IST400/IST600 Science Data Management

Spring 2009

Course Time & Location:
3:45–5:05 p.m. MW
117 Hinds Hall
First Day of Class: January 12th

Instructor Information:
John D'Ignazio
Office: 114E Hinds Hall (in Student Services Suite)
Hours: Wednesdays, 1:00–3:00 p.m.
Phone: 443-5603
Email: jadignaz@syr.edu

Course Description

The Science Data Management course includes three modules:

1. fundamentals of science data and data management,
2. data management in aggregation, and
3. broader issues of science data including tools for management and visualization, as well as quality and publication practices.

The first module provides an overview of science data and data management, including data fundamentals such as forms, scales, types, and levels, data structures and models, data formats, and databases used to store, retrieve, and manage the data. The second module uses case studies regarding data collection, processing, transformation, and management to understand aggregations of data at three levels of organization: research, resource, and reference collections. In the last module, students will be introduced to methods and tools for evaluating data quality and working with data for use in various practice communities. Throughout the semester, students will work in an interdisciplinary team on a comprehensive science data management project under the instructor's guidance. The performance of each student is based on exercises, quizzes, group reports, class participation, and the course project.

This course has been developed as part of the iSchool's Science Data Literacy project with funding from the National Science Foundation's Course, Curriculum and Laboratory Improvement program. For more information about the project, please visit the project website http://sdl.syr.edu.

Course Objectives

At the end of the course, students are expected to understand:

- concepts and characteristics of science data and practice
- principles and practices in data management and use
- technologies used to manage and use data
- procedures and methods of using data for inquiry

At the end of the course, students are expected to have hands-on experience in:

- identifying the needs for organizing, reporting, and managing science data
- describing data sets by using science metadata standards
- designing databases and systems for science data management and use
- evaluating data quality

Who Should Take the Course?

Undergraduate juniors/seniors or graduate students in any science, engineering, and information technology major may take this class. If you find you fit into the following description, this course would be for you:

- an interest in a career in science data management or science and engineering research in the corporate, academic, or government sector
- basic computer skills, including creating simple Web pages, using spreadsheets (e.g. Excel) and/or database (e.g. Access) software, and others
- willingness to work in an interdisciplinary team

Required Readings

There is no required textbook, but required readings will be available in via WebCT as electronic documents for reading and printing or distributed in print form in class. Students are responsible for familiarizing themselves with assigned materials before class to participate in discussions and apply the material in course assignments and projects.

Grading

The work for this class will involve a mixture of quizzes, individual assignments, group reports and a final project:

- **Quizzes** (3 x 5% = 15%) are designed to test your understanding of basic concepts in science data management.
- **Exercises** (4 x 7% = 28%) are designed for you to practice the necessary skills in carrying out data management project.
- **Group reports** (2 x 6% = 12%) are designed to maximize the usefulness of group discussion results.
- **Final project** (30%): the project team will be formed early on in the semester and the team members will work together throughout the semester on many tasks.
- **Participation** (15%) includes your attendance and participation in class discussions and activities.

Course Management and Expectations

I try to make every class worth attending. Students will be responsible for all material covered, handed-out, announced, etc. in class unless told otherwise. Attempts will be made, however, to place important announcements in class and/or on the class web page. Class will be held on all days as scheduled, unless notified otherwise.

Every attempt will be made to return assignments in a timely fashion. Assignments are due at the start of class, unless specified otherwise, and will be annotated with the point count for individual components of the assignment. Late work will be accepted only for two days after the due date, with a 5% penalty per day. This is to facilitate the timely return of graded assignments with answers.

This syllabus (including course requirements, due dates, etc.) may be changed with sufficient notice.

Grading Policy

- Each assigned work will be graded on the scale as specified for the component (e.g. each exercise will receive a maximum points of 7), which will be summed at the end of the semester.
- Grade levels follow the scales below: A = 94-100, A- = 90-93.9, B+ = 85-89.9, B =

80-84.9, B- = 75-79.9, C+ = 70-74.9, C = 65-69.9, C- = 60-64.9, F = below 60

- An incomplete grade, **I,** can be given only if the circumstances preventing the on-time completion of all course requirements were clearly unforeseeable and uncontrollable. If an incomplete is required a written contract must be completed which specifies the nature of the missing work, the date it will be completed, and the default grade that will be given if that deadline is missed.
- It is unethical to allow some students additional opportunities, such as extra credit assignments, without allowing the same options to all students.
- Failure to complete any course requirement will result in a course grade of **C** or lower, regardless of the grades received in other components.
- When a dispute over unequal group participation occurs, group-based assignments will have a component of the final grade based on each group member's assessment of the contribution made by the others in the group.
- To discuss a grade, arrange for a private meeting or attend the instructor's office hours in which you identify the sources of your concern. It is important to bring with you to that meeting the relevant materials (e.g., marked papers). Except for extraordinary circumstances, no appeal for an individual assignment or project will be considered later than two weeks after the graded assignment was returned. For final grades, no appeal will be considered after 5/15/2009.

