

Graduate School of Library and Information Science  
Fall 2000

# Information Storage and Retrieval

(LIS 329 )

---

Section FO Fridays, 1–3:50 PM, CST Room 329, Davenport Hall

---

David Dubin

Office: LIS 222

Office hours: Thursday 2–5 PM CST

Phone: 217-244-3275 (217-BIG-EARL)

E-mail: [dubin@alexia.lis.uiuc.edu](mailto:dubin@alexia.lis.uiuc.edu)

Web: <http://www.lis.uiuc.edu/~dubin/>

*This document is Copyright © 2000 by David Dubin and the Trustees of the University of Illinois. In addition to this syllabus, this course is governed by the rules and guidelines set forth in the document A Handbook for Graduate Students and Advisers which students receive upon admission to the program. Students should also consult, and take to heart, the Professional Guidelines and Codes of Ethics for Library and Information Science Professionals available from the GSLIS main office.*

*This syllabus is provided to UIUC students as part of the materials for a particular class. However, it may be copied, redistributed, and modified under the terms of the [OpenContent License](http://www.opencontent.org) (Version 1.0). The text of that license is available on the Worldwide Web at [www.opencontent.org](http://www.opencontent.org). Resources that are linked to or referenced from within this syllabus (e.g., readings, outlines, discussions) are not covered by the OpenContent License, unless specifically labeled as such.*

## REQUIRED TEXTS

**Robert R. Korfhage.** *Information Storage and Retrieval*, first edition (John Wiley and Sons, Inc. , 1997 ).

**Various authors.** *LIS329 Reading packet* (Campus Publishing Services , 2000 ).

## SCOPE AND OBJECTIVES

This class covers systems for storage and retrieval of documents and references; their characteristics, evaluation, factors affecting their performance, and the mathematical models on which their operations are based. Primary focus is on modern computer-based systems. No prior mathematical background beyond high school algebra and trigonometry is assumed, but during the semester students will become comfortable with elementary matrix and vector arithmetic, logarithms, conditional probability, Boolean algebra, and a few basic elements of projective geometry and graph theory. This class will help prepare students for work in the area of design and development of information retrieval systems.

### *Objectives*

- Critically review research and development of information retrieval systems and services to discern predominant models which can help advance the state of the art.

- Evaluate some of the efforts to improve information access by means of different retrieval mechanisms, document and knowledge representations, intermediary and database designs, or information technologies.
- Provide an opportunity for students to select and study one aspect of the field in depth.
- Prepare for more advanced course work and projects in information retrieval.

## THIS SYLLABUS

The official syllabus for this course is the SGML version that is linked off the class web page. Expressions of the syllabus in other formats are derived from the SGML version. The current SGML version should be consulted to resolve any inconsistencies among other renditions.

## ACCESSIBILITY

To insure that disability-related concerns are properly addressed from the beginning of class, students with disabilities who require reasonable accommodations to participate are asked to contact the instructor as early as possible.

## BASIS FOR EVALUATION

Students are responsible for their performance in meeting their own educational goals and those of the course; instructors are responsible for providing guidance, expertise, and support to help students reach those goals. Students are expected to participate in class exercises and online discussions. In addition to completing all required readings, students will read additional material of their choice in order to gain a solid understanding of each course topic. Satisfactory work will receive a grade in the C range, good work will receive a grade in the B range, and superior work will receive a grade in the A range.

Final grades will be calculated as follows:

- Midterm Exam: 30%
- Research Paper: 30%
- System Case Study Presentation 15%
- Text Processing Exercise 10%
- Class Participation: 15%

### *Midterm Exam*

The midterm exam will be distributed via the class web pages, completed working alone, and submitted to the instructor via email. The Spring 1997 and 1998 exams are available on the web. Students will have two weeks to complete the midterm exam. Students may use books, articles, notes, and computers to complete the problems, but may not solicit or receive assistance from other human beings.

### *Research Paper*

The research paper is a 15 to 20 page project on a topic relevant to information storage and retrieval. The paper should present in-depth research on a topic of interest, such as those listed in the semester outline below. Term papers should demonstrate familiarity with relevant literature and should be documented with appropriate references. Use a standard style manual, such as the *Publication Manual of the American Psychological Association*, as a guide to citation. A written proposal for the research paper must be approved by the instructor no later than the 9th week of class. The proposal should include a title, one paragraph description, and citations for at least four sources. Papers are due during finals week.

### *System Case Study Presentation*

Choose a document retrieval system to which you have access. Prepare an analysis of its weaknesses and strengths, addressing the following issues:

- What is the domain and scope of the documents in the database, and what criteria have been used for their selection?
- How are documents represented within the database?
- What attributes or fields are explicitly represented?
- What kinds of access methods are available to users, and how are they shaped or constrained by the format of the documents? Conversely, you might discuss how the document representations are constrained by the access methods.
- If the system provides ranked output, what is the ranking principle?
- What similarity or relevance estimation formula is employed?
- Contact a user of the system, and ask him or her to discuss a real application of the system. How successful or unsuccessful was the use of the system?

Schedule and deliver your report as a 15 minute oral presentation to the class. Presentations will take place during the last four class meetings. Scheduling of presentations should be finalized no later than November 3.

