INFOSYS 290-1, Spring 2006

After Google, What?
Information Management and the Academic Enterprise
in a Networked Digital Age

Course Info
Instructor: Daniel Greenstein, Ph.D. | daniel.greenstein (at) ucop.edu
Office Hours: Wednesdays after seminar, or by appointment at the CDL.
Contact phyllis.baker (at) ucop.edu to make an appointment.
Teaching Assistant: Alex Dailey, MIMS '07 | alex (at) sims.berkeley.edu
Class Meeting Time: Wednesdays, 1-4pm
Location: South Hall, Rm 110
URL: http://groups.sims.berkeley.edu/agw/
Blog: http://groups.sims.berkeley.edu/agw/blog/

Course Description
Universities are knowledge organizations. Amongst their essential products one counts graduates who are accredited to establish some place for themselves in social and professional hierarchies; and research that powers economic innovation, government and public policy, and cultural and educational understanding. Both of these products depend upon the university’s massive consumption and production of information and thus on sound processes that enable its creation, management, discovery, distribution and use. These processes are not only mission critical, they are essential means by which universities distinguish themselves from one another and compete effectively in a market place where good students and good faculty are keys to its reputation and revenue growth. Yet while a university’s information management processes are mission critical, they are increasingly obsolete. They are held over from an era when information circulated in analog formats and where access to it was largely determined by one’s physical location. These
same processes are increasingly dysfunctional in a world in which massive quantities of information can easily and instantaneously be “published”, discovered and transmitted.

This course evaluates trends in the information industry and how they impact upon the academic enterprise that is so heavily reliant upon outmoded forms of control over the production and flow of information. It will look in at a number of challenges in particular, including:

- the constantly evolving information needs of faculty, researchers, and students
- the changing the economics of scholarly publishing and emerging new norms of scholarly communication
- digital preservation or rather the stewardship of and entitlement to access online information that supports or results from research and learning
- role, governance, and funding of academic libraries and campus-based information services

The course will be practically oriented and informed throughout with lectures from leading practitioners in areas under consideration.
Schedule

Introduction
Jan. 18 - Course aims and overview (and a brief introduction to assessment methods)

The Academic Library
Jan. 25 - The academic library as place, steward, and information service.

Feb. 1 - Changing fundamentally, but in order to remain the same? Part 1
 * Due: Choose term paper/project option.

Feb. 8 - Changing fundamentally, but in order to remain the same? Part 2

Feb. 15 - Project Discussion Session 1

Research: Production and Distribution of Knowledge
Feb. 22 - The economics and other dysfunctions of scholarly communication systems.

Mar. 1 - Scholarly communication in transition

Mar. 8 - Primary Data as a secondary resource – data archives and their use.

Mar. 15 - Cyberinfrastructure / eScience – What is it? Why should we care?

Instruction: Personalized Paradigms in a Networked Space
Mar. 22 - The different modalities of computer-assisted instruction.

Mar. 29 - SPRING BREAK

Apr. 5 - Computer-assisted instruction at UC (guest lecture)

Apr. 12 - The qualities of open

*Open courseware, open learning, open textbooks and the future of the academy as a
gateway to credentials* *The academic enterprise as a business: IT as a business function.*

Apr. 19 - Project Discussion Session 2

Apr. 26 - Guest lecture: Digital asset management in a knowledge organization

*The University's roles and responsibilities, potential risks and rewards.*

May 3 - Strategic planning for the academic information environment

*Academic enterprise in the age of Google. IT planning.*

* Due: Term paper or project (for both longer and shorter assignments), submitted
electronically to daniel.greenstein (a t) ucop.edu by 5pm on Friday, May 5, 2006
List of Assignments

Assignments for term papers and projects are set out below under two headings: research topics, and practical exercises.

Students who wish to pursue problems not identified on the list set out below may be permitted to do so, so long as:

- the problems are related to the work of the course
- of equivalent depth, breadth, and impact, and
- are expressly approved by the instructor by the third class, February 1, 2006

Research topics

**How are academic libraries perceived and being used?** It is incontrovertible that fundamental aspects of the academic library are changing fundamentally. The trajectory of that change is less easy to discern. Thinking in an innovative way about your use of methods and sources, identify, explain and critique some of the developmental trajectories that you see emerging. Don’t hesitate to apply your analysis to UC libraries whether at Berkeley or elsewhere.