Attendance

- Attendance in all class sessions (and throughout each session) is expected, exactly as it would be on a job. If an emergency or illness occurs, have someone notify your team and the course instructor as soon as possible—even if you are out of town. If you are going to be absent from class or from team meetings you need to arrange to catch-up on what you missed and to make sure your part of the workload is covered.
- Too many absences are sufficient cause to lower the final course grade. Exceptions will be made for emergencies and other extenuating circumstances provided they are verified by appropriate documentation that is received no later than 1 week after the absence(s).

Academic Integrity

The academic community of Syracuse University and of the School of Information Studies requires the highest standards of professional ethics and personal integrity from all members of the community. Violations of these standards are violations of a mutual obligation characterized by trust, honesty, and personal honor. As a community, we commit ourselves to standards of academic conduct, impose sanctions against those who violate these standards, and keep appropriate records of violations. The academic integrity statement can be found at: http://supolicies.syr.edu/ethics/acad_integrity.htm

Disability Statement

In compliance with section 504 of the Americans with Disabilities Act (ADA), Syracuse University is committed to ensure that "no otherwise qualified individual with a disability…shall, solely by reason of disability, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program or activity…" If you feel that you are a student who may need academic accommodations due to a disability, you should immediately register with the Office of Disability Services (ODS) at 804 University Avenue, Room 308 3rd Floor, 315.443.4498 or 315.443.1371 (TTD only). ODS is the Syracuse University office that authorizes special

accommodations for students with disabilities.

Schedule

Semester	Course Date	Topics	Activities	Readings / Handout
Module 1: Fundamentals of science data and data management				
Week 1	12-Jan	Introduction to the course	Pre-course assessment survey	Carlson, S. (2006); Guterman, L. (2001); Howe. D. (2008)
	14-Jan	Science data life cycle: collect, store, retrieve, use, present		Anderson, W.L. (2004) / How to identify datasets on the Web and exercise 1 requirements
Week 2	19-Jan	Martin Luther King, Jr. Day. No class		
	21-Jan	Intro to databases & database programs & data attributes		Rob, P. and C. Coronel (2004), chapter 1, pp. 6-25; chapter 2, pp. 30-65; Rob, P. and C. Coronel (2004), chapter 3, pp. 76-109 / on WebCT: Examples of data sets in relational databases,
Week 3	26-Jan	Data relationships	Due: Exercise 1 Identifying datasets In class presentations: Share your information analysis of data repository/dataset	NSF (2007); Cleaveland, J. C. (1986) chapter 1; Scheiner, S. M. (2004) chapter 3
	28-Jan	Fundamentals about data: Forms, Scales, Types, Levels Data structures and models: Physical data, Model data	In class presentations: Share your information analysis of data repository/dataset Quiz 1: Data and database fundamentals (Th-Fri)	NSB (2005), chapters 2&3; Brooks, A. A. (1984) chapter 2 Lab: Designing a database

Week 4	2-Feb	Data formats: Data format standards, Representation of data, Communication of data, Markup languages/Metaformats		Abiteboul, S., P. Buneman, et al. (2000). Chapter 2: A Syntax for Data. & Chapter 3: XML. Data on the Web.
	4-Feb	Describing datasets (1): introduction to Metadata: aboutness, types, elements, schemes, standards	In-class group work: investigate CSDGM profiles Due: Exercise 2 Database design	Zeng, M.L. & J. Qin (2008) Chapter 1; Caplan, P. (2003). Chapter 14; McGlamery, P. (2004). Chapter 13
Week 5	9-Feb	Describing datasets (2): Metadata's role in resource management,		Zeng, M.L. & J. Qin (2008), Chapters 3&4
	11-Feb	Describing datasets (3): metadata elements in schemes, tools. Case study	In-class group work: form use for data description	Hillmann, D. I. and E. L. Westbrooks (2004): Hill & Janée Chapter 8; Arms & Arms Chapter 14
Week 6	16-Feb	Describing datasets (4): Relationship-Making and Relational Database lab	In-class work: Describe dataset & create an extension	Bose, R. and J. Frew (2005)
	18-Feb	Managing data (1): Encoding, Storage, Import/Export, Cleaning, Transformation	Due: Exercise 1 & 2 Revision Quiz 2: Data formats and description (Th-Fri)	Gleason, D. (1997). Chapter 13: Data Transformation; NRC (1995), Chapter 2: Preservation and Use
Week 7	23-Feb	Managing data (2): Data Querying & Retrieval SQL Lab	Practice with SQL statements and data reports	Rob, P. and C. Coronel (2004), chapter 6, pp. 226-316
	25-Feb	Managing data (3): Ownership and access, Data quality		Statistics Canada (2001); Dasu, T. & T. Johnson (2003), chapter 4, pp. 99–137.

