CSCI B669 Scientific Data Management and Preservation

Meets: Wednesdays, 5:30-8:00, 919 E. 10th St. Informatics East 122
Instructor: Beth Plale
Office: 919 10th St., Informatics East 256
Email: plale@cs.indiana.edu
Office phone: 812 855 4373
Home phone: 812 330 0737 (before 9pm)
Office hours: Tue 2:00-3:00 p.m. or by appointment (schedule with Robert Ping robping@indiana.edu)

Credits: 3

Course Overview: Computers, sensors and networks have fundamentally changed the nature of scientific research, going so far as to introduce data-driven research as a 4th paradigm of science. Digital data produced from computation oriented science is not a commodity that is consumed in a single use, but an important and invaluable intellectual asset that can be used repeatedly to fuel new ideas and insights. Managing research data for the long-term, ensuring its continued access, has emerged as a major challenge. But as the well known 2003 “Atkins report” states, “absent systemic archiving and curation of intermediate research results… data gathered at great expense will be lost”. In this course we examine the full lifecycle of digital data with a focus on data generated and used in the course of advancing scholarship and science. That is, research data and data products from their creation through use to long term archival. This lifecycle is sometimes referred to as “data curation”, which has been defined as “the activity of managing data from their point of creation to ensure they are available for discovery and reuse in the future.”

The course is divided into five major sections. In the first part, we look at the larger factors motivating the increasing emphasis on long-term availability of research data. In the second part we examine motivating applications taken from the environmental sciences and digital humanities. In the third part we examine contributing technologies for data curation. In the fourth part we examine systems and trends that are present solutions to persistent availability and accessibility of data. In the fifth part we examine trends that impact solutions in the future.

This is a reading intensive seminar that covers a large and broad body of material. Most of the readings and content of the course are provided below, though you should expect readings will be added. The later weeks are intentionally left vague so as to remain flexible to student’s needs and interests. Shifts in schedule may occur depending on the availability of guest lecturers. While the majority of the students are expected to have a computer science background, not all will. Students should expect to feel more comfortable with some material and less so with others.

Learning Objectives:

- Develop an appreciation of the purpose, interdisciplinary nature, and value of data curation as an integral part of the research process.
- Gain understanding of the motivations driving current interest in data
- Acquire knowledge of key issues in managing and providing access to data over its lifecycle

---

1 This definition from syllabus of SI 754 Data Curation course, University of Michigan Fall 2009, Ann Zimmerman instructor.
• Acquire in-depth knowledge of technical solutions in existence that address the data storage and access, and a conversational grasp of the strengths and weaknesses of each.

Requirement and Expectations:
• Attendance in all course sessions is expected. If you have a conflict that will result in missing class, please notify me and complete work in advance. Only in extenuating circumstances should an extension be asked for.
• Complete required readings at the level of preparation thorough enough to discuss and critique readings for each week.
• Complete all assignments on time.

Evaluation:
Class participation ...... 45%
Mid-term paper .......... 25%
Final project ............. 35%

Class Participation:
Evaluation of class participation will be based on attendance and completion of in-class activities. The latter will consist of things such as in-class writings, reports from in-class group work, etc.

Assignments:

Mid-term paper: the midterm paper will be on a topic covered in the course, and will be 5 pages in length, single spaced, 11 pt. font. It will contain at least 10 high quality references (no citations to unpublished content or to web pages.) It will be a single author paper.

Final project: the final project will be on a topic of the student’s choosing. The instructor will provide projects from which the student can choose or the student can choose a project based on their research. The project will likely have a technology component, and possibly a programming component, but it need not if the student is not comfortable with programming and has a sufficiently interesting project thesis. Students can work in teams on the final project, but the amount of work needs to be equivalent to what one would do alone.

Course Schedule
The 4th Paradigm: Data Intensive Science Discovery text is available for free download, however the bound version has beautifully crisp illustrations. Readings will additionally come from the Scientific Data Management text. This book, priced between $65 and $45 is a good one for your bookshelf. Other readings not easily available for download will be posted to Oncourse.
Week 1 (Jan 12): Overview and Broad Motivation

Readings:


Optional:


Week 2 (Jan 19): Data Centric Computing: Earth and Environmental sciences

Readings:

Gray’s laws: database-centric computing in science, Alexander S. Szalay, José A. Blakeley, 4th Paradigm, pp. 5-12.


