Readings: "Delivering the Maori–Language Newspapers on the Internet", Mark Apperley, Te Taka Keegan, Sally Jo Cunningham, and Ian H. Witten In *Rere atu, taku manu! Discovering history, language & politics in the Maori–language newspapers*, edited by Jenifer Curnow, Ngapare Hopa & Jane McRae. Auckland University Press, 2002, 211–232.

http://www.cs.waikato.ac.nz/~ihw/ papers/02–MA–etal–DeliveringMaori.pdf

Purpose This paper gives background on the *Niupepa* digital library.

Reading 2	Course CD–ROM, Readings: "User behaviour in learning objects repositories: an empirical analysis", J. Najjar, S. Ternier, and E. Duval. *Proceedings of the ED–MEDIA 2004 World Conference on Educational Multimedia, Hypermedia and Telecommunications*, edited by L. Cantoni and C. McLoughlin, pp. 4373–4379, 2004. http://www.cs.kuleuven.ac.be/~najjar/papers/edmedia2004.pdf
Purpose	Describes the influence of educational digital library design on the behavior of users.
Reading 3	Course CD–ROM, Readings: "Lost in Cyberspace: The BBC Domesday Project and the Challenge of Digital Preservation", Douglas Brown, June 2003, Cambridge Scientific Abstracts. http://www.csa.com/hottopics/cyber/oview.html
Purpose	A fascinating story that underlines the importance – and difficulty – of digital preservation.

UNIT 5.1 Metadata standards: METS, MODS, RDF

METS and MODS, which we study in this unit, are contemporary XML-based metadata standards managed by the U.S. Library of Congress and specifically designed for use in digital libraries. METS stands for *Metadata Encoding and Transmission Standard* and MODS stands for *Metadata Object Description Schema*. We also study RDF, the *Resource Description Framework*, which is designed to facilitate the interoperability of metadata, particularly in a web-based environment. All three standards are potentially relevant to digital libraries of educational resources.

To prepare to meet these standards we first extend our knowledge of XML and related technologies. That is the purpose of the first assignment.

Assignment 1

Read Chapter 8 of the textbook up to and including Section 8.3 and answer these questions.

- What do URL, URN, and URI stand for?
- What is the relationship between them?
- What is the purpose of namespaces in XML?
- Explain the difference between a global namespace and a qualified namespace. Give a small XML excerpt that uses tags both from the global namespace and a qualified namespace.
- How many namespaces can an XML document use?
- Compare and contrast the ability of XLink to relate documents with the hyperlinking capabilities of HTML. (You may find it helpful to draw diagrams illustrating different forms of linking.)
- With respect to structure and type information, compare and contrast the abilities of XML Schema with DTD.
- Develop an RDF diagram (along similar lines to the one shown in Figure 8.6 of Section 8.2) that describes this Study Guide.
- (Optional) Use a text editor (or XML editor if available) to specify the same information in XML

Now it is time to introduce MODS and METS. Recall the distinction made at the beginning of Unit 3.1 between internal metadata, which is intended to assist readers in navigating within documents, and external metadata, which is descriptive information about documents. MODS is a standard for external metadata, designed along similar lines to the other metadata schemes we have met elsewhere in the course. METS is concerned with all forms of metadata, both internal and external, and can be conceptualized as a framework within which document content, document structure, and traditional external metadata can be housed.

In many ways the MODS standard can be conceptualized as a contemporary form of MARC library metadata, expressed in XML. Indeed, the MARC Standards Office was instrumental in defining MODS. It is fairly easy to convert MARC metadata to MODS. It is not a straightforward one-to-one mapping, though: more than 30 years elapsed between the inception of MARC and MODS, so it is not surprising that the latter includes some additional concepts. There is a separate MARCXML standard, also administered by the MARC Standards Office at the Library of Congress that gives an exact one-to-one representation of MARC in XML.

In Assignment 2 of Unit 4.3 we learned about the Greenstone archive format, an XML-based scheme used to store documents and metadata between the importing and building phases. Using tags names, such as <*Section*>, <*Content*>, and <*Metadata name="Title"*>, it prescribes a precise XML-compliant syntax that defines a hierarchically nested document structure accompanied by metadata at any level within that hierarchy.

METS fulfills a role similar to the Greenstone archive format, but was designed more recently and is considerably more general. One important generalization is that it is allows parallel hierarchical structures to be defined. For example, a document might decompose into a hierarchy of chapters, sections, subsections, and paragraphs and into a second independent hierarchy of page headers, page bodies, and page footers. This cannot be done in the Greenstone archive format, which is restricted to a single hierarchical decomposition. In addition, any metadata standard that has an XML manifestation can be embedded in METS, including Dublin Core, XMLMARC, and (as we shall see) MODS.

Assignment 2

Read "METS: Metadata Encoding and Transmission Standard" and "MODS: Metadata Object Description Schema", both by Richard Gartner. Now answer the following questions.

- What are the main top level sections to METS? Write a brief paragraph for each that explain its purpose.
- What dual demands does the MODS standard try to reconcile?
- How many top-level elements and sub-elements does MODS have?
- Give an excerpt of XML that shows how sub-elements in MODS can be used within a top-level element to qualify the data present.
- Which of the top-level elements are mandatory?
- Name five top-level elements and write a brief paragraph describing each.

Greenstone can import and export METS documents. More precisely, given that METS is not a particular format but a framework for describing documents and metadata, it can import and export documents that conform to a particular METS "profile". A METS profile describes a class of METS documents in sufficient detail to provide both authors and programmers the guidance they require to create and process it.[4]

METS documents can be imported into Greenstone using a METS plug-in, while in the other direction any Greenstone collection can be exported to METS using the Export option on the Librarian Interface's *File* menu.

Greenstone can also use the METS representation internally as an alternative to the older Greenstone Archive Format. This is done by specifying METS in a special *saveas* switch to Greenstone's *import* process that we learned about in Unit 4.3 (Assignment 2). This switch instructs Greenstone to convert the documents processed by the importing phase into the METS format. The files are stored in the *archives* folder as before. In the building phase METSPlug detects and process these files – which store essentially the same information as the Greenstone Archive files – and builds exactly the same collection as before.

Enrichment exercise: Exporting as METS

1. In GLI, open the Tudor collection.

To be able to substitute METSPlug for GAPlug you need to be in Expert mode.

2. Click *File→Preferences→Mode* and change to Expert mode.

3. Switch to the **Design** panel select **Document Plugins.** Remove **GAPlug** from the list of plug–ins and add **METSPLug.**

4. Now change to the **Create** panel, locate the options for the import process, and set – *saveas* to *METS*. Import options are not available unless you are in *Expert* mode.

5. Rebuild the collection.

6. In your Windows file browser, locate the *archives* folder for the Tudor collection. For each document in the collection, Greenstone has generated two files: *docmets.xml*, the core METS description, and *doctxt.xml*, a supporting file. (Note: unless you are connected to the Internet you will be unable to view *doctxt.xml* in your web browser, because it refers to a remote resource.) Depending on the source documents there may be additional files, such as the images used within a web page. One of MET's many features is the ability to reference information in external XML files. Greenstone uses this to tie the content of the document, which is stored in the external XML file *doctxt.xml*, to its hierarchical structure, which is described in the core METS file *docmets.xml*.

[4] More information on METS profiles can be found in http://www.loc.gov/standards/mets/profile_docs/METS.profile.requirements.rtf

Unit closing

If you will be working with digital libraries over the coming years (or decades), keeping up with standards is going to have to be an ongoing activity. There's an old saying that "the great thing about standards is that there are so many different ones to choose from", to which we might wryly add, "and they're changing all the time". None of the standards described in this unit existed ten years ago (including XML itself, whose specification was first released in February 1998) and all will seem antiquated in ten years time. The textbook for this course, which was published in 2003, mentions neither METS nor MODS.

Modern standards tend to be complex, reflecting the complexity of information and how we deal with it. Also, they tend to rely heavily on other standards. Because of these two features, implementers – pressed for time and with limited resources – throw up their hands in despair and define subsets of the standard that their implementation copes with. In METS, subsets have been institutionalized as METS "profiles". The net result is that two systems may apparently use the same standard but fail to interoperate in any meaningful way. You have to read the fine print.

Fortunately, all these standards are represented in XML. To translate from one dialect or "profile" to another involves transforming an XML document into a different form. Maybe something will inevitably be lost because the two profiles do not express exactly the same information, but much of the gist can probably be retained using an XML transformation. It is likely that XSLT, the XSL transformation language that you met in Unit 3.1 (Assignment 3), will play a central role in ensuring practical interoperability between different implementations of these standards.

UNIT 5.2 Institutional repositories and interoperability

In this unit we learn about institutional repositories as a form of digital library, and about the complexities of interoperating between them. Interoperability is the ability to share digital library sources and services. Sources are typically documents, metadata, or some combination thereof. Services encapsulate a wide variety of functional behaviours, from searching to retrieving a document based on its identifier and beyond. In Unit 5.1 we studied MODS and METS, two standards that express the structure and content of documents and metadata. These facilitate the interoperability of sources. To support the interoperability of services, it is necessary to access them using standard protocols.

In Assignment 1 we will meet two standard protocols for digital library services: Z39.50 and the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH). Next, in Assignment 2, we learn about DSpace and EPrints, two open source software systems that are specifically designed to support institutional repositories. Then we examine various possibilities for interoperability between DSpace and Greenstone. Some practical exercises that illustrate interoperability conclude the unit. First we demonstrate source interoperability between DSpace and Greenstone, in both directions. Then we illustrate service interoperability by building a Greenstone collection from metadata (and corresponding documents) that has been exported from an external digital library using OAI-PMH. Nothing is known about the software used to serve the external library except that it complies with the OAI protocol.

Assignment 1

Read from Section 8.5 to the end of Chapter 8 of the textbook. (Section 8.4 is left as an optional reading exercise.) Now answer the following questions.