**Re-centralization of academic information services?** The paradigm for managing and funding the university’s enterprise-wide information systems seems to cycle between centralized and distributed modes. The mainframe computer environments that were so prevalent in the mid-1980s characterize the extremely centralized mode. The distributed cluster-based scientific computing environment and the departmental networks and service providers characterize the more distributed approach to information service provision that is more prevalent today. There is evidence in the past few years that the pendulum is swinging back from distributed to the more centralized or at least coordinated norms. Basing your assessment on the real experience as reported or observed at a handful of academic institutions, explain what services that are being considered in a more coordination or centralized fashion and account for this shift in their provision.

**Managing risk inherent in our digital scholarly assets.** The academic establishment is becoming increasingly concerned about the risks inherent in the volatility of digital scholarly information that has no analog in more stable formats such as print or even film. A great deal of attention has been paid to the persistent management of scholarly publications and to digital research data. Are there other digital assets that the academy should be concerned with? If so, what are they? To whom do they have some long-term value (what is the business case for
investment in their preservation)? What strategies technical, organizational, and financial, should be considered in capturing, managing, and ensuring appropriate subsequent use of such assets.

**Search in an academic context.** The assignment is intended to gather information about how academic users locate scholarly information and is designed to assist in the design of effective resource discovery systems. Students who choose this assignment will in effect be asked to participate in and then document their experience of a series of user protocols where they will either:

- locate information that support an academic inquiry in a field they are already very familiar with
- locate information that supports an academic inquiry in a field they are not already familiar

Students will document their experience of the protocols and comparing the search systems they encountered during it, with a view to identifying essential and desirable attributes of resource discovery systems as used in different contexts described above.

As part of this assignment, students will have an opportunity to assess features of either a relevance ranking or recommender service that is being developed to prototype by the California Digital Library (CDL).

**The economics of scholarly publishing.** “Postprint services will not significantly affect the economics of scholarly publishing. Worse, they will divert scarce university resources away from strategic task capable of re-shaping scholarly communication processes in general”. Discuss

**Massive digitization.** A well known company is interested in working with the UC libraries to scan a 100,000 out of copyright books and making them available online where they may accessed openly by anyone with an Internet connection. The company gives the library only some very general guidance… that the works selected for digitization should be of public and educational interest and focus broadly on Americana: literature, the arts, and history (people, places, events, and ideas) as broadly construed. What are the key approaches that the libraries should consider for selecting the 100,000 books? For each approach identify strengths and weaknesses as well as critical obstacles that would need to be overcome. Finally, make a recommendation to the libraries about how they should proceed.
Practical exercises

The following assignments are for the more technically inclined. CDL computing resources may be made available in some circumstances.

Evaluating text mining and information extraction techniques and tools. Choose from one of the following:

- Evaluate Emory's MetaCombine and Dave Newman's TopicSeek in light of other available clustering algorithms
- Evaluate the efficiency of selected open source classifiers

Resource location before discovery. Federated search integrates access to pre-selected targets. In an ideal world, targets would be dynamically selected as being those most likely to return results appropriate to a user-supplied query. In this project students will investigate methods for a recommending service that selects targets based on a user-supplied query in advance of conducting a federated search. Problems that need to be solved include determining the nature, level, and type of information that the recommendation service would need to gather from potential targets in order to make “good” and “reliable” recommendations, and of course how to gather that information and maintain it as current.

Low-cost digital preservation? As digital preservation repositories emerge on the academic information landscape, one cannot help but wonder what will happen to their contents if we find ourselves 10 or 20 years out with vast collections of archived objects, but insufficient funding to effectively render any but the most treasured ones for the then current generation of computer hardware and software. What would we do to access the rank and file objects.

Cooperative harvesting of web-based content. Web harvesting is a grossly inefficient means of capturing Internet-accessible assets. It misses deep web materials and is either too costly (where conducted with excessive manual intervention) or too inclusive (where conducted less discriminately across whole web domains). What methods might be used to enable data providers that wish to (a small set, to be sure), to “push” their content periodically into waiting repositories. What server-side standards and practices (eg, MOD, OAI, Google sitemaps) might be used?

