

# Document Processing

(LIS 450 )

---

Section DP Thursday, 9–11:50 AM Room 111, Speech and Hearing Building

---

*This document is Copyright © 2001 by David Dubin, Allen Renear, 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](#) (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.*

*PostScript ® is a registered trademark of Adobe Systems Incorporated.*

## INSTRUCTORS

### **Allen Renear**

Office: LIS 213

Office Hours: Thursdays, 1–3 PM

Phone: 217–244–3297

Email: [renear@uiuc.edu](mailto:renear@uiuc.edu)

Web: <http://www.stg.brown.edu/staff/allen.html>

### **David Dubin**

Office: LIS 222

Office Hours: Wednesdays, 12–2 PM and by appointment

Phone: 217–244–3275 (217–BIG–EARL)

Email: [dubin@alexia.lis.uiuc.edu](mailto:dubin@alexia.lis.uiuc.edu)

Web: <http://www.lis.uiuc.edu/gslis/people/faculty/dubin.html>

## REQUIRED TEXTS

**Neil Bradley.** *The XML Companion* (Addison-Wesley, 2000).

**Wynter Snow.** *T<sub>E</sub>X for the Beginner* (Addison-Wesley, 1992).

**Various authors.** *LIS450-DP Reading Packet* (Campus Publishing Services, 2000).

## SCOPE AND OBJECTIVES

This class is an introduction to the technology of electronic document processing and electronic publishing. Topics include typography, page description languages, procedural and declarative markup, file formats, specifications for document style and semantics, standards for electronic document interchange.

The objective of the class is to help prepare students for careers in document design, electronic publishing, and document database administration. Students will learn to use tools at different levels of the electronic document hierarchy, from page description languages to generalized markup.

### *Objectives*

- Introduce, through demonstrations and hands-on exercises, issues of robustness, platform independence, and standardization in the representation and processing of electronic documents.
- Illuminate the representational and computational problems in document processing, and how those problems are solved in application software.
- Familiarize students with specific *de facto* and *de jure* document processing standards, including PostScript and XML.

## 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.

## 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 discussions. 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:

- PostScript programming assignment: 20%
- T<sub>E</sub>X formatting macros: 20%
- XML DTD and valid document instance: 20%
- Final Project: 30%
- Class Participation: 10%

### *On Adapting the Work of Others*

Criteria for grading assignments include (but are not limited to) creativity and the amount of original work demonstrated in the assignment. However, students are permitted to use and adapt the work of others, provided that the following guidelines are followed:

- Use of other people's material must not infringe the copyright of the original author, nor violate the terms of any licensing agreement. Know and respect the principles of fair use with respect to copyrighted material.

- Students must scrupulously attribute the original source and author of whatever material has been adapted for the assignment. Summarize (e.g. using sourcecode comments) the changes or adaptations that have been made. Make plain how much of the assignment represents original work.

### *Assignment 1: PostScript Program*

Write a PostScript application and document it with appropriate comments. The program should be a small application or simple document type, but the exact functionality of the program is up to you. The program should include examples of looping and/or branching logic, the definition of at least one subroutine, and a non-trivial graphics state transformation. The instructors will present an example of an application with complexity and scope appropriate for this assignment.

Submit the assignment via email as a set of machine-readable files. The email address to which you should send it is docproc@edfu.lis.uiuc.edu. Encode the PostScript file (or files) and the documentation as MIME attachments. PostScript file names should end with a '.ps' or '.eps' suffix. Documentation files (in plain text) should be named with a '.txt' suffix. The subject line of the message should read as follows: "LIS450-DP Assignment 1 *email-address*" (where *email-address*@uiuc.edu is your email address in the UIUC domain).

### *Assignment 2: T<sub>E</sub>X Formatting Macros.*

Write a set of T<sub>E</sub>X macros for formatting a simple document class of your choice. Wherever possible, write declarative, structural macros rather than presentation macros. Turn in the macros and T<sub>E</sub>X source for a document that illustrates them.

Submit the assignment via email as a set of machine-readable files. The email address to which you should send it is docproc@edfu.lis.uiuc.edu. Encode the T<sub>E</sub>X macros, application, and documentation as MIME attachments. T<sub>E</sub>X file names should end with a '.tex' suffix. Documentation files (in plain text) should be named with a '.txt' suffix. The subject line of the message should read as follows: "LIS450-DP Assignment 2 *email-address*" (where *email-address*@uiuc.edu is your email address in the UIUC domain).

### *Assignment 3: XML DTD and Conforming Document*

Design an XML Document Type Definition for a simple document class. The instructor will present an example of an application with complexity and scope appropriate to this assignment. Submit (via email) the DTD and a non-trivial conforming document, valid under the DTD you have designed.

Submit the assignment via email as a set of machine-readable files. The email address to which you should send it is docproc@edfu.lis.uiuc.edu. Encode the DTD, conforming document, and documentation as MIME attachments. XML document instance file names should end with a '.xml' suffix. DTD files should end with a '.dtd' suffix. Documentation files (in plain text) should be named with a '.txt' suffix. The subject line of the message should read as follows: "LIS450-DP Assignment 3 *email-address*" (where *email-address*@uiuc.edu is your email address in the UIUC domain).