Module 2: Managing Data in Aggregation				
Week 8	2-Mar	Managing data (4): Applying XML to Data and Metadata	Exercise 3: Querying data & metadata	Goldfarb, C.F. & P. Prescod (2000). Chapter 53 & 59–62.
	4-Mar	Data aggregation scenario (1): research collection	Guest speaker	Muller, R. J. (1999). Chapter 3: Gathering Requirements.
Spring Break				
Week 9	16-Mar	Data aggregation scenario (2): reference collection	Guest speaker	Karasti, H., K. S. Baker, et al. (2007). & Muller, R. J. (1999). Chapter 4 Modeling Requirements with Use Cases
	18-Mar	Data aggregation scenario (3): resource collection	Guest speaker Due: Exercise 4 XML for metadata and data records	Parsons, M. A. and R. Duerr (2005); Morris, S. P. and J. Tuttle (2007).
Week 10	23-Mar	Data and users (1): requirements Interviewing faculty about data management needs & practices analysis Data and users (2): data set characteristic analysis, needs assessment	Group report 1 / data and users	Muller, R. J. (1999). Chapter 6
	25-Mar	Data and users (3): Case Study	In-class group work	Case handouts to be delivered in class
Week 11	30-Mar	Organizational planning (1): Goals and objectives, Procedures, Quality control		Muller, R. J. (1999). Chapter 7
	1-Apr	Organizational planning (2): metadata issues, long-term preservation	Report to class findings from interviews	Gray, J., D. T. Liu, et al. (2005)
Learning module 3: Broader issues in science data management				

Week 12	6-Apr	Enabling technologies: organizing and managing data, storing and retrieving, using data	Group report 2 on interview result	
	8-Apr	Understanding Data Curation: metadata description, quality criteria, archival concepts		
Week 13	13-Apr	Data repositories and discovery: Directory services, controlled vocabularies	Quiz 3: Using data	Green, D. and T. Bossomaier (2002)
	15-Apr	Data analysis: data mining, meshing		Treinish, L. A. (1997)
Week 14	20-Apr	Data presentation: Visualization, Tools, Formatting for Publication		Sieber, J. E. (2005); Gleick, P. H. (2007)
	22-Apr	Sharing data: Ethics, Publishing, Citation		Uhlir, P. F. (2003)
Week 15	27-Apr	Project presentations and discussions, Wrap-up	Final Project Due; Post-course assessment survey	

Reading List

Book Chapters:

Abiteboul, S., P. Buneman, et al. (2000). Data on the Web. San Francisco, Morgan Kaufmann:

Chapter 2: A Syntax for Data. pp. 11–26

Chapter 3: XML. pp. 27–50

Brooks, A. A. (1984). Chapter 2: Data Types and Structures in Science and Technology. Database Management in Science and Technology: a CODATA Sourcebook on the Use of Computers in Data Activities. J. R. Rumble. New York, Elsevier Science: pp. 15–37.

Caplan, P. (2003). Chapter 14: Metadata for Geospatial and Environmental Resources. Metadata fundamentals for all librarians. Chicago, American Library Association: pp. 136–144.

Cleaveland, J. C. (1986). Chapter 1: What is a Type? An Introduction to Data Types. Reading, Mass., Addison-Wesley. pp. 1–10.

Dasu, T. & T. Johnson. (2003). Chapter 4: Data Quality. Exploratory Data Mining and Data Cleaning. Hoboken, N.J., John Wiley & Sons. pp. 99–137.

Gleason, D. (1997). Chapter 13: Fundamental Types of Data Transformation. Data Warehouse: Practical Advice from the Experts. Bischoff, J. & T. Alexander. Upper Saddle River, N.J.: Prentice Hall. pp. 160–173.

Goldfarb, C.F. & P. Prescod, 2000. The XML Handbook. Upper Saddle River, N.J.: Prentice Hall PTR.

Chapter 53: XML Basics; pp. 722–747

Chapter 59: XML Path Language; pp. 844–871

Chapter 60: Extensible Stylesheet Language(XSL); pp. 844–893

Chapter 61: XML Pointer Language (XPointer); pp. 894–900

Chapter 62: XML Linking Language (XLink); pp. 902–917

Green, D. and T. Bossomaier (2002). Chapter 10: Data Warehouses. Online GIS and Spatial Metadata. New York, Taylor & Francis: 167–187.

Hillmann, D. I. and E. L. Westbrooks (eds). (2004). Metadata in practice. Chicago, Ill., American Library Association.