Demographics of the .edu domain. Web harvesting techniques may provide an opportunity to capture and leverage valuable at-risk scholar information assets that reside on the .edu domain. The promise, however, requires a better understanding of the domain’s demographics. What kind of information assets are available there? Can they be detected automatically? Which are sufficiently valuable to justify the investment involved in their long-term preservation? And how may some of them be leveraged?
Recommender systems based on course syllabuses. This may relate logically to the demographics of the.edu domain (above) in a manner that makes an obvious assignment pairing. The .edu domain is littered with course syllabuses, assignment lists, and faculty and student papers and publications. Embedded in these are numerous references to scholarly publications of all kinds (journal articles, monographs, textbooks, etc), and consequently a source of data that may usefully be employed in recommender systems that guide scholarly resource discovery. How would such a recommender system be developed. How would data be gathered and maintained? How would data be weighted and/or what algorithms would be used to deploy it in a recommender system?
Assessment

Students’ work will be assessed as follows:

Term paper(s) / project(s)  70%
Class presentations       20%
Participation             10%

1. Term paper(s) / project(s) (contributing 70% of assessment)

By Wednesday, February 1st (third class) students will choose whether they wish to do:

**one** term paper or project, 25-30 pages in length (or equivalent) and submitted electronically to daniel.greenstein@ucop.edu by 5pm on Friday, May 5, 2006

OR

**two** term papers or projects, each 12-15 pages in length (or equivalent) and submitted electronically to daniel.greenstein@ucop.edu respectively by 5pm on Friday, March 10, 2006 and 5pm on Friday, May 5, 2006

Term papers and projects will be selected from the list of assignments that is supplied.

Term papers will be based upon review of a relevant and current literature. They may also include original research, for example, as may be available in some aspect of assessment (of users’ needs, information resources or services, evaluation of current events, trends, news, and credible web-based discussions and postings).

Practical work will be technically oriented and submitted with a clearly written statement explaining the approach taken to the assigned problem, why that approach was taken, as well as results achieved (strengths, weaknesses, areas for future work). The statement will be submitted with documented code, and/or any functional and technical requirements statements or specifications that have been produced in the course of the work.

For both term papers and projects, emphasis will be given in assessment to originality, creativity, clarity, and thoughtfulness.
Work may be submitted electronically in an appropriate format that may be rendered easily on the screen via readily accessible desktop software and where appropriate, on paper (e.g. as a Word file).

2. Class presentations (contributing 20% of assessment)

Each student will present twice to the seminar: once on a term paper or project they are preparing, and once on an issue of their choice that relates somehow to “After Google What?”.

Presentation on term paper or project. Two full seminars have been reserved for students to gain input from colleagues about the approaches they have chosen to their term paper or project. Twenty (20) minutes will be available to each student (depending on class-size) and students will be left to determine the balance between presentation and discussion. In their presentations students will introduce the term paper / project they have chosen and focus on the approach they are taking to address it. Students should use the opportunity to gain input from colleagues that may guide and enhance their work

Student-led discussion. Every Wednesday beginning on January 25 (second class), we will spend 20 or 30 minutes discussing an issue introduced by a student with a brief (5-10 minute) presentation.

Students may introduce any issue they choose so long as it relates somehow to an aspect of “After Google What?” . Students should feel free, in this regard, to pick up on some item “in the news”, on an issue they are engaging with in another SIMS class, on a theme they are encountering in the AGW seminar, or on a topic they are grappling with in a some independent or group-based project they are conducting.

On the Monday preceding their presentations, students should use the class listserv to describe the issue they will be introducing (a title and a sentence or two), pointing to any useful background information that might be useful for colleagues.

In the assessment of presentations, emphasis will be on creativity and clarity, as well as on facility engaging and encouraging lively, constructive, and collegial discussion and debate.