- Who administers the Z39.50 protocol? What is the earliest reference to it and in which year was the first version of the standard approved?
- What is the purpose of a Z39.50 registry and what elements of the standard does it cover? How many categories are there in the protocol and what is a minimum (baseline) implementation?
- Where did the motivation come for the Open Archives Initiative? The Open Archives Initiative was launched in 1999. The first version of the standard was released in January 2001, with a minor update (version 1.1) in July 2001, and a new version (2.0) in June 2002. Compare this pattern of development with Z39.50 and comment on any differences.
- Compare and contrast the approaches taken by Z39.50 and OAI-PMH.
- Think of a digital library project related to your local university library (or any other library you are familiar with) that would require service-level interoperability. Which of these two protocols would you recommend, and why?

Since the textbook was published a new version of the Z39.50 standard has been ratified (in 2003).
Now we turn from protocols to institutional repositories.

Assignment 2

Read "Institutional repositories: hidden treasures" by Miriam Drake, which introduces the general topic of institutional repositories, and answer these questions.

- What is the main purpose of an institutional repository and in what kind of institution have most repositories been developed?
- According to the article, what are the ten document categories typically found in university repositories? For your host institution, how well do these match your requirements?
- Name two issues that affect the sustainability of an institutional repository.

Now read "DAEDALUS: Initial experiences with EPrints and DSpace at the University of Glasgow" by William Nixon, which describes one UK university's experience when working with two prominent open source institutional repository systems, and answer these questions.

- In DAEDALUS, a record level description (metadata) was augmented to include the type and status of a document, in order to automatically determine the appropriate collection for a given document. What are the five collection categories? How do they correspond to the list of ten document types in Drake's article?
- The article identifies three key areas in the development and maintenance of an institutional repositories: configuration, submission, and administration. For you host institution, who would have the suitable skills to undertake the work in each area? If there is no-one currently with the necessary skills to work in a particular area, who is the most suitable person to undergo training to take on the work?

As noted at the beginning of this unit, digital library systems can interoperate by exchanging sources (documents and metadata) or by using each other's services. The next reading focuses on two prominent open source digital library systems and discusses various ways in which they might interoperate.

Assignment 3

Read "StoneD: A bridge between Greenstone and DSpace" by Ian Witten and others and answer the following questions.

- For both DSpace and Greenstone, name six key points that characterize the systems.
- The paper itemizes several differences between DSpace and Greenstone. Name two of them and write a brief paragraph describing each.
- Name four ways in which links can be formed between the two systems. For each one, specify whether it is an example of source interoperability or service interoperability.

The article you read in Assignment 3 discusses how to construct a Greenstone collection from files exported from DSpace. We now build this very collection, as a practical exercise to demonstrate source-level interoperability. In this course you have not installed DSpace (it requires considerable technical expertise to install and, at the time of writing, does not run on Windows systems). Therefore the assignment begins at the point where documents have already been exported from the DSpace collection.

This exercise uses PDF, Word, and MP3 files that represent a mixture of articles, interviews, and transcript of interviews. Unlike other exercises in this course the documents are not sourced from the real world but have been generated in the lab to make an interesting example for a hypothetical academic music department. Remember this when you examine the finished collection.

Practical exercise: From DSpace to Greenstone

1. Launch the Greenstone Library Interface (if is not already running).
2. Change to *Library System Specialist* mode (using File→Preferences), because you will need to change the order of plug–ins in the **Design** panel.
3. Start a **new collection** called **StoneD** and fill out its fields appropriately. Leave the metadata set at Dublin Core, the default.
4. Switch to the **Design** panel and select the **Document Plugins** section on the left–hand side. **Remove TEXTPlug, EMailPlug**, and **HTMLPlug.** Strictly speaking we do not need to remove these, however it reduces clutter.
5. Now add **DSpacePlug.** Leave the plugin options at their defaults and press <**OK**>.
6. Using the up and down arrows, **Move** the position of **DSpacePlug** to above **PDFPlug** and below **GAPlug.**
7. Now add **MP3Plug**, with the default configuration options. Its position in the plug–in pipeline need not be changed.

8. In the **Gather** panel, locate on the course CD–ROM the folder **sample_files\dspace\exported_docs.** It contains five example items exported from a DSpace institutional repository. Copy them into your collection by dragging them over to the right–hand side of the panel.

9. **Build** the collection and **preview** it to see the basic defaults exhibited by a DSpace collection.

If you browse by titles a–z, you will find seven documents listed, though only five items were exported from DSpace. Two of the original items had alternative forms in their directory folder. DSpace plug–in options control what happens in such situations: the default is to treat them as separate Greenstone documents.

*Below we use a plug–in option (*first_inorder_ext*) to fuse the alternative forms together. This option has the effect of treating documents with the same filename but different extensions as a single entity within a collection. One of the files is viewed as the primary document – it is indexed and metadata is extracted from it if possible – while the others are handled as "associated files".*

The first_inorder_ext *option takes as its argument a list of file extensions (separated by commas): the first one in the list that matches becomes the primary document.*

10. Select **DSpacePlug** and click <**Configure Plugin**>. Switch on its configuration option **first_inorder_ext.** Set its value to *pdf,doc,mp3* in the popup window that appears and press <**OK**>.

11. **Build** and **preview** the collection.

    There are now only 5 documents, because only one version of each document has been included – the primary version.

The DSpace exported files contain Dublin Core metadata for title and author (amongst other things). Next augment the collection with corresponding indexing and browsing capabilities.

12. In the **Design** panel, select **Search Indexes.** Delete the *ex.Title* and *ex.Source* indexes and add one for **dc.Title** called "titles" and another for **dc.Contributor** called "authors".

13. Stay within the **Design** panel, select **Browsing Classifiers**, and **delete** both **AZList** classifiers (*ex.Title* and *ex.Source*). Add an **AZList** classifier for **dc.Title** and another for **dc.Contributor**.

14. Now select the **Format Features** section of the **Design** panel and replace the **VList** format statement with this:

```
<td valign=top>
  [srclink]{or}{[thumbicon],[srcicon]}[/srclink]
</td>
<td valign=top>
[highlight]{or}{[dls.Title],[dc.Title],[ex.Title],Untitled}[/highlight]
  {If}{[ex.Source],<br>
  <i>([ex.Source])</i>}{If}{[equivlink],<br>
  Also available as:[equivlink]}
</td>
```

    You will find this text in the file *format_tweak.txt* in the *dspace* folder of *sample_files* and you can copy and paste this as you did before (in Unit 4.2). Remember to press <**Replace Format**> when finished.

15. **Build** collection once again and **preview** it.

There are still only 5 documents, but against some of the entries – for example, *Interview with Bob Dylan* – the line "Also available as", appears followed by icons that link to the alternative representations.

In the exercise that follows you export a Greenstone collection in a form suitable for Dspace. It is possible to do this from the Librarian Interface's File menu, which contains an item called *Export...* that allows you to export collections in different forms. However, to gain a deeper understanding of Greenstone, we perform the work by invoking a program from the Windows command-line prompt. This requires some technical skill and is therefore set as an enrichment exercise. If you are not used to working in the command-line environment we recommend that you skim this enrichment exercise.

Enrichment exercise: Greenstone to DSpace

16. Open a DOS window to access the command–line prompt. This facility should be located somewhere within your Start→Programs menu, but details vary between different Windows systems. If you cannot locate it, select Start→Run and enter *cmd* in the popup window that appears.

17. In the DOS window, move to the home directory where you installed Greenstone. This is accomplished by something like:

    cd C:\Program Files\greenstone

18. Type:

    setup.bat

to set up the ability to run Greenstone command–line programs.

19. Change directory into the collection you built in the last exercise:

    cd collect\stoned

Note: even though the collection name used capital letters the directory generated by the Librarian Interface is all lowercase.

20. Run the following command to export the collection using the DSpace import/export format:

    perl –S export.pl –saveas DSpace –removeold stoned

Exporting in Greenstone is an additive process. If you ran the export.pl command once again, the new files exported would be added – with different folder names – to those already in the export folder. For the kind of explorations we are conducting we might re–run the command several times. The removeold option deletes files that have previously been exported.

21. This command has created a new subfolder, *collect\stoned\export*. Use the file browser to explore it. In it are the files needed to digest this set of documents into DSpace. Not surprisingly, they are very similar to those with which we began the previous Practical Exercise.

You could equally well run the export.pl command on a different Greenstone collection, such as the Course Readings collection, and transfer the output to a DSpace installation by using DSpace's batch–import facility.

The next exercise explores service-level interoperability using the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH). The practical exercises in this course use a stand-alone computer, so we will not actually connect to the external server that is acting as the data provider. Instead we have provided an appropriate set of files on the course CD-ROM. These take the form of XML records that would be produced by the OAI-PMH protocol. The enrichment exercise that follows describes how the connection can be accomplished.

We have already met a pre-built OAI based Greenstone collection in Assignment 1 of Unit 4.3, when examining the documented example collections supplied with your Greenstone installation. The exercise that follows is based on this

collection: it takes you through the steps necessary to reconstruct it. (Note: this example is a collection of images: you will not be able to build it unless ImageMagick is installed on your computer as described in Unit 2.2 Assignment 1.) You may wish to take a look at the documented example collection *OAI demo* now to see what the following practical exercise will build.

Practical exercise: OAI collection

1. Start a new collection called **OAI Service Provider.** Fill out the fields with appropriate information. You can leave the default metadata set as Dublin Core, although we do not make use of it.

2. In the **Gather** panel, navigate to the *sample_small* folder in *sample_files/oai*. Drag this folder into the collection and drop it there.

3. During the copy operation, a popup window appears asking whether to add **OAIPlug** to the list of plug–ins used in the collection, because the Librarian Interface has not found an existing plug–in that can handle this file type. Press the <**Add Plugin**> button to include it.

When files are copied across like this, the Librarian Interface studies each one and uses its filename extension to check whether the collection contains a corresponding plug–in. Up until now, the answer has always been yes, so all file transfers have proceeded without interruption. This time, however, no plug–in in the list is capable of processing the OAI file records that are copied across (they have the file extension .oai).

Sometimes there is more than one plug–in that could process a file – for example, the .xml extension is used for many different XML formats. The popup window, therefore, offers a choice of all possible plug–ins that matched. It is normally easy to determine the correct choice. If you wish, you can ignore the prompt (click <Don't Add Plugin>), because plug–ins can be added later, in the Document Plugins section of the Design panel.

