Moving Web Archiving into the Classroom

Edward A. Fox
Digital Library Research Laboratory
Virginia Tech
fox@vt.edu
http://fox.cs.vt.edu

Digital Curation Curriculum Conference, Chapel Hill, NC, April 1-3, 2009

Outline

- Graduate Student Studies
- Undergraduate Honors Class
- DL Curriculum Project
- CS6604, Digital Libraries, Fall 2008
- Curatorial Work and Learning in Virtual Environments

Graduate Student Studies

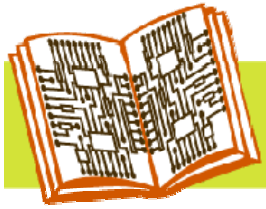
- Li Wang, "Crawling on the World Wide Web", June 2002, MS independent study report, http://eprints.cs.vt.edu:8000/archive/00000572/
- Yuxin Chen, "A Novel Hybrid Focused Crawling Algorithm to Build Domain-Specific Collections", Feb. 2007, PhD dissertation, http://scholar.lib.vt.edu/theses/available/etd-02162007-005107/

Undergraduate Honors Class

- http://collab.dlib.vt.edu/runwiki/wiki.pl?HonorsApril16
- UH3004 – Students mostly juniors, seniors
- “A Digital Library Related to 4/16/07 at Virginia Tech” in Spring 2008
- http://www.dl-vt-416.org/
- Found seeds to help with Internet Archive crawl
- Projects with media digitization and archiving
- Working with Web and DL systems

Digital Library Curriculum Project

- http://curric.dlib.vt.edu/
- With funding from the National Science Foundation, grants IIS-0535057 (to Virginia Tech) and IIS-0535060 (to the University of North Carolina at Chapel Hill)
- PI: Edward A. Fox; GRA: Seungwon Yang, at Virginia Tech
- Co-PIs: Barbara M. Wildemuth, Jeff Pomerantz; GRA: Sanghee Oh, at UNC-CH

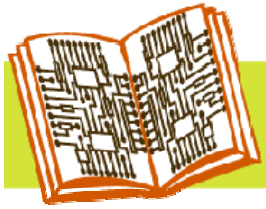


DIGITAL LIBRARIES
Curriculum Development

Curriculum Framework, 1

CORE TOPICS

1	Overview	1-a (10-c): Conceptual frameworks, models, theories, definitions	1-b: History of digital libraries and library automation
2	Digital Objects	2-a: Text resources 2-b: Multimedia 2-b (1): Images	2-c (8-c): File formats, transformation, migration
3	Collection Development	3-a: Collection development selection policies 3-b: Digitization 3-c: Harvesting	3-d: Document and e-publishing presentation markup 3-e (7-e): Web (push) publishing 3-f (7-f): Crawling
4	Info/ Knowledge Organization	4-a: Information architecture (e.g., hypertext, hypermedia) 4-b: Metadata 4-c: Ontologies, classification, categorization	4-d: Subject description, vocabulary control, thesauri, terminologies 4-e: Object description and organization for a specific domain
5	Architecture (agents, mediators)	5-a: Architecture overviews 5-b: Application software 5-c: Identifiers, handles, DOI, PURL	5-d: Protocols 5-e: Interoperability 5-f: Security



DIGITAL LIBRARIES
Curriculum Development

Curriculum Framework, 2

CORE TOPICS

6	User Behavior/ Interactions	6-a: Info needs, relevance 6-b: Online information seeking behavior and search strategy	6-c: Sharing, networking, interchange (e.g., social) 6-d: Interaction design, usability assessment 6-e: Info summarization and visualization
7	Services	7-a: Search engines, IR, indexing methods 7-a (1): Image retrieval 7-b: Reference services 7-c: Recommender systems	7-d: Routing, community filtering 7-e (3-e): Web (push) publishing 7-f (3-f): Crawling 7-g: Personalization
8	Preservation	8-a: Approaches to archiving and repository development 8-b: Web archiving	8-c: Sustainability 8-c (2-c): File formats, transformation, migration
9	Management and Evaluation	9-a: Project management 9-b: DL case studies 9-c: DL evaluation, user studies 9-d: Bibliometrics, Webometrics	9-e: Intellectual property 9-f: Cost/economic issues 9-g: Social issues
10	DL education and research	10-a: Future of DLs 10-b: Education for digital librarians	10-c (1-a): Conceptual framework, theories, definitions 10-d: DL research initiatives