Instrumenting the earth: next-generation sensor networks and environmental science, Michael Lehning, Nicholas Dawes, Mathias Bavay, Marc Parlange, Suman Nath, Feng Zhao, from 4th Paradigm, pp. 45-51.

Optional:

A Comparison and Critique of Eucalyptus, OpenNebula and Nimbus, Peter Sempolinski & Douglas Thain, 2nd IEEE Int'l Conference on Cloud Computing Technology and Science (CloudComm), Indianapolis, IN 2010

Week 3 (Jan 26): Data Centric Computing: Scholarly Communication

Readings:

Jim Gray’s fourth paradigm and the construction of the scientific record, Clifford Lynch, 4th Paradigm, pp. 177-183

Text in a data-centric world Paul Ginsparg. 4th Paradigm, pp. 185-191.

The future of data policy Anne Fitzgerald, Brian Fitzgerald, Kylie Pappalardo, 4th Paradigm, pp. 201-208.
Week 4 (Feb 2) : Data Lifecycle - Provenance

Readings:


Optional:

Open Provenance Model v1.1 http://eprints.ecs.soton.ac.uk/21449/


Week 5 (Feb 9) : Data Lifecycle – Data Formats and Lifecycle Models

Guest Speaker: Stacy Kowalczyk, Instructor, Spring 2011 SLIS 31401 Data Curation

Readings:


Interoperability and Data Integration in the Geosciences, Michael Gertz, Carlos Rueda, and Jianting Zhang, from Scientific Data Management, Ch. 10, pp. 369-398.

Week 6 (Feb 16) : Data and Semantics on the Web

Readings:


Semantic eScience: Encoding Meaning in Next-Generation Digitally Enhanced Science, Peter Fox and Jim Hendler, 4th Paradigm, pp. 147-152

Ontologies : A contribution to the DL/DB debate, Nadine Cullot, Christine Parent, Stefano Spaccapietra, and Christelle Vangenot, The first International Workshop on Semantic Web and Databases, Sept 2003

Optional:
Relational Databases on the Semantic Web
http://www.w3.org/DesignIssues/RDB-RDF.html

Week 7 (Feb 23) : Data Lifecycle – Contributing Technologies

Demos of Apache Solr and GeoPortal.

Readings:
Apache Solr service http://wiki.apache.org/solr/FrontPage

Discovering Emergent Behavior from Network Packet Data: Lessons From the Angle Project, Robert L Grossman, Michal Sabala, Yunhong Gu, Anushka Anand, Matt Handley, Raymond Sulo and Lee Wilkinson, in Next Generation Data Mining, edited by Hillol Kargupta, Jiawei Han, Philip S Yu, Rajeev Motwani and Vipin Kumar, CRC Press, Boca Raton, 2009, pages 243-260.

Week 8 (Mar 2) : Persisting Data On the Web

Guest Speaker: Scott Jensen, Postdoctoral Fellow, Data To Insight Center.

Scott will be speaking on the XMC Cat metadata catalog. His talk will include demo. http://pti.iu.edu/d2i/xmccat

Readings:


*** Friday Mar 5: Assignment 1 due at 5:00 p.m. ***

Week 9 (Mar 9) : Persisting Data in eScience

Readings:
Parallel Data Storage and Access, Robert Ross, Alok Choudhary, Garth Gibson, And Wei-keng Liao, Scientific Data Management, pp. 35-72.

Week 11 (Mar 16) : Spring break week – no class

Week 10 (Mar 23) : Data Persistence and Movement in eScience

Readings:
    Dynamic Storage Management, Arie Shoshani, Flavia Donno, Junmin Gu, Jason Hick, Maarten Litmaath, and Alex Sim, Scientific Data Management, pp. 73-114.


Week 12 (Mar 30): Archiving and Preservation Repositories

Readings:


Week 13 (Apr 6): Persisting Data to Trusted Repositories

Readings:

Week 14 (Apr 13): Policy and Economic Issues in Data Preservation

Readings:


Week 15 (Apr 20): Future Directions

Readings:


*** Apr 22: Assignment 2 due at 5:00 p.m. ***

**Week 16 (Apr 27): Student Presentations**

**Academic Integrity**

Academic integrity requires that students take credit only for their own ideas and efforts. Misconduct, including cheating, fabrication, plagiarism, interference, or facilitating academic dishonesty, are prohibited because they undermine the bonds of trust and cooperation among members of this community and between us and those who may depend on our knowledge and integrity. Complete details are contained in the Indiana University *Code of Student Rights, Responsibilities, and Conduct*. 