4. You need to configure the Image plug–in. In the **Design** panel, select the **Document Plugins** section, then select the **plugin ImagePlug** line and click <**Configure Plugin**>. In the resulting popup window locate the **screenviewsize** op–tion, switch it on, and type the number 300 in the box beside it to create a screen–view image of 300 pixels. Click <**OK**>.

5. Now switch to the **Create** panel and **build** and **preview** the collection.

Like other collections we have built by relying on Greenstone defaults, the end result is passable but can be improved. The next steps refine the collection using the metadata harvested by OAI–PMH into the .oai files.

6. In the **Browsing Classifiers** section of the **Design** panel, delete the two **AZList** classifiers (*ex.Title* and *ex.Source*).

7. Add an **AZCompactList** classifier based on **ex.Subject** metadata.

8. Now add an **AZCompactList** classifier based on **ex.Description** metadata. In its configuration panel select **mincompact = 1, maxcompact = 10,** and **buttonname = Captions.**

9. In the **Search Indexes** section of the **Design** panel, delete all indexes and add a new one called "captions" based on *ex.Description* metadata.

10. **Build** collection and **preview** it.

Further improvements can be made by tweaking the presentation.

11. In the **Design** panel, select **Format Features.** First replace the **VList** format statement with this:

```
<td>
{If}{[numleafdocs],[link][icon][/link],[link][thumbicon]
    [/link]}
```

```
</td>
<td valign=middle>
  {If}{[numleafdocs],[Title],<i>[Description]</i>}
</td>
```

You will find this text in the file *vlist_tweak.txt* in the *oai* folder of *sample_files*. Remember to press <**Replace Format**> when finished

This format statement customizes the appearance of vertical lists, such as the search results and captions lists to show a thumbnail icon followed by Description *metadata*. *Greenstone's default is to use extracted metadata, so* [Description] *is the same as* [ex.Description].

Next, select **DocumentHeading** from the **Choose Feature** pull–down list and make its format statement (which is currently blank) read

```
<h3>[Subject]</h3>
```

The document heading appears above the detach and no highlighting *buttons when you get to a document in the collection*. *By default* DocumentHeading *displays the document's* ex.Title *metadata*. *In this particular set of OAI ex–ported records, titles are filenames of JPEG images, and the filenames are particularly uninformative (for example,* 01dla14). *You can see them in the* ***Enrich*** *panel if you select an image in* sample_small→oai→JCDLPICS→ srcdocs *and check its filename and* ex.Title *metadata*. *The above format statement displays* ex.Subject *metadata instead*.

Finally, you will have noticed that where the document itself should appear, you see only *This document has no text*. To rectify this, select **DocumentText** in the **Choose Feature** pull–down list and use the following as its format statement (which is currently blank) (this text is in *doctxt_tweak.txt* in the *format_tweaks* folder mentioned above):

```
<center><table width=_pagewidth_ border=1>
<tr><td colspan=2 align=center>
<a href=[OrigURL]>[screenicon]</a></td></tr>
<tr><td>Caption:</td><td> <i>[Description]</i> <br>
(<a href=[OrigURL]>original [ImageWidth]x[ImageHeight] [ImageType] available</a>)
</td></tr>
<tr><td>Subject:</td><td> [Subject]</td></tr>
<tr><td>Publisher:</td><td> [Publisher]</td></tr>
<tr><td>Rights:<td> [Rights]</td></tr>
</table></center>
```

This format statement alters how the document view is presented. *It includes a screen–sized version of the image that hyperlinks back to the original larger version available on the web*. *Factual information extracted from the image, such as width, height, and type, is also displayed*.

12. Recall that format statements are processed by the runtime system (Unit 4.3), so the collection does not need to be rebuilt for these changes to take effect. Switch to the **Design** panel and press <**Preview Collection**> to see the changes.

To expedite building, this collection contains fewer source documents than the pre–built version supplied with the Greenstone installation. *However, after these modifications, its functionality is the same*.

In the above exercise we did not obtain the data from an external OAI-PMH server. This missing step is accomplished by running a command-line program (like the exporting process described in the previous enrichment exercise). To do this, your computer must have a direct connection to the Internet – being behind a firewall may interfere with the ability to download the information. If your working environment does not meet this requirement, skim the enrichment exercise and move on to the assignment that follows it.

Enrichment exercise: Downloading over OAI

13. **Save** your collection. Note its directory name, which should be oaiservi (it appears in title bar of the Librarian Interface), and **quit** GLI.

14. Perform the first four steps of the previous enrichment exercise: open a command window, change directory to where Greenstone is installed, run *setup.bat*, and change directory once again, this time into *collect\oaiservi*, the folder containing the OAI Service Provider collection you built in the last exercise.

15. In a text editor, open the collection's configuration file, which is in *oaiservi\etc\collect.cfg*. Add the following line (all on one line):

    acquire OAI –src rocky.dlib.vt.edu/~jcdlpix/
    cgi–bin/OAI1.1/jcdlpix.pl –getdoc

    Although the position of this line is not critical, we recommend you to place it near the beginning of the file, after the public and creator lines but before the index line. Save the file and quit the editor.

16. Delete the contents of the collection's *import* folder. This contains the canned version of the collection files, put there during the last Practical Exercise. Now we want to witness the data arriving anew from the external OAI server.

17. Back at the DOS prompt, run *perl –S importfrom.pl oaiservi*.

Greenstone will immediately set to work and generate a stream of diagnostic output. The importfrom.pl *program connects to the OAI data provider specified in collection configuration file (it does this for each "acquire" line in the file) and exports all the records on that site*.

18. The downloaded files are saved in the collection's import folder. Once the command is finished, everything is in place and the collection is ready to be built. Confirm you have successfully acquired the OAI records by rebuilding the collection.

There is no provision in OAI-PMH for describing document content – in fact, the standard makes it quite clear that this is outside its remit. However, there is an informal convention that the *Identifier* metadata field may contain a URL for the document it describes. Greenstone has a *getdoc* option, used in the previous Enrichment exercise that exploits this. It checks the *Identifier* field of each downloaded record to see if it contains a URL and if so downloads it into a file and associates the file with that record.

Behind the scenes, Greenstone's OAI importing program (*importfrom.pl*) communicates with the external server over the HTTP protocol, which is the backbone of the web. If metadata records (and associated documents) have been downloaded before, they are only downloaded again if the copy on the server is newer than the existing one.

The following assignment draws on your newly-acquired knowledge about OAI and Greenstone. It assumes that even if you haven't actually performed the above enrichment exercise, you have read through it and digested the instructions.

Assignment 4

How would you use Greenstone to implement a virtual digital library that harvests a set of OAI repositories maintained by different institutions and provides a single point of access to them?

So far, we have learned how to make Greenstone collections from material that has been obtained from an OAI repository elsewhere. But what if you would like others to access your Greenstone collections using the OAI protocol? To permit this, Greenstone incorporates the ability to serve any collection or collections over OAI. To do this, you must ensure that an "OAI server" program, which is called *oaiserver*, is active. Then anyone can access your Greenstone

collections over OAI by using the same URL as they use to access your Greenstone installation but with *library* replaced by *oaiserver*. That is, if your Greenstone collections are accessed with the URL
http://127.0.0.1/cgi-bin/library

then your OAI server is accessed using the URL
http://127.0.0.1/cgi-bin/oaiserver (in your case, the "127.0.0.1" will probably be replaced by something different).

You can specify which collections will be accessible over OAI. The OAI setup is determined by the file *oai.cfg* file in the Greenstone *etc* directory. This file specifies general information about the repository and lists the collections that are to be made accessible. By default, collections are not accessible: to enable one, add its name to the *oaicollection* list. (Collections built with versions of Greenstone earlier than 2.52 must be rebuilt before they can be served over OAI.) The OAI server only supports Dublin Core metadata. For collections that use other metadata sets, rules should be provided to map the existing metadata to Dublin Core as described in the *oai.cfg* file.

As mentioned earlier your collections will only be accessible if the *oaiserver* program is active. When Greenstone was installed in Unit 2.2 (Assignment 4), the Windows local library server, which is the default, was used. To run the OAI server you must instead run the Web Library version of Greenstone and for this you need to install a web server (such as Apache). Instructions for this are included in the Greenstone Installer's Guide, which is on your course CD-ROM: if you have difficulty doing this you should seek assistance from a computer specialist. You need to install a web server and then re-install Greenstone just as in Unit 2.2 Assignment 4, but this time choosing the *web library* instead of the *local library*. When you use the web library version of Greenstone the *oaiserver* program is automatically activated. (It runs as a "CGI" program, just like the Greenstone *library* program does, because it resides in the Greenstone *cgi-bin* directory. Your local computer specialist will explain what this means.)

Unit closing

This unit – and indeed the entire course – has focused on technical issues rather than human, organizational, and political ones – although the latter are, of course, certain to be key factors in the creation and adoption of digital library technology. The application contexts of educational digital libraries are so varied that it is difficult to say much about non-technical issues. However, some useful points can be made in the specific context of institutional repositories.

Institutional repositories provide an infrastructure for capturing the intellectual output of an institution, storing and preserving it, and making it accessible over the long term to the broadest possible readership. In setting them up special attention must be paid to major stakeholders. For example, typical constituencies in a university setting include:

- faculty members, whose scholarship and research is being captured and disseminated;
- library staff, who will generally have responsibility for deploying, running, and maintaining the repository and associated services;
- long-term planners and decision-makers, because the venture will require a sustained financial commitment;
- members of the wider university community, whose intellectual and political influence will and develop momentum to support the project;
- early adopters.

In order to succeed, the project must be marketed and supporters recruited from throughout the community. Opinion leaders, key administrators, and respected faculty members must be persuaded; their endorsement will promote credibility. A number of critical questions must be asked. For example, can old versions of rewritten articles be removed? Indeed, can anything be withdrawn from the repository, and under what circumstances? Can faculty deposit huge items – e.g. terabyte-scale datasets – into the repository?

Early adopters play crucial roles by testing all aspects of the repository and providing feedback. Benefits for them include the visibility that comes with blazing new trails in the institution. They influence the design of the procedures and user interfaces and receive prompt attention in solving problems. They can showcase their research ahead of later entrants and reach worldwide audiences immediately; their names figure in promotional and publicity materials.