Chapter 8: Hill, L. L. and G. Janée The Alexandria Digital Library Project: Metadata Development and Use. pp. 117–138.

Chapter 14: Arms, C. R. and G. Arms. Mixed Content and Mixed Metadata: Information Discovery in a Messy World. pp. 223–237

McGlamery, P. (2004). Chapter 13: Metadata and spatial data. Metadata applications and management. Lanham, Md., Scarecrow Press: 274–305.

Muller, R. J. (1999). Database Design for Smarties: Using UML for Data Modeling. San Francisco, Morgan Kaufmann:

Chapter 3: Gathering Requirements, pp. 55–74

Chapter 4: Modeling Requirements with Use Cases, pp. 75–98

Chapter 6: Building Entity-Relationship Models, pp. 105–125

Chapter 7: Building Class Models in UML, pp. 127–184

Rob, P. and C. Coronel (2004). Database Systems. Danvers, Mass., Boyd & Fraser: Chapter 1: File Systems and Databases, pp. 4–27

Chapter 3: Data Models, pp. 28–71

Chapter 3: The Relational Database Model pp. 74–121

Chapter 6: Structured query language, pp. 226–316

Scheiner, S. M. (2004). Chapter 3: Experiments, Observations, and Other Kinds of Evidence. The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations. M. L. Taper and S. R. Lele. Chicago, University of Chicago: pp. 51-71.

Zeng, M. L. and J. Qin (2008). Metadata. New York, Neal-Schuman:

Chapter 1: Introduction pp. 3–13

Chapter 3: Schemas—Structure and Semantics pp. 87–129

Chapter 4: Schemas and Syntax pp. 131–147

Published Articles:

Anderson, W. L. (2004). "Some Challenges and Issues in Managing, and Preserving Access to, Long-lived Collections of Digital Scientific and Technical Data." Data Science Journal **3**: 191–202.

Bose, R. and J. Frew (2005). "Lineage Retrieval for Scientific Data Processing: A Survey." ACM Computing Surveys **37**(1): 1–28.

Carlson, S. (2006). "Lost in a Sea of Science Data: Librarians are called in to archive huge amounts of information, but cultural and financial barriers stand in the way." The Chronicle of Higher Education. **52:** A35.

Gray, J., D. T. Liu, et al. (2005). "Scientific Data Management in the Coming Decade." SIGMOD Record 34(4): 34–41.

Guterman, L. (2001). "Learning to Swim in the Rising Tide of Scientific Data: from astronomy to zoology, researchers face an unprecedented wealth of information." The Chronicle of Higher Education. **47:** A14.

Howe, D., M. Costanzo, et al. (2008). "Big data: The future of biocuration." Nature **455**(7209): 47-50.

Karasti, H., K. S. Baker, et al. (2007). "Digital Data Practices and the Long Term Ecological Research Program." 3rd International Digital Curation Conference. Washington, D.C.**:** 13.

Morris, S. P. and J. Tuttle (2007). "Curation and Preservation of Complex Data: The North Carolina Geospatial Data Archiving Project." DigCCurr2007: An international symposium on Digital Curation, Chapel Hill, N.C., University of North Carolina, Chapel Hill.

Parsons, M. A. and R. Duerr (2005). "Designating User Communities for Scientific Data: Challenges and Solutions." Data Science Journal 4: 31–38.

Sieber, J. E. (2005). "Ethics of Sharing Scientific and Technological Data: A Heuristic for Coping with Complexity and Uncertainty." Data Science Journal **4**: 165–170.

Reports (Institution & Government):

Gleick, P. H. (2007). **The Political and Selective Use of Data: Cherry-Picking Climate Information in the White House**. Oakland, California, Pacific Institute**:** 5.

Treinish, L. A. (1997). **Scientific Data Models for Large-Scale Applications.** Yorktown Heights, N.Y., IBM**:** 12.

Uhlir, P. F. (2003). **Scientific Data for Decision-Making Toward Sustainable Development: Senegal River Basin Case Study -- Summary of a Workshop**. Washington, D.C., National Academies Press: 8–47, 62–68.

(1995). **Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources**. Washington, D.C.**:** Commission on Physical Sciences, Mathematics, and Applications, National Research Council**:**

Chapters 2: The Challenge: Preservation and Use of Scientific Data, pp. 13–32.

(2001). **Statistics Canada's Quality Assurance Framework**. Ottawa, Statistics Canada.

(2005). **Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century**. Washington, D.C., National Science Board, National Science Foundation**:**

Chapters 2: The Elements of the Digital Data Collections Universe, pp. 17–23

Chapters 3: Roles and Responsibilities of Individuals and Institutions, pp. 25–30

(2007). **Cyberinfrastructure Vision for 21st Century Discovery**. Washington, D.C., Cyberinfrastructure Council, National Science Foundation: 56.