LIS901-04

Collecting digital documents

2008–12–22

See the course web site at http://openlib.org/home/krichel/courses/lis654p09s for the latest online version of this file.

Course Description

More and more archives and libraries are trying to build collections of digital documents to support the parent institutions' mission of dissemination of information. Many of these documents have a long-run value. And many of them come as complex object. Therefore these documents require curatorial efforts that go beyond simple storage on a web site. Therefore documents are typically stored in formal repositories. Repository building and maintenance is a crucial avenue for libraries to increase relevance in the digital age.

There are three basic challenges in collecting digital docments. First there is the issue of maintaining a server to house the collection. Second there is the issue of understanding the complicated structure of digital documents. For example, a digitatized book may come as a set of scanned images, some text that has been extracted from the pages, some cover art and some metadata on the origin of the book, the particular edition and print. The book therefore is a composite object of distinct files, in different format, but that have be kept together. Third, there is the issue of managing the collection itself, and the implementation of the management through the software. The collection policy has do be precise defined and formally encoded in such a way that the collection management software can enforce it.

This is hands-on course that develops repositories for document collection from scratch. Each student will build a server and host the server at home or, as a last resort, in the Palmer School. The server hardware can come from stocks of surplus to requirement machines, or can be purchased as a barebones computer system for about $200. All software used is open-source. Broadly, there are two parts to the course. First, the students set up computers running an open source software operating system. Second the students set up an open source repository system. The repository implements a collection scenario that the student defines.

At this time, the operating system is the testing version of Debian GNU/Linux. The repository software is Fedora Commons, version 3.

Course objectives

After taking this course students

- will understand basic computer hardware and computer networking issues. This will be tested by quizzes.

- will be able to setup and maintain a Linux server for the storage of digital document. This will be measured by a rubric on the final server.

- will be able to set up a reasonably complete set of functional requirements for a repository. This will be measured by a written submission.

- will be able to translate an institution requirements into a functional system based on respository software used. This will be measured by comparing the written submission with the final server outcome.

Prerequisites

There are no other formal prerequisites for this course. However this course is not suitable for computer neophytes or technophobes.

Students should have an old computer that they can use to run the server on. The computer should be an Intel or AMD processor based PC. Installing on another computer is possible, but would add difficulty. An Intel-based computer needs to have at least a Pentium processor. It should have 500 megabytes of RAM, and 4 gigabyte of disk space. Most old computers

will do a lot better than than. The computer needs to be dedicated to the course during the run-time of the course but can be put to other usage once the course is over. The instructor will try to collect old computers for those who have difficulty finding a computer.

Students should have a network connection at home. It is best if the network goes via a cable mode connection, that leads to an Ethernet connection. The server can be connected to a router that is common in home networking scenarios. Or it can replace such a routing system. Each scenario has its own set of challenges, and the course will be addressing both. If a student does not have such a connection at home, it is possible to host the computer in the instructor's office, but public service on such computers would, bar an extensive and very complicated effort, only be visible on campus. In addition the server would be at the risk on networking policy changes at the CW Post campus.

Instructor
Thomas Krichel

Palmer School of Library and Information Science
C.W. Post Campus of Long Island University
720 Northern Boulevard
Brookville, NY 11548–1300
krichel@openlib.org
work phone: +1–(516)299–2843
skype: thomaskrichel
Private contact details may be obtained from the online CV at /home/krichel/cv.html.

Class structure
Classes are held on the CW Post campus on Saturdays between 11:30 and 16:30. The instructor promises to be there shortly after 11:00 for extra help and questions.

Most classes will have lengthy presentation by the instructor. Some class time is spent by students working directly with their computers under the supervision of the instructor. However, give the hefty weight of the class material, students are expected to do much of the work on their repositories at home. To support the students in this process, the instructor will be on campus in extra sessions for students who need additional support. Support via skype is available pretty much around the clock, i.e. unless the instructor is riding his bicycle or is asleep.

Class details:

2009–03–07	11:30 to 16:30	Important concepts of open source software, computer hardware and computer networks Inspection of ha
2009–03–21	11:30 to 16:30	Installing the operating system. Transfer of servers home.
2009–03–28	11:30 to 16:30	Installing and configuring required software. Security and backup. Project plans and service models.
2009–04–04	11:30 to 16:30	Setup of repostitory software.
2009–04–18	11:30 to 16:30	Configuring and collecting.
2009–04–25	11:30 to 16:30	Presentation of finished work.

Slides for all classes are downloadable from the course web site. The slides on the course website are drafts until the time that the class is held.

Assessment

Mailing list
There will be a mailing list for the course at https://lists-1.liu.edu/mailman/listinfo/cwp-lis654-krichel. All students are encouraged to subscribe. As a rule, answers to email sent to the instructor are copied to the list. There are exceptions to this rule

- if the question writer requests the answer not to be posted;

- if the question is a purely private matter.

Literature

There is no text for this course. The expertise acquired in the course is very difficult to find in existing literature because it spans a wide area of subjects. Some Internet sites include

- Debian – Debian-Installer

- DebianInstaller - Debian Wiki

-

Formally authored documents include

- Maurice Vanderfeesten, Friedrich Summann, and Martin Slabbertje, (editors) "The Driver Guidelines version 2.0", available at `http://www.driver-support.eu/documents/DRIVER_Guidelines_v2_Final_2008-11-13.pdf`

- Mary R. Barton, "Project Planning Matrix", available at `http://www.dspace.org/images/stories/service-model.pdf`

- Mary R. Barton, "Defining a Service Model", available at `http://www.dspace.org/images/stories/project-plan.pdf`

- Javier Fernández-Sanguino Peña, "The Debian GNU/Linux and Java FAQ", available at `http://www.debian.org/doc/manuals/debian-java-faq/`.

- W. Martin Borgert, "Debian GNU/Linux Reference Card. The 101 most important things when using Debian GNU/Linux", available at `http://xinocat.com/refcard/refcard-en-lt.pdf`

- The Fedora Development Team, "Tutorial #1: Introduction – Basic Concepts in Fedora (Fedora 3.0)", available at `http://www.fedora.info/documentation/3.0/userdocs/tutorials/tutorial1.pdf`

- The Fedora Development Team, "Tutorial #2: Getting Started – Creating Fedora Objects and Using the Content Model Architecture (Fedora 3.0)", available at `http://www.fedora.info/documentation/3.0/userdocs/tutorials/tutorial2.pdf`

Assessment

There will be a quiz at the beginning of each session except the first. These will count for 40% of the grade. By the third session, students will hand in a description of basic functionality of the installation that they will built, based on the "stories" in the papers by Mary R. Burton. The repository is most likely fictions, but real collections could also be built. This will count for 20% of the grade. By the fifth session, the instructor will publish a set of requirements that the DSpace installation will have to conform to. These will be assessed on the server by the week following the class. This counts for 20% of the final grade. Finally class participation is a crucial component of the course. This is includes demonstrated ability to build their own server as well as willingness and capability to help struggling comrades with their servers. This will be counted for 20% of the grade.