Setting up an institutional repository is an open-ended, long-term project. What will it cost to run, support, maintain, and upgrade over decades? One of the biggest unknowns is assessing the costs of data preservation, because it

is hard to assess the feasibility and cost of large-scale data migration. Managing the storage capacity for massive datasets requires careful financial planning. Forecasting staffing is essential, because it will become a major expense. Finally, institutional repositories are often introduced into a library environment that already suffers from severe funding restrictions.

Institutions seeking to implement repositories quickly learn that building awareness and user demand among relevant constituencies before determining technological specifications raises issues that help define services and set policy. They will have to elaborate a governance structure for the repository and define the content it will handle – whether scholarship, research data, theses, metadata, or administrative data. An operations plan spelling out system backup and recovery policies and procedures needs to be devised. A management, staffing, and advisory structure that suit the institution's culture must be put into place.

UNIT 5.3 Case studies of educational digital libraries

This unit presents case studies of existing systems that illustrate different ways of disseminating information in educational digital libraries. As you examine them, think of the potential legacy that each system leaves to future generations who wish to continue its operation in a radically new computing environment. That is the strongest argument of all for open standards and open source digital library software.

Case study: Merlot: Multimedia educational resource for learning and online teaching

Our first case study is a catalogue of materials on the web. It doesn't contain the learning objects themselves. Like the catalogue of a bricks–and–mortar library, it provides information about material that is shelved elsewhere. It's a gateway that consists of pure metadata, whereas digital libraries generally hold the actual content, too – as in the collection built from LOM records that you examined in Unit 3.3.

The Multimedia Educational Resource for Learning and On–Line Teaching (MERLOT) is a high quality collection of interactive learning materials, assignments, reviews, and people. MERLOT is also a national network of online discipline communities that will be selecting and peer–reviewing learning materials in their specific disciplines. MERLOT serves as a national gateway to web–based peer–reviewed learning materials.

From: http://www.merlot.org/help/FAQ.po

This is a rich multidisciplinary project that engages its community with peer review processes. Merlot integrates reviews and comments into its searching and browsing facilities. These user feedback channels help build a community of educators to sustain and enrich the catalogue of educational materials.

The snapshots overleaf illustrate the system. Starting at the home page, the user enters the search term "music" (off screen) and presses "go" to initiate a search. The query term is intentionally broad: the user is trying to get a feel for what the repository contains. Such usage is frequently observed when analyzing user logs – even when, as in Merlot, there is a "browse by category" feature that could be used instead. The search returned 155 matches and the user continues to explore by homing in on the top item, *Interactive Music Skill Checks* by David Megill. The initial view of this learning object summarizes metadata for the resource and includes links to peer reviews. This particular item has been awarded "Merlot classic" by the subject area's editorial board.

Clicking the *location* field brings up the resource itself. It comprises a set of tests divided into *theory* and *aural* and further categorized by pitch, rhythm, and so forth. The user chooses *name the note (keyboard)*, which runs a Java applet that displays a note using music notation in another window and asks the user to play it on a piano keyboard. Having done so, the user can select *check* to see if they were right.

Merlot home page

Search results for “music”

The top matching item

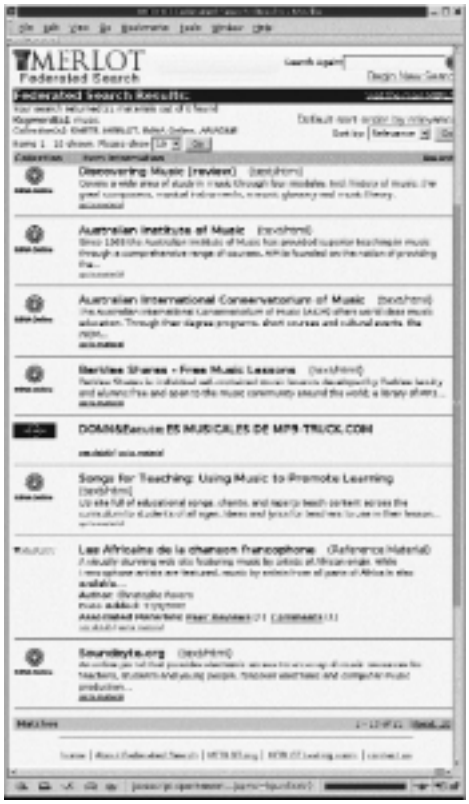
Metadata for the item

Choosing a skill test

Merlot supports federated search with the *search more libraries* button on the home page. In the example below the user has repeated the query term *music*. The default settings for federated search limit the number of items returned from each site to 10, and only 21 matching documents are returned: 10 from Merlot, 10 from EdNA, and 1 from Ariadne (information about these libraries is given below). The sites queried by the federated search are by no means uniform and the information displayed for matching items varies. For EdNA the only option is to go directly to material, Ariadne provides a brief description, and Merlot gives further details, such as author and date, along with links to reviews.

Federated search in Merlot

Search results for “music”



Like Careo (described in Unit 3.3), much of Merlot’s functionality can be accessed by anyone, but only registered users can perform operations, such as storing objects in a personal area and writing reviews. The following assignment asks you to look at the system, but you will only be able to do this if you have Internet access.

Assignment 1 (online only)

If you have online access, browse and search the Merlot catalogue at http://www.merlot.org

- Can you find material that is relevant to any of your courses?
- Are the browse categories useful?
- Are the reviews and user comments valuable?

Consider the problem of maintaining a metadata-only catalogue.

- How does the catalogue get updated when the resources are changed?
- How might this work when a resource appears in many catalogues?

Merlot’s federated search service is at http://fedsearch.merlot.org/main/search.jsp

Using it, you can search four learning repositories simultaneously:

- MERLOT;
- EdNA Online, which aims to support and promote the benefits of the Internet for learning, education and training in Australia;

- SMETE, a gateway to resources that advance, strengthen and improve the teaching, learning and comprehension of Science, Mathematics, Engineering and Technology in the U.S.;
- ARIADNE, which is designed to increase European citizens' awareness of existing ICT-based training channels.

Repeat the first question of this assignment using the federated catalogue. What differences do you notice?

You can read about the activities of the Merlot community in the online proceedings of their conferences at http://conference.merlot.org/

Case study: Niupepa: Newspapers in Māori

Māori are the indigeneous people of New Zealand and Niupepa is a collection of historic newspapers published primarily for a Māori audience between 1842 and 1932, which is of particular interest because it covers the period of European colonization (New Zealand, being remote, was discovered rather late by Europeans). The newspapers can be searched (full text), browsed (by series), and accessed by date. The collection has been made available by the New Zealand Digital Library Project at the University of Waikato.

The Niupepa collection contains over 17,000 newspaper pages taken from 34 separate periodicals. Some were government sponsored, others were initiated by Māori, and the remainder came from religious groups. It is based on a microfiche collection produced by the Alexander Turnbull Library in New Zealand. Most of the collection is written solely in Ma–ori (70%), some is bilingual (27%), and a small proportion is in English only (3%).

The collection has four components:

- facsimile images of the original pages;
- text extracted from the newspapers (for searching);
- bibliographic commentaries for each newspaper title;
- English abstracts for each issue.

Unlike the other digital libraries and repositories referred to in this course (e.g. DLESE, NLVM, and MERLOT), Niupepa has no explicit educational metadata. Its educational value stems from the greatly increased availability of the resource. The project has repositioned these Māori newspapers from extremely restricted access using mi–crofiche readers in particular libraries to global availability from any Internet terminal. In addition, and equally im–portantly, new access mechanisms like full text searching have been added. These changes provide a baseline that educators, historians, and researchers can exploit to design further educational activities involving this mate–rial. It can be viewed as a first step towards a fully–featured educational digital library that includes both raw ma–terial and specific activities based on it.

Various snapshots are shown of the collection in use. The user begins by viewing the home page to learn what is in the collection, first in Māori and then in English. Next they search the full text of the newspapers for occurrences of the word *waka* (Māori for *canoe*). Scanning down the list of matching documents, they select an item of interest – one that happens to mention *waka* in the title. Clicking on this brings up an initial view of the document. From here various views are possible, including the facsimile image of the newspaper's first page. In the next snapshot the user is browsing by series title, first looking at the newspaper series, which also shows how many issues each series has, then expanding the bookshelf for *Anglo Māori Warder* to see the individual items.

Niupepa home page

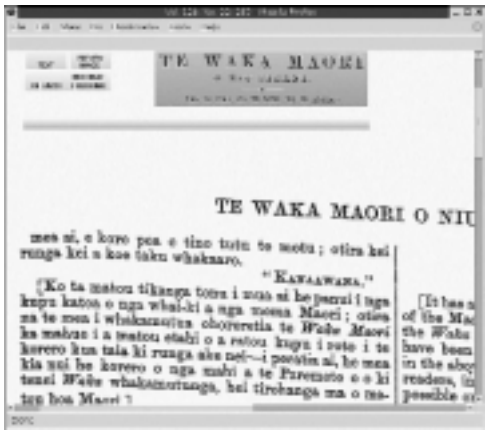
Home page in English

Searching the collection for waka

Search resul

Viewing a document: plain text

Viewing a document: scanned image



Browsing by series: top–level newspaper titles

Browsing by series: Anglo Ma–ori Warder 1848

Assignment 2

Read "Delivering the Maori-Language Newspapers on the Internet" by Mark Apperley and others for background on the Niupepa project and then answer these questions.

- What document collections that you are familiar with could benefit from this approach?
- List some educational applications of this material that digital library technology would enable.
- Can you think of a way to handle errors from the OCR process? Describe it.

You can explore Niupepa online at http://www.nzdl.org/niupepa

The exercise that follows asks you to use Greenstone to build a small replica of Niupepa, using five newspapers taken from two newspaper series. It allows full text searching and browsing by title and date. When a newspaper is viewed, a preview image and its corresponding plain text are presented side by side, with a *goto page* navigation feature at the top of the page.

The collection involves a mixture of plug-ins, classifiers, and format statements. The bulk of the work is done by *PagedImgPlug*, a plug-in designed precisely for the kind of data we have in this example. For each document, an "item" file is prepared that specifies a list of image files that constitute the document, tagged with their page number, and (optionally) accompanied by a text file containing the machine-readable version of the image, which is used for full text searching. Three newspapers in our collection (all from the series *Te Whetu o Te Tau*) have text representations and two (from *Te Waka o Te Iwi*) have images only. Item files can also specify metadata. In our example the newspaper series is recorded as *ex.Title* and its date of publication as *ex.Date*. This metadata is extracted as part of the building process.