### *Final Project*

All students enrolled in the class must complete a final project that has been approved by the instructor. The final project represents in-depth investigation of a topic related to electronic document processing. The project can take one of three forms:

- A software development project, with supporting documentation and a 1–2 page report/explanation. An example of such a project would be a Perl program for transforming (via an existing parser) documents conforming to the DTD of assignment 3 into the T<sub>E</sub>X format designed for assignment 2 (assuming both assignments modeled the same document class). The program must be submitted to the instructor as machine-readable source code.
- A research paper (approximately 20 pages in paper form) on an approved topic related to document processing. The paper should present in-depth research on the topic, demonstrate familiarity

with relevant literature and should be documented with appropriate references. The research paper *must* be prepared in a declarative markup language, either application-specific (e.g., a  $\text{\LaTeX}$  document class) or conforming to an SGML/XML DTD expressive enough for academic writing (e.g. DocBook). The paper must be submitted as a machine-readable ASCII file.

- A short research paper (approximately ten pages in paper form) related to a modest programming project. Format and submission guidelines as per the first two options.

Students must write a one-page proposal for their project, and approve it with an instructor no later than the eighth week of class. The proposal should include details of the format in which the project will be submitted.

### *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: **Introduction** *January 18*

*Overview of the class*

**Readings:** Syllabus

Part II: **Page Description Language / Human Language** *January 25*

*Device independence, Introduction to PostScript*

**Readings:** Spring and Dubin, Weingartner

Part III: **PostScript Programming / Writing** *February 1*

*Coordinate space transformations*

**Readings:** Spring and Dubin, Weingartner

Part IV: **Page Elements / Printing** *February 8*

**Readings:** Snow, chapters 1–5

Part V: **Document Formatting 1 / Word Processing** *February 15*

*Typography as constraint satisfaction*

**Readings:** Snow, chapters 6–9

Part VI: **Document Formatting 2 / Publishing Systems** *February 22*

*Knuth's box and glue model*

**Readings:** Snow, chapters 10, 11, 28, 29

**Assignment 1:**  
**PostScript Program:** Due at 5 PM.

**Part VII: Document Formatting 3 / The Publishing Business**

*March 1*

*Processing instructions, device independence revisited*

**Readings:** Snow, chapters 30–33

**Part VIII: An Introduction to Markup**

*March 8*

*Structure vs. presentation*

**Readings:** Coombs, Renear, and DeRose; Bradley, chapters 1–3

**Part IX: Spring Break**

*March 15*

**Part X: Character Sets and Encodings**

*March 22*

*Unicode*

**Readings:** Bradley, chapter 16; Erickson

**Part XI: XML / Document Analysis**

*March 29*

*entities, elements, content models, document type definitions*

**Readings:** Bradley, chapters 4–6

**Assignment 2: T<sub>E</sub>X  
macros and document**

**instance:** Due at 5 PM.

**Part XII: XML Processing 1**

*April 5*

*whitespace, namespaces, tree and event-based processing*

**Readings:** Bradley, chapters 7–10

**Part XIII: XML Processing 2 / Text Ontology**

*April 12*

*DOM, SAX, XSL*

**Readings:** Renear, Mylonas, and Durand; Bradley, chapters 13–15

*The Worldwide Web*

**Readings:** Bradley, chapters 17–18

**Assignment 3: DTD  
and document**

**instance:** Due at 5 PM.

Part XV: **Emerging Trends**

April 26

*schemas, XLink, XPointer, RDF and “The Semantic Web”*

**Readings:** Bradley, chapters 11–12

Part XVI: **Reading Day (no class meeting)**

May 3

Part XVII: **Wrapup and Evaluation**

May 10

**Final Projects:** Due at 5 PM.

READING ASSIGNMENTS

[Bradley, 2000] Bradley, N. (2000). *The XML Companion*. Addison-Wesley, London.

[Coombs et al., 1987] Coombs, J. H., Renear, A. H., and DeRose, S. J. (1987). Markup systems and the future of scholarly text processing. *Communications of the Association for Computing Machinery*, 30:933–947.

[Erickson, 1997] Erickson, J. C. (1997). Options for presentation of multilingual text: Use of the Unicode standard. Published on the Worldwide Web at <http://dns.hti.umich.edu/htistaff/pubs/1997/janete.01/>. Prepared for John Price-Wilkin as a Digital Information Associate Research Project.

[Renear et al., 1996] Renear, A., Mylonas, E., and Durand, D. (1996). Refining our notion of what text really is: The problem of overlapping hierarchies. In Hockey, S. and Ide, N., editors, *Research in Humanities Computing 4*, pages 263–280. Oxford University Press, Oxford.

[Snow, 1992] Snow, W. (1992). *T<sub>E</sub>X for the Beginner*. Addison-Wesley, Reading, MA.

[Spring and Dubin, 1992] Spring, M. B. and Dubin, D. S. (1992). *Hands-on PostScript*. Hayden Books, Carmel, IN.

[Weingartner, 1997] Weingartner, P. J. (1997). A first guide to PostScript. Published on the Worldwide Web at <http://www.cs.indiana.edu/docproject/programming/postscript/postscript.html>.