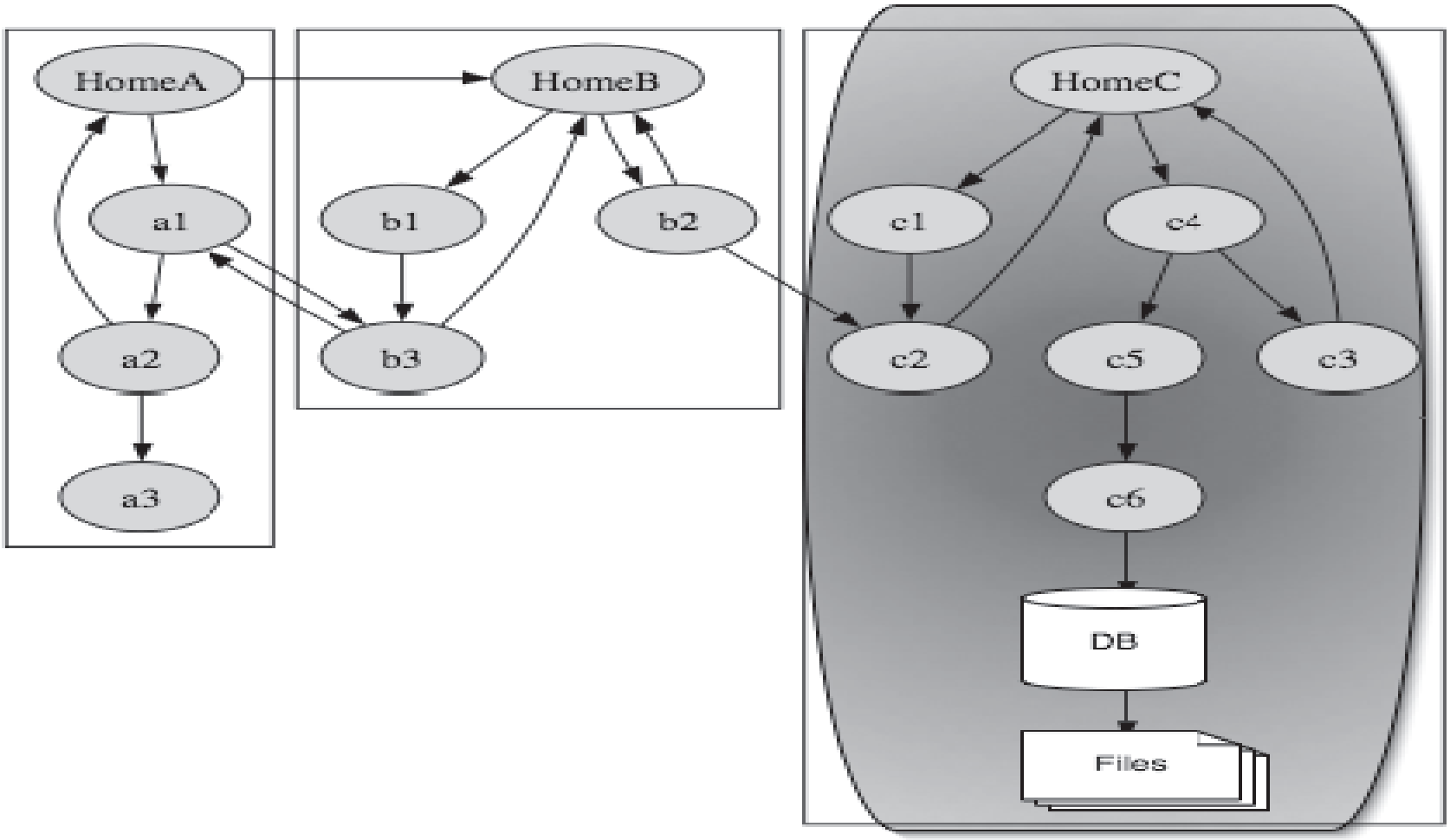
CS6604 Digital Libraries

- Fall 2008; Ph.D. students mostly; some MS students; 16 students, 4 teams
- Covered 13 modules from DL curriculum, 1 presented by instructor
- Each team
  - Presented 2 modules over 2 class sessions
  - Prepared 1 new module, and presented it in 1 class session
- 2 modules relate (see next, plus “File formats, transformation, and migration”)

Module 8b: Web Archiving

Team 3:
Jian Jiao,
Tarek Kan’an,
Spencer Lee

Intensive Archiving (Shaded Area)



Interested parties

- Schools and libraries have an interest in educating the next generation of creators of information and knowledge by providing them with access to the documentary record; this means access based on the need to learn rather than on the ability to pay.