Practical exercise: Scanned image collection

1. Start a new collection called **Paged Images** and fill out the fields with appropriate information: it is a collection sourced from an excerpt of Niupepa documents; the only metadata used is document title and date, and these are extracted from the "item" files included in the source documents so no metadata set need be stipulated.

2. Add **PagedImgPlug** and set its **screenview** configuration option to **on.** The source images we use were scanned at high resolution and are large files for a browser to download. The *screenview* option generates smaller screen–resolution images of each page when the collection is built.

3. In the **Gather** panel, open the *niupepa\sample_items* folder in *sample_files* and drag it into your collection on the right–hand side.

4. Some of the files you have just dragged in are text files that contain the text extracted from page images. We want these to be processed by **PagedImgPlug**, not **TEXTPlug.** Switch to the **Design** panel and delete **TEXTPlug.** While you are at it, you could tidy things up by deleting **HTMLPlug, EMAILPlug, PDFPlug, RTFPlug, WordPlug**, and **PSPlug** as well, since they will not be used.

5. Now go to the **Create** panel, **build** the collection and **preview** the result. Search for *waka* and view one of the titles listed (all three appear as *Te Whetu o Te Tau*). Browse by *titles a–z* and view one of the *Te Waka o Te Iwi* titles.

This collection was built with Greenstone's default settings. You can locate items of interest, but the information is less clearly and attractively presented than in the Niupepa collection discussed above. For instance, under *titles a–z* documents from the same series are repeated without any distinguishing features, such as date. It would be better to group them by series title and display dates within each group. This can be accomplished using an *AZCompactList* classifier rather than *AZList* and tuning the *VList* format statement.

A second quirk is that when you reach a newspaper, only its associated text is displayed. When either of the *Te Waka o Te Iwi* newspapers is accessed, the document view presents the message *This document has no text.* No scanned image information (screen-view resolution or otherwise) is shown, even though it has been computed and stored with the document. This can be fixed by a format statement that modifies the default behaviour for *DocumentText*.

The next exercise takes you through a series of refinements to the collection design.

Practical exercise: Improved image collection

1. In the **Design** panel, under the **Browsing Classifiers** section, delete the **AZList** classifiers for *ex.Source* and *ex.Title*.

2. Now add **AZCompactList** for *ex.Title* and **DateList** for *ex.Date*.

3. **Modify** the format statement for **VList.** Find the part of the default statement that says

   {If}{[ex.Source],
([ex.Source])}

and change it to

   {If}{[ex.Date],: [ex.Date]}

This has the effect of displaying the extracted date information, if present.

4. At the end of this format statement, where is says:

   </td>

append

   {If}{[numleafdocs],<td>([numleafdocs] items)</td>}

As a consequence of using the AZCompactList classifier, bookshelf icons appear when titles are browsed. This revised format statement has the effect of specifying in brackets how many items are contained within a bookshelf. It works by exploiting the fact that only bookshelf icons define [numleafdocs] metadata.

5. Staying within the **Format Features** section of the **Design** panel, under "Choose Feature" select **DocumentText.** Its HTML format string is empty, triggering the default behaviour of displaying the document's plain text, or, if there is none, "This document has no text". Change this to:

```
<center>
  <table width=_pagewidth_>
    <tr>
      <td valign=top>[srclink][screenicon][/srclink]</td>
      <td>[Text]</td>
    </tr>
  </table>
</center>
```

(available as *niupepa\doc_tweak.txt*)

Including [screenicon] has the effect of embedding the screen–sized image generated by switching the screensize option on in PagedImgPlug. It is hyperlinked to the original image by the construct [srclink]…[/srclink].

6. Switch to the **Create** panel, **build** and **preview** the revised collection.

7. If you like, add a logo and change the background as you did in the Enrichment Exercise near the end of Unit 4.2. You will find a suitable image in the file *niupepa\images* that is activated through *macros\extra.dm*.

In the collection you have just built, newspapers are grouped by series title, and dates are supplied alongside each one to distinguish it from others in the same series. Users can browse chronologically by date and when a newspaper page is viewed a preview image is shown on the left that displays the original high-resolution version when clicked, accompanied on the right by the plain-text version of that newspaper (if available).

Like the other practical exercises in this course, this is designed to illustrate features of the Greenstone software step by step. As you become more experienced you will find yourself completing more in one go. Remember that by basing a new collection on an existing one, you inherit its settings. Having completed this exercise, if you want to establish a new collection of scanned page images you can get a head start by basing it on this one.

Case study: What do digital library users actually do?

In a physical library you can guess the interests of users by the books they borrow. In an electronic catalogue or digital library the users' queries and their access to documents can be recorded or *logged*. These records, called *transaction logs*, can help you determine how the electronic resource is being used. They often highlight striking differences between how the designers thought the resources would be used and what real users actually do with them.

Transaction logs for a digital library help you monitor and evaluate its impact on communities of learners and teachers. When designing the system it is important to consider whether the library staff will have access to usage information. It is necessary to understand local laws on maintaining data that can be traced to individual users. Usually, valuable information can be recovered without identifying users.

The box below gives a tiny excerpt of a Greenstone transaction log that records a few moments in April 2003. It was taken from the New Zealand Digital Library website *http://www.nzdl.org* and has been anonymized to remove any identifying information. Every time Greenstone generates a web page, which it does whenever a user clicks a hyperlink or button in the reader's interface, details are recorded in the log file. The excerpt tracks a single user over the course of a few web page requests.

The log records the domain name of the computer requesting the page, or – if that is not present – its IP (Internet Protocol) address, a numeric code that identifies the originating computer. This is the part that has been anonymized in the excerpt as *79ec8da3e7dc19b0*. There follows an alphabetical list of all the arguments that can be used in Greenstone URLs. There are many – over 100 in all – but for any given page request most are unused and therefore empty. The first record in the transaction log (which stretches over a dozen lines or so) shows them all, but after that only pertinent arguments are shown. These entries do not represent consecutive lines in the log file: other users (elsewhere in the world) were interacting with Greenstone at the same time.

The format of Greenstone's transaction log is intended for processing by software rather than perusal by people. Log analysis software might generate summary statistics, such as the proportion of transactions that were queries, document accesses, help requests, and so on. Although the log looks cryptic, with a bit of patience you can glean information from it.

This sequence of transactions shows a user accessing the Humanity Development Library (*c=hdl*: c gives the collection name and *hdl* is the short name of this collection), starting at its *about this collection* page (*p=about*). The interface language is English (*l=en*) and the user is running their web browser on a Windows NT 5.1 system. In the next record of the log (which is about the 15th line in the box, because the first record is so long) they move on to browse by classifier (*cl=CL1*, then in the next record *CL1.4*) before retrieving a document (*d=HASH0157382f65cca62098138be4*) located within that part of the clas–sifier. You can see how much time passes between each access. Skipping to the last entry shown in the log (*Wed Apr 02 14:10:47 +1200 2003*) you can find a query (*a=q* indicates that Greenstone's *action* is *query*) for the text *dig well*.

/cgi–bin/library 79ec8da3e7dc19b0 [Wed Apr 02 13:57:08 +1200 2003] (a=p, b=0, b1=0, b2=0, bc1aboutdesc=, bc1cfgchanged=0, bc1clone=0, bc1clonechanged=0, bc1clonecol=, bc1contactemail=, bc1dirname=, bc1dodelete=0, bc1econf=0, bc1esrce=0, bc1fromsrce=0, bc1fullname=, bc1infochanged=0, bc1input=, bc1inputnum=3, bc1inputtype=, bc1tmp=, bcp=, beu=, bft=, bl=english, bnu=, bp=, bt=0, c=hdl, cc=, ccp=0, ccs=0, cfgfile=, cl=, cm=, cq2=, ct=0, d=, de=, debc=0, ds=, dsbc=0, e=, el=prompt, er=, f=0, fc=1, fqa=0, fqc=, fqf=, fqk=, fqn=4, fqs=, fqv=, g=Document, gc=0, gt=0, h=, h2=, hd=0, hl=1, hp=, hs=0, il=l, j=, j2=, k=1, ky=, l=en, m=50, n=, n2=, nl=, o=20, p=about, pc=, pfd=0, pfe=0, pfl=0, pld=10, ple=10, pll=10, ppnum=0, pptext=, pw=, pxml=0, q=, q2=, qb=0, qt=0, r=1, rd=0, s=0, st=1, t=1, tlng=french, u=0, ua=, uan=, ug=, uma=listusers, umc=, umnpw1=, umnpw2=, umpw=, umug=, umun=, umus=, un=, us=invalid, v=0, w=utf–8, x=0, z=084bf801ce71d63f) "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.0.3705)"

/cgi–bin/library 79ec8da3e7dc19b0 [Wed Apr 02 13:57:20 +1200 2003] (a=d, ... c=hdl, ... cl=CL1, ... z=084bf801ce71d63f) "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.0.3705)"

/cgi–bin/library 79ec8da3e7dc19b0 [Wed Apr 02 13:57:48 +1200 2003] (a=d, ... c=hdl, ... cl=CL1.4, ... z=084bf801ce71d63f) "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.0.3705)"

/cgi–bin/library 79ec8da3e7dc19b0 [Wed Apr 02 13:58:10 +1200 2003] (a=d, ... c=hdl, ... cl=CL1.4, ... d=HASH0157382f65cca62098138be4, ... z=084bf801ce71d63f) "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.0.3705)"

...

/cgi–bin/library 79ec8da3e7dc19b0 [Wed Apr 02 14:10:47 +1200 2003] (a=q, ... c=hdl, ... q=dig well, ... z=084bf801ce71d63f) "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.0.3705)"

Logging in Greenstone can be switched on and off and is normally off when the software is installed. However, it is on by default in the version of Greenstone supplied on the course CD–ROM, so when you reach this point you can look at the log of what you have been doing in the course so far. The log file is in

greenstone\etc\usage.txt

Transaction logging is controlled by an entry in Greenstone's configuration file main.cfg, which resides in the etc directory. There are instructions in this file: set logcgiargs to true to keep a log of usage information.

