

Graduate School of Library and Information Science  
Spring 2000

# Document Processing

(LIS 450 )

---

Section DP   Wednesday, 8–10:50 AM   Room 143, Henry Building

---

**David Dubin**

Office: LIS 222

Office hours: Tuesdays, 1–3 PM

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.*

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

## REQUIRED TEXTS

**R. C. Turner, T. A. Douglass, and A. J. Turner.** *Readme 1st: SGML for Writers and Editors* (Prentice Hall , 1996).

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

**Spring and Dubin, Weingartner.** *PostScript Reading Packet* (Campus Publishing Services, 1998).

## RECOMMENDED TEXTS

**Bob DuCharme.** *SGML CD* (Prentice Hall, 1998).

## 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 SGML.

## 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%
- SGML 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 instructor will present an example of an application with complexity and scope appropriate for this assignment. Submit the assignment electronically via email.

### *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 electronically via email.

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

Design an SGML 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.

### *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 L<sup>A</sup>T<sub>E</sub>X document class) or conforming to an SGML 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.

### *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 19*

*Overview of the class*

**Readings:** Syllabus

**Part II: Page Description Languages 1**

*January 26*

*Device independence*

**Readings:** Spring and Dubin, Weingartner

**Part III: Page Description Languages 2**

*February 2*

*Coordinate space transformations*

**Readings:** Spring and Dubin, Weingartner

**Part IV: Document Formatting 1**

*February 9*

*Elements of the page*

**Readings:** Snow, chapters 1–5

**Part V: Document Formatting 2**

*February 16*

*Character sets and encodings, typography as constraint satisfaction*

**Readings:** Snow, chapters 6–9

**Part VI: Document Formatting 3**

*February 23*

*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 4**

*March 1*

*Processing instructions, device independence revisited*

**Readings:** Snow, chapters 30–33

**Part VIII: Style Sheets**

*March 8*

*CSS, DSSSL, XSL*

**Readings:** Stylesheet reading

**Part IX: Spring Break**

*March 15*

Part X: **SGML 1: Document Analysis**

March 22

*Generalized markup, structure vs. presentation*

**Readings:** Turner, et al, chapters 1–5

Part XI: **SGML 2: Elements and Content Models**

March 29

*content models, document type definitions*

**Readings:** Turner, et al chapter 6

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

**instance:** Due at 5 PM.

Part XII: **SGML 3: Entities and Attributes**

April 5

*reusable content, element properties*

**Readings:** Turner, et al chapters 7–8

Part XIII: **SGML Concluded**

April 12

*portability, reusability, marked sections, documentation and revision*

**Readings:** Turner, et al chapters 11–12

Part XIV: **Hypermedia 1**

April 19

*The Worldwide Web*

**Readings:** Turner, et al chapters 9 and 13

**Assignment 3: DTD  
and document**

**instance :** Due at 5 PM.

Part XV: **Hypermedia 2**

April 26

*Hytime*

**Readings:** Turner, et al chapter 14

Part XVI: **Wrapup and Evaluation**

May 3

**Final Projects:** Due May 10 at 5 PM.