Three Main Questions

- **When building a Web Archive the problems translate into three questions:**
  - What should be collected?
  - How do we preserve its authenticity?
  - How do we build the technology needed to access and preserve it?

Learning Activities

Discussion activity (15 minutes): In class, break students into groups of 3~4.

a. Why do we need to archive web pages and what should be collected?

b. Have them discuss Web archiving regarding:

- Personal perspective
- Commercial perspective
- Academic perspective
- Social impacts
- Technological impacts

* 10 minutes for discussion, 5 minutes for brief presentation. (1~2 minutes per group, no slide required)

Digital Library Curriculum Development
Module 8-b: Web archiving

- **1. Module Name:** Web archiving
- **2. Scope:** This module covers the general ideas, approaches, problems and needs of web archiving to build a static and long term collection consisting of web pages.
- **3. Learning objectives:** Students will be able to:
  - Explain basic concepts and methods related to Web Archiving.
  - Explain simple archiving strategies using Web Archiving approaches and overcoming current difficulties.
  - Describe the necessity (for example, everyday losses of some web pages, etc.) and problems (for example, cultural, economic, and legal) of Web Archiving.

4. 5S characteristics of the module

- Stream: Collect data and ingest into Digital Library
- Structure: The archiving process should follow certain structures according to where the gathering process happens. The organization and storage of archived data also involves particular structures.
- Space: The physical storage for keeping this archived data. Also, the varied distributed locations to ensure no single physical event destroys all copies.
- Scenario: Process of collecting and saving web content
- Society: Individuals, groups and organizations involved in setting policy and in carrying out archiving policies

- **5. Level of effort required:**
  - Prior to class: 4 hours for readings
  - In class: 2 hours
- **6. Relationship with other modules:** Close connections with:
  - 8-a: Preservation. The Web archiving module follows the preservation module. Students should know how web archiving supports preservation of digital objects.
  - 2-c (8-c): File formats, transformation, migration
  - 9-e: Intellectual property
  - 9-f: Cost/economic issues

9. Body of Knowledge

- **Definition and Problems:**
- Why archive the Web?
- What is to be collected?
- Acquisition methods
- Organization and storage
- Quality and completeness
- Scope
  - Site-Centric Archiving
  - Topic-Centric Archiving

9. Body of Knowledge – cont'd

- **Methods and Approaches**
- **Difficulties and Limitations**
- **Selection for Web Archives**
- **Copying Websites**
- **Mining Web Collections**
  - Materials for Web Archives
  - Use cases

10. Resources

- Required
  - Lyman, P. (2002). Archiving the World Wide Web. Building a National Strategy for Preservation: Issues in Digital Media Archiving. Council on Library and Information Resources, Page 38-51.
  - Masanes, J. (2005). Web Archiving Methods and Approaches: A Comparative Study. Library Trends, Vol. 54, No. 1, Summer 2005
  - Masanes, J. (2006). Web archiving: issues and methods. In J. Masanes (Ed.), Web archiving., Berlin Heidelberg New York: Springer, page 1-46

13. **Evaluation of learning achievement**

- In their answers to the discussion questions, students demonstrate an understanding of
  - Different Web Archiving perspectives
  - Different Web Archiving problems and limitations
  - Why Web Archiving is needed

Curatorial Work and Learning in Virtual Environments

- Support for community (incl. IEEE TCDL)
  - Posters from here and JCDL 2009
- Education, teaching and training related to digital preservation
- PI: Gary Marchionini; GRA: Javier Velasco-Martin, at UNC-CH, IIS-0910465
- Co-PI: Edward A. Fox; GRA: Spencer Lee, at Virginia Tech, IIS- 0910183

Summary

- Graduate Student Studies
- Undergraduate Honors Class
- DL Curriculum Project
- CS6604, Digital Libraries, Fall 2008
- Curatorial Work and Learning in Virtual Environments
  - How can we use SL in support of digital curation and preservation, especially related to Web Archiving?