Assignment 3

Read "User behaviour in learning objects repositories: an empirical analysis" by J. Najjar and others to learn how transaction logs of the ARIADNE educational digital library mentioned in Assignment 1 can be analyzed.

- Do you think this sort of study would be useful in evaluating an educational digital library?
- What interesting questions might it answer?
- What sort of questions about users would a transaction log study be unable to answer?
- Can you think of ways in which a digital library could automatically include records of its usage in the interface it presents to users?
- Could this help users find resources?

Case study: The Domesday Project

IWe end this unit with a sombre warning. Building digital libraries may be exciting and rewarding, but the mere act of creating an electronic collection does not guarantee longeviry. The Domesday Project (pronounced Dooms–day), organised by the BBC (British Broadcasting Corporation) with the support of the British Government and the European Community, collected a fabulous resource of data but had no long–term view of how it might be preserved for future access.

The Domesday Project was a landmark multimedia resource which was produced to celebrate the 900th anniversary of the original Domesday book. School children and researchers from across the country collected together a massive amount of material which was recorded on two special Video Discs.

From http://www.si.umich.edu/CAMILEON/domesday/what.html

Assignment 4

Read Brown (2003) for background on the Domesday Project.

- Can you still access the data from all your previous projects?
- What steps could you take to ensure that your data can be accessed in the future?

A newspaper article from 2002 captures the irony that the 11th Century content is now more accessible than that from 1986:

> "The special computers developed to play the 12-inch video discs of text, photographs, maps, and archive footage of British life are – quite simply – obsolete.
>
> As a result, no one can access the reams of project information – equivalent to several sets of encyclopaedias – that were assembled about the state of the nation in 1986. By contrast, the original Domesday Book – an inventory of eleventh-century England compiled in 1086 by Norman monks – is in fine condition in the Public Record Office, Kew, and can be accessed by anyone who can read and has the right credentials".

From "Digital Domesday Book lasts 15 years not 1000" by Robin McKie and Vanessa Thorpe, *The Observer*, Sunday March 3, 2002.

http://books.guardian.co.uk/news/articles/0,6109,661585,00.html

The 1986 Domesday Project was supposed to produce a resource that would be widely used for educational purposes and remain available for future generations. Yet, despite the achievements of the original participants, the entire resource is in danger of disappearing forever.

There are two sides to the problem of preserving Domesday: technical and legal. From a technical point of view, one preservation strategy is to write software that allows modern computers to emulate the hardware on which Domesday ran, making it possible to run the original software and thereby retrieve, display and reuse digital documents through the original interface. A project called CAMiLEON is testing the feasibility and effectiveness of emulation for preserving the intellectual content, structure, and look-and-feel of this material.

The legal problems are thornier. Many different copyright owners contributed data for inclusion on the Domesday Project disks, such as:

- professional photographs;
- photos from a national competition;
- text from published sources such as newspapers and magazines;
- movie sequences of news and sports events;
- maps licensed from the UK Ordnance Survey.

No records were kept of who contributed and under what conditions their data can be used. Intellectual property issues concerning differing aspects of the work are a key stumbling-block to restoring Domesday as a viable resource. Most (if not all) the components, including both software and content, are still within their term of copyright, and part of the restoration work will be to identify relevant rights holders and obtain permission to copy, alter, or emulate those components. The problem is far from trivial: it has been estimated that over a million people took part in the project in one way or another.

There are no easy solutions to the general problem of preservation. However, from a technical standpoint open-source software based on open standards seems to offer a far better chance of future-proofing your information than closed commercial software. Many people have old document files they can no longer read because the software they wrote them with is obsolete and unobtainable. Deciphering such files is a job for a cryptologist. At least open source software lays bare the underlying program, the cryptographic key to all those otherwise meaningless binary bits.

Unit closing

We have come a long way together in our exploration of digital libraries in education. Our closing message to you is that this course is not about studying the past, it's about constructing the future – a better future for learners and teachers.

Digital libraries are organized collections of information. Our experience of the World Wide Web – vibrant yet haphazard, uncontrolled and uncontrollable – daily reinforces the impotence of information without organization. Likewise, experience of using online public access library catalogues from the desktop – impeccably but stiffly organized and distressingly remote from the actual documents themselves – reinforces the frustrations engendered by organization without fingertip-accessible information.

Digital libraries let you have it both ways. The technology is new – barely a decade old – and the idea that people might build libraries using end-user software is even newer and still perhaps a little startling. You can take the high road and view digital libraries as grand national or international resources, or the low road and see them as grassroots collections created by individuals or groups in response to local needs. Ultimately the two will co-exist, of course. Standard protocols will coordinate remote access to many local collections, forming a grand resource from individually and lovingly curated pieces. Other protocols will allow you to withdraw material from large centralized digital libraries and meld them with local documents to build your own collections, tailored to your own needs.

Without losing sight of the high road, this course has emphasized the low one: how you, the educator, can build your own digital library collections to use in the courses you teach – and, if you wish, get your students to build theirs, too. When our teachers were at school and university years ago, their textbooks became their friends. They spent years reading them, studying them, annotating them, getting to know them. Some have them on their shelves still – dated,

no doubt, but nevertheless a valued part of their intellectual baggage. Despite all the years that have elapsed, they refer to their old textbooks occasionally even today. They know them so well that when some long-forgotten fact or formula must be recalled, it can be located instantly.

We, the authors of this course, were lucky: we could afford to buy books and to keep them. Many students today are not so lucky – not necessarily because they can no longer afford books, but because much of the material that their teachers refer them to is on the web. Will it still be there, decades later, for them to consult so conveniently? Hardly. And many others are not so lucky because not only can they not afford books, but they have no access to the web. Digital libraries promise to be able to level out the playing field. How would it be if every student left school or university with a digital library collection of all the material they had studied, all the books, all the notes, in every course, in a convenient, portable, stable, long-lasting format? A resource for life.

These are novel ideas. And, in fact, in units like the present one where we want to inspire you by showing what others have done, we have been hard pressed to find illustrative examples. Indeed, to tell the truth, it has been difficult to find examples of any useful and inspirational educational digital libraries.

This is a future that we challenge you to create.

APPENDIX A GLOSSARY OF TERMS

ASCII	American Standard Code for Information Exchange, a 1968 standard 7-bit code for representing the Roman alphabet plus numerals and special symbols
Boolean query	Query to an information retrieval system that may contain AND, OR, NOT
Browsing	Accessing a collection by scanning an organized list of metadata values associated with the documents (such as author, title, date, keywords)
Classifier	Greenstone code module that examines document metadata to form an index for browsing
Collection	Set of documents that are brought together under a uniform searching and browsing interface
CSS	Cascading Style Sheets, a way of controlling the presentation of HTML and XML documents
Digital library	Collection of digital objects (text, audio, video), along with methods for access and retrieval, and for selection, organization, and maintenance
Document	Basic unit from which digital library collections are constructed; it may include text, graphics, sound, video, etc.
DTD	Document Type Definition: a specification used in XML (and also SGML) to express the structure of a particular set of documents
Dublin Core	An intentionally minimalist standard for describing metadata, designed to be applied to resources on the web
Encapsulated PostScript	Variant of PostScript designed for expressing graphics of a single page or less that are to be included in other documents
Format string	A string that specifies how documents and other listings are to be displayed in Greenstone
GIF	Graphics Image Format, a widely-used compression scheme for lossless images specified in 1987
GNU Public License	Software license that permits users to copy and distribute computer programs freely and modify them – so long as all modifications are made publicly available
Greenstone	The name of the digital library software used as an example in the book (greenstone.org)
HTML	HyperText Markup Language, the language in which web documents are written
HTML Tidy	Software utility that converts older HTML formats to XHTML
HTTP	Hypertext transfer protocol
Importing	Process of bringing collections of documents into the Greenstone system
Index	Information structure that is used for searching or browsing a collection
ISBN	International Standard Book Number

JPEG	Standard for (mainly) lossy image compression, named after the "Joint Photographic Experts Group"
JPEG-2000	Later version of the JPEG image compression standard
LCSH	Library of Congress Subject Headings, a controlled vocabulary for assigning subject descriptors
LOM	Learning Object Metadata standard
MARC	Machine-readable cataloguing format, a metadata scheme designed in the late 1960s for use by professional library cataloguers
Metadata	Structured information, such as author, title, date, keywords, and so on, that is associated with a document (or document collection)
METS	The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library
MIDI	Musical Instrument Digital Interface, a representation of music used by music synthesizers
Mirror	The process of copying a web site, or part of a web site, to another location and making it available there
MIME	Multipurpose Internet Mail Extensions, a standard for including different types of file – text, images, audio, video, or application-specific data – in E-mail messages
MODS	Metadata Object Description Schema
MPEG	Standard for representing multimedia material, named after the "Motion Picture Experts Group"
New Zealand Digital Library Project	Research project in the Computer Science Department at the University of Waikato, New Zealand, that created the Greenstone software (*nzdl.org*)
OAI	Open Archives Initiative, the name of a protocol designed for the efficient dissemination of digital library content
OCR	Optical Character Recognition, the process of producing a digital representation of the textual content of a document image
PDF	Portable Document Format, a page description language designed for interactive use as a successor to PostScript
Perl	Programming language used for many of the text-processing operations that occur during the Greenstone building process
Plugin	Code module for handling documents of different formats, used during the importing and building processes
PNG	Portable Network Graphics, an open standard for lossless images
PostScript	The first page description language, released in 1985
Protocol	Set of conventions according to which two systems communicate (for example, a Greenstone receptionist and collection server)

RDF	Resource Description Framework, a scheme designed to facilitate the interoperability of metadata
RTF	Rich Text Format, a standard format for interchange of text documents
SCORM	Sharable Content Object Reference Model. SCORM acts as a wrapper to various components specified using existing standards, including LOM
SGML	Standard Generalized Markup Language, a metalanguage for describing markup formats that was standardized in 1986 and forms a precursor to XML
SQL	Structured Query Language, an industry standard database query language
SVG	XML-based language for describing two-dimensional graphics
TEI	Text Encoding Initiative, a project founded in 1987 that developed SGML DTDs for representing scholarly texts in the humanities and social sciences
TIFF	Tagged Image File Format, a public-domain file format for raster images that incorporates facilities for descriptive metadata
UCS	Unicode Character Set, the set of characters supported by Unicode
Unicode	Standard scheme for representing the character sets used in the world's languages
URI	Uniform Resource Identifier, a generic name for URLs and URNs
URL	Uniform Resource Locator, a standard way of addressing objects on the web (but this term is supposed to be superseded by URI)
URN	Uniform Resource Name, a way of naming resources instead of specifying their locations
UTF	UCS Transformation Format, a scheme for representing Unicode characters with three variants: UTF-32, UTF-19 and UTF-8
XHTML	Modern version of HTML that incorporates the stricter syntactic rules of XML
XML	Extensible Markup language, a metalanguage for describing markup formats for structured documents and data on the web
XML Schema	Way of specifying the structure of a particular set of documents that provides more expressive facilities for structures and data types than DTDs
Z39.50	International standard communication protocol developed for use by library catalogue systems

APPENDIX B BIBLIOGRAPHY, JOURNALS, AND WEBSITES

The bibliography is divided into articles that we have cited in the course, including the course readings, and a bibliography of items that might be useful as further reading.