3. Participation (contributing 10% of assessment)

Ten percent of the students’ assessment will be based on participation in seminar discussions (both in class and via the blog). Emphasis will be placed on originality, thoughtfulness, preparedness, and facility encouraging and engaging constructive and collegial discussion and debate.
Syllabus

Introduction. The subjects we will be discussing are complex and not likely fully to be available in any one article or text. The reading list reflects this fact and provides a number of alternative or possible perspectives for each of the topics we will be covering. In some cases (overview of assessment methods) it is possible to identify a single most important work which it is essential that everyone become familiar with, in most others you will want to dip into a variety of works that are listed, grazing, perhaps, rather than intensively digesting.

In addition and more passively, you will want to use RSS feeds to keep up with “the news” as it becomes available. A selection of serial publications, blogs, and news sources is supplied below.

Assigned and suggested readings are organized by seminar date and topic. Required readings are indicated with an asterisk.

1/18/2006 Course overview

Practical work of the course will require some working knowledge of assessment techniques/methods and how they are used in the design, assessment, and continued evaluation of academic information services.

For an overview of assessment methods see:


As Covey points out, different methods have very different strengths and weaknesses, and are accordingly fit for different purposes. Good assessment studies often involve the use of multiple methods. To get a feel, you’ll want to familiarize yourself with how different techniques are used (well but also poorly) in different studies. Many of the readings for this course rely in some way or another on the use of different assessment techniques. You can quick-start your inquiries by sampling from some of the following:


• Any of the numerous assessment and evaluation reports conducted by the California Digital Library in the course of its designing and continually evaluating new information services, available from http://www.cdlib.org/inside/assess/evaluation_activities/index.html

1/25/2006 Historic roles

The literature on the changing roles of academic libraries is already long and continually growing.
*Graze through and get a feel for reactions to this topic in some of the following.


- *Perceptions of Libraries and Information Resources* (OCLC, 2005)
  [http://www.oclc.org/reports/pdfs/Percept_intro.pdf](http://www.oclc.org/reports/pdfs/Percept_intro.pdf)


- Leo Waaijers, “From Libraries to ‘Libratories’” *First Monday* 10:12(December, 2005),

**2/1/2006 Care and feeding of collections**

“Interesting” trends in library collection management are so new they are difficult to spot in the literature. Since UC seems to be well out in front in this area, the required readings are highly localized. See

*Relevant section of Systemwide Strategic Directions for Libraries and Scholarly Information at the University of California* (April 2004) from [http://libraries.universityofcalifornia.edu/planning/library_strategy.pdf](http://libraries.universityofcalifornia.edu/planning/library_strategy.pdf)


A less parochial more general picture can be obtained from

2/8/2006 Search and public service

* Clay Shirky, *Ontology is overrated: categories, links, and tags*  

* Tim O’Reilly, “What is Web 2.0?”  

[http://www.oclc.org/research/staff/dempsey/recombinant_library/default.htm](http://www.oclc.org/research/staff/dempsey/recombinant_library/default.htm)

Other relevant readings include:

  (Nov 2006)  


[http://www.dlib.org/dlib/april05/hammond/04hammond.html](http://www.dlib.org/dlib/april05/hammond/04hammond.html)

- S. Powers, “Cheap Eats at the Semantic Web Café”, *Burningbird* (January 27, 2005) from  

- Relevant sections of *Perceptions of Libraries and Information Resources* (OCLC, 2005)  
[http://www.oclc.org/reports/pdfs/Percept_intro.pdf](http://www.oclc.org/reports/pdfs/Percept_intro.pdf)


2/22/06 Scholarly communication in transition

The prolific literature that has grown up around this topic seems sometimes to generate more heat than light.

*Get a flavor by reading selectively from the brief articles contributed to “Web Focus. Access to the Literature” in *Nature* (March-September, 2004) from  
[http://www.nature.com/nature/focus/accessdebate/archive.html](http://www.nature.com/nature/focus/accessdebate/archive.html)

Every point of view is
represented there and it is best not to gravitate toward the one(s) with which you feel most
naturally inclined.

