Content Management at Grainger Engineering Library

Case studies from various digital library research projects

Tom Habing
thabing@uiuc.edu
Outline

• Introduction

• Case studies
  – DeLiiver project
  – Open Archives Initiative (OAI) projects
  – Simultaneous Search
  – Institute of Physics (IoP) archive

• Challenges / Conclusion
Intro: DL Research Focus

• Engineering / Scientific Resources
  – Access and Discovery
  – Full-Text
    • Rendering (especially mathematics, MathML)
    • Markup-based (SGML / XML)
    • Standard Tools: XSLT, XML Schema
  – Metadata
    • Schemas: MARC, RDF, DC
  – Linking
    • DOI, OpenURL
  – Search
    • Distributed databases (Grainger Search Aid)
    • Aggregated databases (Open Archives Initiative)
Intro: The Digital Library

• ‘Digital’, ‘Virtual’, ‘Electronic’ Library as network-based library without regard to place and time.
• Tendency to apply term to collections and resources.
• Digital Collections vs. Digital Library.
• Emphasis on the integration of collections and services (NSDL).
• Application of standards and protocols is important.
DeLiiver
http://dli.grainger.uiuc.edu/

• Testbed funded under DLI-I by NSF, DARPA, and NASA, 1994--1998. Awards made to 6 universities.

• Large-Scale testbed, distributed repository models, evaluation, web software.


• Collaborating Partners Program. AIP, APS, ASCE, IEE, NRL, ASM, ACM, NTT Learning Systems, Elsevier.
DeLLiver - Testbed

• American Institute of Physics--APL, JAP, RSI
  – 16,000+ articles, 1995--.
• American Physical Society--PRL
  – 10,000+ articles, 1995--, weekly updates.
• ASCE Journals (25 titles)
  – 9,000+ articles, 1995--.
• IEE Proceedings and Electronics Letters
  – 8,500+ articles, 1993--.
• ACM (Association for Computing Machinery).
• Elsevier Science.
DeLlver - Project Objectives

• Construct large-scale, multipublisher, markup-based full-text journal testbed.
• Investigate processing, indexing, normalization, retrieval, rendering and linking.
• Study end-user searching behavior and needs.
• Develop one-stop-shopping retrieval techniques (Aggregation, Resource Linking).
• Identify models for effective retrieval in distributed repository environment.
DeLIver - Accomplishments

• Process and retrieve from multiple publishers and heterogeneous DTDs.
• Cross-repository searching.
• SGML to XML conversion.
• Metadata extraction, representation, merging.
• Transformation and rendering technologies.
• Dynamic linking: forward/backward, from/to A & I services.
• End-user studies
DeLiver – Workflow

• SGML files from publishers (FTP, CD, Tape)
• Convert to XML
• Extract and process metadata using custom scripts and XSLT
  – Create reference links
  – Normalize
• Process mathematics
• Build search indices
• Move files to web server
• Tape backups
Deliver - Demos

- http://dli.grainger.uiuc.edu
- http://dli.grainger.uiuc.edu/~asm/
- http://dli.grainger.uiuc.edu/~acm/
DeLIver – Details

• Web Server
  – Dell PowerEdge 4300, Dual Pentium II, 512 MB
  – 145 GB across 5 HDs
    • ~80 GB used by DeLIver content
  – Windows 2000 (just upgraded from NT)
  – IIS 5.0 (Active Server Pages, VBScript)
  – Access is controlled via campus Bluestem service
DeLiver - Details

• Database Server
  – HP 9000 J200, HP-UX
  – OpenText LiveLink database for full-text search capability
  – Older Netscape web server CGI application

  – Also MS SQL Server for metadata only search
Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)

http://oai.grainger.uiuc.edu
http://www.openarchives.org

• HTTP and XML based protocol
• Data providers
  – Share metadata about their collections
• Service providers
  – Harvest the metadata and use it to develop different services (i.e. search portal)
OAI-PMH - Demos

• Grainger Data Provider
  – http://g118.grainger.uiuc.edu/engdocoai/oai.asp

• Service Providers
  – http://oai.grainger.uiuc.edu/CandI303/search/
OAI-PMH - Details

• Data Providers
  – Many open source toolkits for various platforms
  – We have developed both ASP and JSP implementations
  – Metadata can reside in various databases, as XML files on a file system, or a combination
OAI-PMH - Details

• Service Providers
  – Also various open source implementations of OAI harvesters
  – Cultural Heritage search is running on Dell PowerEdge 4600, 4 GB Ram, 180 GB Disk, RedHat Linux 7.3, U. Michigan DLX Software.
  – Engineering search is running on a Dell Poweredge 6300, Quad Pentium, IIS ASP application, MS SQL Server database
Grainger Search Aid
http://g118.grainger.uiuc.edu/searchaid/nopx3.asp

• Distributed search across multiple resources with a common interface
  – Google, Library Catalog, A & I databases

• Integrating A & I services with full-text resources (OpenURL, DOI)
Institute of Physics (IoP) Archive
http://gita.grainger.uiuc.edu/iop/

• Recently acquired a local copy of the full text of the IoP archive back to 1874
  – PDF, XML Metadata, GIF and JPEG Images
  – 550,000 files in 160 GB

• Integrated with the OAI search interface

• How to integrate this with the DeLlver material?
Misc. Challenges

• Full-text rendering across different browsers, especially SciTech material with math and special characters (MathML)
• Integrating heterogeneous resources
• Maintaining code across software and OS updates
• Typical source code control issues, especially for research projects which are transitioned into production
Conclusion

• For a digital library the biggest challenge isn’t managing one’s own content (although this is still a big challenge), but integrating, managing, and making accessible different content from a wide variety of sources, many of which are outside your direct control.

• XML and related standards are helping enormously.

• Many other standards such as DOI, OpenURL, OAI are also critical to the problem.