References

Apperley, M., Keegan, T., Cunningham, S.J., and Witten, I.H. (2002) "Delivering the Māori-Language Newspapers on the Internet". In: *Rere atu, taku manu! Discovering history, language & politics in the Maori-language newspapers*, edited by Jenifer Curnow, Ngapare Hopa, & Jane McRae. Auckland University Press, 211-232.

Brown, D.(2003) *Lost in Cyberspace: The BBC Domesday Project and the Challenge of Digital Preservation*. Cambridge Scientific Abstracts, June. http://www.csa.com/hottopics/cyber/oview.html

Drake, M.A. (2004) "Institutional repositories: hidden treasures". *Searcher*, 12(5), May. http://www.infotoday.com/searcher/may04/drake.shtml

Duval, E. (2004) "Learning technology standardization: making sense of it all". *International Journal on Computer Science and Information Systems*, No. 1, pp. 33–43. http://www.comsis.fon.bg.ac.yu/ComSISpdf/Volume01/InvitedPapers/ErikDuval.pdf

Duval, E. and Hodgins, W. (2003) "A LOM research agenda". *Proc International Conference on World Wide Web*, edited by Hencsey, G., White, B., Chen, Y., Kovacs L., and S. Lawrence, pp. 1–9. http://www2003.org/cdrom/papers/alternate/P659/p659-duval.html

Friesen, N., Mason, J., and Ward, N. (2003) "Building educational metadata application profiles". *Proc International Conference on Dublin Core and Metadata for e-Communities*, pp. 63–69. http://www.bncf.net/dc2002/program/ft/paper7.pdf

Gartner, R.(2002) METS: *Metadata Encoding and Transmission Standard*. Oxford University Library Services, JISC. http://www.jisc.ac.uk/uploaded_documents/tsw_02-05.pdf

Gartner, R.(2003) MODS: *Metadata Object Description Schema*. Oxford University Services, JISC. http://www.jisc.ac.uk/uploaded_documents/tsw_03-06.pdf

Marchionini, G. and Geisler, G. (2002) "The Open Video digital library". *D-Lib Magazine*, 8(12), December. http://www.dlib.org/dlib/december02/marchionini/12marchionini.html

Marchionini, G. and Maurer, H. (1995) "The roles of digital libraries in teaching and learning". *Communications of the ACM* (1995), 38(4), 67–75.

Masullo, M. and R. Mack. (1996) "Roles for digital libraries in K-12 education". *D-Lib Magazine*. 2(8), September. http://www.dlib.org/dlib/september96/eduport/09masullo.html

Mendel, J.M. (1999) "Education using Digital Libraries". In: *WTEC Panel Report on Digital Information Organization in Japan*. 13-22. World Technology Division, International Technology Research Institute, Loyola College, Baltimore, MD, USA. http://www.wtec.org/loyola/pdf/dio.pdf

Najjar, J., Ternier, S., and Duval, E. (2004) "User behaviour in learning objects repositories: an empirical analysis". *Proc ED-MEDIA World Conference on Educational Multimedia, Hypermedia and Telecommunications*, edited by Cantoni, L., and McLoughlin, C., pp. 4373–4379. http://www.cs.kuleuven.ac.be/~najjar/papers/edmedia2004.pdf

Nixon, W.J. (2003) "DAEDALUS: Initial experiences with EPrints and DSpace at the University of Glasgow". *Ariadne*, No. 37, October. http://www.ariadne.ac.uk/issue37/nixon/

Roes, H. (2001) "Digital libraries and education: trends and opportunities". *D-Lib Magazine*, 7(7/8), July/August. http://www.dlib.org/dlib/july01/roes/07roes.html

Wallace, R., Krajcik, J., & Soloway, E. (1996) "Digital libraries in the Science classroom: an opportunity for inquiry". *D-Lib Magazine*, 2(8), September. http://www.dlib.org/dlib/september96/umdl/09wallace.html

Witten, I.H. and Bainbridge D. (2003) *How to Build a Digital Library*. Morgan Kaufmann, CA.

Witten, I.H., Bainbridge, D., Tansley, R., Huang, C-Y., and Don, K.J. (2005) "StoneD: A bridge between Greenstone and DSpace". Working Paper 2005/02, Department of Computer Science, University of Waikato, New Zealand.

Bibliography

ABBYY Software (2000) *FineReader User's Guide*. ABBYY Software, 123015 Moscow, P.O. 72, Russia.

Adobe Systems Incorporated (1985) *PostScript language tutorial and cookbook*. Addison Wesley, Boston, MA.

Adobe Systems Incorporated (1999) *PostScript language reference*. Addison Wesley, Boston, MA, third edition.

Adobe Systems Incorporated (2000) *PDF reference*. Addison Wesley, Boston, MA, second edition (version 1.3).

American National Standards Institute (1968) *American Standard Code for Information Interchange (ASCII)* Standard No. X3.4–1968; updated as X3.4–1986.

Andrews, N. (1987) "Rich text format standard makes transferring text easier". *Microsoft Systems Journal*, Vol. 2, No. 1, pp. 63–67, March.

Arms, W.Y. (2000) *Digital libraries*. MIT Press, Cambridge, Massachusetts.

Atkinson, R. (1986) "Selection for preservation: a materialistic approach". *Library Resources and Technical Services 30*, pp. 344–348, October/December.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999) *Modern information retrieval*. ACM Press, New York.

Bainbridge D. and Cunningham S.J. (1998) "Making oral history accessible over the World Wide Web". *History and Computing*, Vol. 10, No. 1/3, pp. 73–81.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001) "The Semantic Web". *Scientific American*, Vol. 284, No. 5, pp. 34–43, May.

Borgman, C.L. (2000) *From Gutenberg to the global information infrastructure: Access to information in the networked world*. MIT Press, Cambridge, Massachusetts.

Bryan, M. (1988) SGML: *An author's guide to the Standard Generalized Markup Language*. Addison Wesley, Boston, MA.

Chang, S.J. and Rice, R.E. (1993) "Browsing: A multidimensional framework". *Annual Review of Information Science and Technology*, Vol. 28, pp. 231–276.

Chapman, N. and Chapman, J. (2000) *Digital multimedia*. Wiley, New York.

Chen, S.S. (1998) *Digital libraries: The life cycle of information*. BE (Better Earth) Publisher, Columbia, MO.

Committee on Intellectual Property Rights, Computer Science, and Telecommunications Board (2000) *The digital dilemma: Intellectual property in the information age*. National Academy Press, Washington DC.

Cooper, M.D. (1996) *Design of library automation systems*. Wiley, New York.

Cox, I., Miller, M., and Bloom, J. (2001) *Digital watermarking*. Morgan Kaufmann, San Francisco, CA.

Crane, G. (1998) "The Perseus project and beyond: How building a digital library challenges the humanities and technology". *D-Lib Magazine*, Vol. 4, No. 1, January.

Crawford, W. and Gorman, M. (1995) *Future libraries: Dreams, madness, and reality*. American Library Association, Chicago, IL.

Dartois, M., Maeda, A., Sakaguchi, T., Fujita, T., Sugimoto, S., and Tabata, K. (1997) "A multilingual electronic text collection of folk tales for casual users using off-the-shelf browsers". *D-lib Magazine*, Vol. 3, No. 10, October.

de Stefano, P. (2000) "Selection for digital conversion". In *Moving theory into practice: digital imaging for libraries and archives*, edited by Kenney, A.R., and Rieger, O.Y. Research Libraries Group, Mountain View, CA, pp. 11–23.

Frakes, W.B. and Baeza-Yates, R. (Editors) (1992) *Information retrieval: Data structures and algorithms*. Prentice Hall, Englewood Cliffs, NJ.

Friesen, N. (2003) "Three objections to learning objects, Learning objects and metadata". Kogan, London. http://phenom.educ.ualberta.ca/not, vert, similarnfriesen/

Gaines, B.R. (1993) "An agenda for digital journals: The socio-technical infrastructure of knowledge dissemination". *Journal of Organizational Computing*, Vol. 3, No. 2, pp. 135–193.

Gapen, D.K. (1993) "The virtual library: Knowledge, society, and the librarian". In: *The virtual library: visions and realities*, edited by Saunders, L.M. Information Today, Medford, NJ, pp. 1–14.

Giles, C.L., Bollacker, K.D., and Lawrence, S. (1998) "CiteSeer: An automatic citation indexing system". *Proc ACM Digital Libraries*, Pittsburgh, PA, pp. 89–98; June.

Goldfarb, C.F. (1990) *The SGML Handbook*. Oxford University Press, New York.

Gore, D. (Editor) (1976), *Farewell to Alexandria*. Greenwood Press, Westport, CT.

Gorman, M. and Winkler, P.W. (Editors) (1988) *Anglo-American Cataloguing Rules*, American Library Association, Chicago, IL, second edition.

Gorn, S., Bemer, R.W., and Green, J. (1963) "American standard code for information interchange". *Communications of the ACM*, Vol. 6, No. 8, pp. 422-426, August.

Harold, E.R. (2001) *XML Bible*. IDG Books, Boston, MA, Gold edition.