### *Text Processing Exercise*

Assemble a collection of 200–500 short text documents in machine-readable form. Using text processing utilities demonstrated in class, investigate which words, word parts, or other units of indexing seem most promising as representatives of the documents for retrieval purposes. Prepare a 3–5 page written summary of your findings. Include whatever graphical summaries of the data are appropriate for conveying your results.

### *Class Participation*

The class participation grade is based on consistent attendance, contribution to in-class and/or online discussions, and providing assistance to classmates outside of class. Please alert the instructor if a classmate has been of help to you outside of class.

## SEMESTER OUTLINE

### **Part I: The Art of Unreasonable Demands**

*August 25*

*Databases, IR vs. DBMS*

**Readings:** Syllabus

### **Part II: Overview of Information Retrieval**

*September 1*

*Abstraction, System Roles, User Roles*

**Readings:** Korfhage, Chapter 1; Schuler et al

### **Part III: Document and Query Forms**

*September 8*

*Documents, Surrogation*

	<b>Readings:</b> Korfhage, Chapter 2; Wenger et al	
Part IV: <b>Query Structures</b>		<i>September 15</i>
	<i>Queries: Boolean, vector, probabilistic, fuzzy</i>	
	<b>Readings:</b> Korfhage, Chapter 3;	
Part V: <b>Matching methods</b>		<i>September 22</i>
	<i>Matching, Relevance estimation, Weighting, Relevance</i>	
	<b>Readings:</b> Korfhage, Chapter 4	
Part VI: <b>Text Analysis</b>		<i>September 29</i>
	<i>Lexical analysis, Term weighting, Similarity measures</i>	
	<b>Readings:</b> Korfhage, Chapter 5; Jones and Furnas	
Part VII: <b>Reference Points and User Profiles</b>		<i>October 6</i>
	<i>Profiles, Query modification, VIRIs</i>	
	<b>Readings:</b> Korfhage, Chapters 6–7	
Part VIII: <b>Review and discussion</b>		<i>October 13</i>
	<b>Take-home midterm distributed 10/13:</b> The exam is due October 27	
Part IX: <b>Text Processing Utilities</b>		<i>October 20</i>
	<b>Research Paper Proposals:</b> Due 5 PM	
Part X: <b>Retrieval Effectiveness</b>		<i>October 27</i>
	<i>Recall, Precision, Expected search length, Relevance Feedback</i>	
	<b>Readings:</b> Korfhage, Chapter 8–9	
	<b>Midterm exam:</b> Due 5 PM	
Part XI: <b>Alternative Retrieval Techniques</b>		<i>November 3</i>
	<i>Document clustering, Hypertext, Citation searching, NLP</i>	
	<b>Readings:</b> Korfhage, Chapter 10; Liddy; Bergström et al	
Part XII: <b>Presentation and Access</b>		<i>November 10</i>

**Readings:** Korfhage Chapters 11–12

**Text Processing**

**Exercise:** Due 5 PM

**Part XIII: The Ecosystem and Policy**

*November 17*

*Copyright, Privacy, Security, Standardization*

**Readings:** Korfhage, Chapter 13

**Part XIV: Thanksgiving Break**

*November 24*

**Part XV: Structured Documents and Information Interchange**

*December 1*

*Declarative markup, Retrieval from structured documents*

**Readings:** Fernandez, et al, Buneman

**Part XVI: Wrapup and Evaluation**

*December 8*

**Research Paper:** Due December 15, 5 PM

## READING ASSIGNMENTS

- [Bergström et al., 1996] Bergström, P., Haitto, H., Helander, E., Ran, A., and Ling, P.-Å. (1996). Quick guide to HyTime basics. Technical Report TR1V1, Swedish SGML User's Group, HyTime Working Group.
- [Buneman, 1997] Buneman, P. (1997). Semistructured data. In *Proceedings of the 16th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS '97)*, pages 117–121, New York. Association for Computing Machinery, ACM Press.
- [Fernandez et al., 1999] Fernandez, M., Siméon, J., and Wadler, P. (1999). XML query languages: Experiences and exemplars. Draft manuscript, communication to the XML Query W3C Working Group.
- [Jones and Furnas, 1987] Jones, W. P. and Furnas, G. W. (1987). Pictures of relevance: a geometric analysis of similarity measures. *Journal of the American Society for Information Science*, 38(6):420–443.
- [Liddy, 1998] Liddy, E. D. (1998). Natural language processing for information retrieval and knowledge discovery. In Cochrane, P. A. and Johnson, E. H., editors, *Visualizing Subject Access for 21st Century Information Resources*, pages 137–147. University of Illinois Graduate School of Library and Information Science, Champaign, IL.
- [Schuler et al., 1996] Schuler, G. D., Epstein, J. A., Ohkawa, H., and Kans, J. A. (1996). Entrez: molecular biology database and retrieval system. In Doolittle, R. F., editor, *Computer methods for macromolecular sequence analysis*, volume 266 of *Methods in Enzymology*, pages 141–162. Academic Press, San Diego.
- [Wenger et al., 2000] Wenger, M., Ochsenbein, F., Egret, D., Dubois, P., Bonnarel, F., Borde, S., Genova, F., Jasiewicz, G., Laloë, S., Lesteven, S., and Monier, R. (2000). "the SIMBAD astronomical database: The CDS reference database for astronomical objects". *Astronomy and Astrophysics Supplement*, 143:9–22.