In addition…

- Information about how the economics of scholarly publishing impact directly at UC, and
  about UC’s various responses are available from the website maintained by The Office of
  Scholarly Communication http://osc.universityofcalifornia.edu/

- A good list of credible additional information resources is available from the OSC from
  http://osc.universityofcalifornia.edu/resources/

- Cf Ted Bergstrom’s journal pricing page from http://www.econ.ucsb.edu/~tedb/

3/1/2006 Scholarly communication in transition

* Selected readings from *“Web focus”

* Each of the five white papers and copyright resolution issued by UC’s Academic Council’s
  Special Committee on Scholarly Communication (from http://www.universityofcalifornia.edu/senate/committees/scsc/reports.html) pursues an important
  strategic or tactical direction with regard for example to faculty copyright management, peer
  review, etc

* The “UC Libraries scholarly communication program and priorities statement” (from
  http://libraries.universityofcalifornia.edu/scholarly/) gives an indication of what libraries can
  hope to achieve in response

IN addition, you might want to look at:

    http://www.arl.org/newsltr/210/principles.html

  o Joseph J. Esposito, “The devil you don’t know: The unexpected future of Open Access
    publishing”, *First Monday*, 9:8(August 2004) from

Primary data as a secondary resource – data archives and their use


More tactical approaches are available from

1. Beagrie’s and Greenstein’s, *A strategic policy framework for creating and preserving digital collections. Version 5.0* (July 2001) makes the case for preserving scholarly digital collections and provides practical and policy guidance to those in a position to do so.


3. And from the guides to developing persistent digital collections as written for different communities of scholarly data producers

4. Various publications of the AHDS [http://www.ahds.ac.uk/about/publications/index.htm](http://www.ahds.ac.uk/about/publications/index.htm)


7. Various publications of the Technical Advisory Service for Images (UK) on the management and sustainability of digitization projects (from [http://www.tasi.ac.uk/advice/managing/managing.html](http://www.tasi.ac.uk/advice/managing/managing.html))

3/15/2006 Cyberinfrastructure / e-Science

The required NSF (“Atkins”) report on cyberinfrastructure has helped to set the tone for debate nationally about the nature and level of investment in basic science (that is, in the infrastructure
to support it). Follow-on reports (listed below) extend the discussion to the social sciences, and to
the arts and humanities respectively and are worth a look.

*”The Atkins report”  

Other “cyberinfrastructure” reports


Other relevant reading on this topic


3/22/2006 Different modalities

* Saul Fisher, “Teaching And Technology: Promising Directions for Research on Online Learning and Distance Education in the Selective Institutions“ http://curry.edschool.virginia.edu/forprofit/SWP-05.htm

An excellent list of useful reports and etc on aspects of this subject is maintained by Educause at “Instructional Technologies”. See the Archived Resources in particular (from http://www.educause.edu/InstructionalTechnologies/645?Parent_ID=383)

4/2/2005 Computer-assisted instruction at UC (guest lecturer)

Readings TBD

4/12/2006 The qualities of open


4/19/2006 - Project Discussion Session 2

4/26/2006 Digital stewardship at UC. Guest lecture. Report from the coal face


5/3/2006 Strategic Planning for the HE Academic Information Environment


- Annual Educause Survey (there are five of these 2000-2004; three can still be found easily on the Net !) published annually in *Educause Quarterly*. They provide a good description of the hot issues on campuses as they plan their IT investments.
  - 6th (http://www.educause.edu/apps/er/erm05/erm0530.asp)
  - 5th (http://www.educause.edu/apps/eq/eqm04/eqm0422.asp)
  - 4th (http://www.educause.edu/ir/library/pdf/EQM0322.pdf)
  - 3rd (http://www.educause.edu/ir/library/pdf/EQM0322.pdf)


**Sources of information worth checking regularly**

Serials that may be worth watching
- DLib Magazine
- Educause Quarterly
- Educause Review
- First Monday
- Library Journal

News sources that may be worth watching
- Business Week Online (technology section) http://www.businessweek.com/technology/
- CNET http://www.news.com
- SearchEngineWatch http://blog.searchenginewatch.com/blog/

Blogs that may be worth watching
- Peter Brantley’s blog http://ono.cdlib.org/shimenawa/
- Lorcan Dempsey’s Weblog http://orweblog.oclc.org/
- O’Reilly Radar http://radar.oreilly.com/
- Peter Suber, Open Access News http://www.earlham.edu/~peters/fos/fosblog.html