Hyman, R.J. (1972) *Access to library collections: An inquiry into the validity of the direct shelf approach, with special reference to browsing*. Scarecrow Press, Metuchen, NJ.

Jones, S., McInnes, S., and Staveley, M.S. (1999) "A graphical user interface for Boolean query specification". *International J Digital Libraries*, Vol. 2, No. 2/3, pp 207–223.

Kahle, B. (1997) "Preserving the Internet". *Scientific American*, Vol. 276, No. 3, pp. 82–83, March.

Kientzle, T. (1995) *Internet file formats: your complete resource for sending, receiving, and using Internet files*. Coriolis Group, Scottsdale, AZ.

Kientzle, T. (1997) *A programmer's guide to sound*. Addison Wesley, Boston, MA.

Korfhage, R.R. (1997) *Information storage and retrieval*. Wiley, New York.

Kuny, T. (1998) "A digital dark ages? Challenges in the preservation of electronic information". *International Preservation News*, No. 17; May.

Lagoze, C. and Fielding, D. (1998) "Defining collections in distributed digital libraries". *D-Lib Magazine*, Vol. 4, No. 11; November.

Lagoze, C. and Payette, S. (2000) "Metadata: Principles, practices and challenges". *In Moving theory into practice: digital imaging for libraries and archives*, edited by Kenney, A.R., and Rieger, O.Y. Research Libraries Group, Mountain View, CA, pp. 84–100.

Lagoze, C. and Van de Sompel, H. (2001) "The open archives initiative: building a low-barrier interoperability framework". *Proc Joint Conference on Digital Libraries*, Roanoke, Virginia, pp. 54-62, June.

Lesk, M. (2005) *Understanding digital libraries*. Second Edition, Morgan Kaufmann, San Francisco.

Library of Congress (1998) *Library of Congress Subject Headings*. Library of Congress Cataloging Policy and Support Office, Washington DC, 21st edition.

Lovins, J.B. (1968) "Development of a stemming algorithm". *Mechanical Translation and Computation*, Vol. 11, No. 1–2, pp. 22–31.

Lynch, C. (1999) "Canonicalization: A fundamental tool to facilitate preservation and management of digital information". *D-Lib Magazine*, Vol. 5, No. 9, September.

Mason, J., Mitchell, S., Mooney, M., Reasoner, L., and Rodriguez, C. (2000) "INFOMINE: Promising directions in virtual library development". *First Monday*, Vol. 5, No. 6, June.

Miller, E. (1998) "An introduction to the resource description framework". D-Lib Magazine, Vol. 4, No. 5, May.

Miller, P. (Editor) (2000) *D-Lib Magazine Special Issue on Collection-Level Description*, Vol. 6, No. 9, September.

Murray, J.D. and van Ryper, W. (1996) *Encyclopedia of graphics file formats*. O'Reilly and Associates, Sebastopol CA, second edition.

Nack, F. and Lindsay, A. (1999) "Everything you wanted to know about MPEG-7: Part I". IEEE Multimedia, Vol. 6, No. 3, pp. 65–77, July–September.

Nack, F. and Lindsay, A. (1999) "Everything you wanted to know about MPEG-7: Part II". *IEEE Multimedia*, Vol. 6, No. 4, pp. 64–73, October–December.

Paepcke, A., Baldonado, M., Chang, C.-C. K., Cousins, S., and Garcia-Molina, H. (1999) "Using distributed objects to build the Stanford digital library". Infobus, *IEEE Computer*, Vol. 32, No. 2, pp. 80-87, February.

Pennebaker, W.B. and Mitchell, J.L. (1993) JPEG: *Still image data compression standard*. Van Nostrand Reinhold, New York.

Pohlmann, K.C. (2000) *Principles of digital audio*. McGraw-Hill, New York, fourth edition.

Polsani, P.R. (2003) "Use and abuse of reusable learning objects". *Journal of Digital Information* Vol. 3 (4), February. http://jodi.ecs.soton.ac.uk/Articles/v03/i04/Polsani/

Price-Wilkin, J. (2000) "Access to digital image collections: system building and image processing". In *Moving theory into practice: digital imaging for libraries and archives*, edited by Kenney, A.R. and Rieger., O.Y. Research Libraries Group, Mountain View, CA, pp. 101–118.

Ranganathan, S.R. (1931) *The five laws of library science*. Madras Library Association, Madras.

Rothenberg, J. (1995) "Ensuring the longevity of digital documents". *Scientific American*, Vol. 272, No. 1, pp. 42–47, January.

Rothenberg, J. (1997) "Digital information lasts forever – or five years, whichever comes first". Rand Corporation Video V-079.

Salton, G. and McGill, M.J. (1983) *Introduction to modern information retrieval*. McGraw Hill, New York.

Salton, G. (1989) *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Prentice Hall, Englewood Cliffs, NJ.

Samuelson, P. (1998) "Encoding the law into digital libraries". *Communications of the ACM*, Vol. 41, No. 4, pp. 13–18, April.

Samuelson, P. and Davis, R. (2000) "The digital dilemma: a perspective on intellectual property in the information age". Presented at the Telecommunications Policy Research Conference, Alexandria, Virginia, September.

Sanders, L.M. (Editor) (1999) *The evolving virtual library II: Practical and philosophical perspectives*. Information Today, Medford, NJ.

Shank, J.D. (2005, to appear) "The emergence of learning objects: The reference librarian's role". *Research Strategies*.

Sperberg-McQueen, C.M. and Burnard, L. (Editors) (1999) *Guidelines for Electronic Text Encoding and Interchange*. Chicago and Oxford.

Sun Microsystems (2000) *The digital library toolkit*. Sun Microsystems, Palo Alto, CA. Available at http://www.sun.com/edu

Svenonius, E. (2000) *The intellectual foundation of information organization*. MIT Press, Cambridge, MA.

Thiele, H. (1998) "The Dublin Core and Warwick Framework: A Review of the Literature, March 1995–September 1997". *D-Lib Magazine*, Vol. 4, No. 1, January.

Unicode Consortium (2000) *The Unicode standard*, Version 3.0. Addison Wesley, Reading, MA.

U.S. Congress (1990) *Taking a byte out of history: The archival preservation of Federal computer records*. House Committee on Government Operations Report 101-987, Washington DC.

van Rijsbergen, C.J. (1979) *Information retrieval*. Butterworths, London, second edition.

Weibel, S. (1999) "The state of the Dublin Core metadata initiative". *D-Lib Magazine*, Vol. 5, No. 4, April.

Weller, M.J., Pegler, C.A., and Mason, R.D. (2003) "Putting the pieces together: What working with learning objects means for the educator". *Elearn International*. Scotland: Edinburgh, February. http://iet.open.ac.uk/pp/m.j.weller/pub/

White, J. (Editor) (1999) *Intellectual property in the age of universal access*. ACM Press, New York.

Witten, I.H., Moffat, A., and Bell, T.C. (1999) *Managing gigabytes: Compressing and indexing documents and images*. Morgan Kaufmann, San Francisco, CA, second edition.

Witten, I.H., McNab, R., Jones, S., Cunningham, S.J., Bainbridge, D., and Apperley, M. (1999) "Managing complexity in a distributed digital library". *IEEE Computer*, Vol. 32, No. 2, pp. 74–79, February.

Witten, I.H., Loots, M., Trujillo, M.F., and Bainbridge, D. (2001) "The promise of digital libraries in developing countries". *Communications of the ACM*, Vol. 55, No. 5, pp. 82–85, May.

World Bank (2000) *World development indicators 2000.* World Bank, Washington DC.

Yeates, S., Bainbridge, D., and Witten, I.H. (2000) "Using compression to identify acronyms in text". *Proc Data Compression Conference*, IEEE Press Los Alamitos, CA, p. 582.

Journals

Computers and Education

D-Lib Magazine http://www.dlib.org/

The Electronic Library

International Journal of Digital Libraries

Journal of the American Society for Information Science and Technology

Library Hi-Tech

Conferences

American Society for Information Science and Technology (ASIST) Annual Meeting

Digital Libraries for Knowledge Communities (DLKC)

European Conference on Digital Libraries (ECDL)

International Conference of Asian Digital Libraries (ICADL)

International Conference of Digital Libraries (ICDL)

Joint Conference on Digital Libraries (JCDL)

Library Information Technology Association (LITA) Forum

Websites mentioned in the course

http://google.com

http://greenstone.org

http://ltsc.ieee.org/

http://matti.usu.edu/nlvm

http://nzdl.org

http://w3c.org

http://www.cancore.ca/

http://www.careo.org

http://www.dlconsulting.co.nz/

http://www.dlese.org/

http://www.loc.gov/standards/mets/

http://www.merlot.org/

http://www.moodle.org

http://www.nzdl.org/niupepa

http://www.opensource.org

http://www.reload.ac.uk

http://www.si.umich.edu/CAMILEON/

Other recommended websites

http://acm.org/dl/
The digital library of the Association for Computing Machinery (ACM) is a large collection of academic research on computer science.

http://memory.loc.gov/ammem/
The American Memory site at the Library of Congress provides free and open access through the Internet to written and spoken words, sound recordings, still and moving images, prints, maps, and sheet music that document the American experience. It is a digital record of American history and creativity.

http://scholar.google.com/
Google Scholar is a collection of resources specifically selected for an academic audience. It includes peer-reviewed papers, theses, books, preprints, abstracts, and technical reports from many areas of research.

http://www.adlnet.org/
The Advanced Distributed Learning (ADL) Initiative, sponsored by the Office of the Secretary of Defense (OSD), is a collaborative effort between government, industry, and academia to establish a new distributed learning environment that permits the interoperability of learning tools and course content on a global scale (see Unit 3.3).

http://www.archive.org/
The Internet Archive is building a digital library of Internet sites and other cultural artefacts in digital form. This is one way in which some of the content of web sites that are removed from the Web can sometimes be retrieved.

http://www.diglib.org/
Digital Library Federation: an international association of libraries and allied institutions. Its mission is to enable new research and scholarship of its members, students, scholars, lifelong learners, and the general public by developing an international network of digital libraries.

http://www.perseus.tufts.edu/
The Perseus Digital Library aims to improve accessibility to sources for the study of the humanities. Originally its collections concentrated on classical Greek and Roman content but they are evolving into other areas.