UB LIS 571 Supplement

UB LIS 571 Supplement

Lecture Notes

SLecture 1.1c
Sample CVs to illustrate different information professional jobs.

Phone: 202.421.7609
Email: allison.denny@gmail.com

Allison Denny

E x p e r i e n c e

DataStream Content Solutions 9/2006-Present
Senior process engineer
- Project lead: development of an XML-based electronic publishing system for the Office of the Legisla-tive Counsel of the U.S. House of Representatives. Analyze new and legacy legislative data and define data model requirements. Map content into data models; create, maintain and document DTDs; define conversion specifications to XML and print formats. Gather and document business and technical re-quirements and support editorial workflow processes. Manage software release schedules and testing.

Discovery Communications, Inc. 3/2006-8/2006
XSLT subcontractor
- Wrote XSLT stylesheets for converting SQL data into wordprocessingML. Defined input XML schema and XML data targets, documented data transforms.

Education Resources Information Center, U.S. Department of Education 7/2004-3/2005
Lexicography subcontractor
- Facilitated the implementation of automatic indexing software and a revised workflow process.

LexisNexis 1/2001-6/2004
XML data architect
- Responsible for consistency and traceability of XML data elements through the development of a data-driven publication management system. Built and maintained an environment of related XML data models, including DTDs, UML diagrams, XSLT stylesheets, templates and technical specifications. Acted as project lead for data-intensive software release cycles. Supported users defining data requirements and business processes. Wrote documentation and training materials.

National Public Radio 1/1998-8/1999
Broadcast librarian
- Catalogued and indexed daily NPR programming. Provided reference service using NPR database.

Education

Georgetown University 9/2003-12/2007

Communication, Culture, and Technology

- Master of Arts. Interdisciplinary program exploring media and technology from social, economic, political, and cultural perspectives. Seminars included Knowledge Management, Computer-Mediated Communication, Technologies of the Text.

University of Maryland, College Park 9/1999-12/2000

College of Information Studies

- Master of Library Science. Concentration in Information Organization with coursework including Information Structure, Construction of Thesauri, Abstracting and Indexing, Database Design.

Northwestern University 9/1991-6/1995

- Bachelor of Arts, honors. Major in American Studies, minor in French.

Technology

- XML-related technologies: XML, XPath, XSLT, DTD, XSD, RELAX NG, schematron, CSS, UML; familiar with RDF/OWL, topic maps, DITA.
- XML-related tools: XMetaL, XML Spy.

Tamar Donovan

6058 SUNNY SPRING • COLUMBIA MD 21044
PHONE 410-772-9049 • E-MAIL KORODON@AOL.COM

WORK EXPERIENCE

05/2005 –
Independent contractor /consultant

- Currently: Metadata Librarian/Information Architecture Consultant, The Electronic Scriptorium Ltd.

Some projects I have participated in:

- Organizing a taxonomy of job titles for a job-search website;
- Designing metadata structure for the digital repository of a symphony orchestra;
- Designing a database to support construction of a names memorial;
- Designing metadata structure for the digital repository of an academic journal.

08/2004 – 05/2006 Gibson Library, J.H.U. Applied Physics Laboratory. Laurel, Maryland
Circulation Clerk (part-time)

- Circulation; technical services; reference and research assistance; special projects.

American Embassy, Tashkent. Tashkent, Uzbekistan. *Consular Associate*
2001 - 2002

- Conducted non-immigrant visa interviews; adjudicated visas; assisted in anti-fraud investigations.

American Consulate, Vladivostok Vladivostok, Russia
1998 - 2000
Consular Associate

- Conducted non-immigrant visa interviews; adjudicated visas; assisted in anti-fraud investigations.

Ankara Community Library. Ankara, Turkey. *Volunteer*
1996-1998

- Performed circulation and shelving duties.

Indiana University Library Systems Bloomington, Indiana
1987 - 1988
Assistant to Modern Languages Librarian

EDUCATION

2003 - 2005 University of Maryland, College of Library and Information Science. *M.L.S., May, 2005.* Track: Information Access and Use.

Indiana University, Department of Central Eurasian Studies. Major fields of study: Hungarian language and
1986 – 1988
literature, comparative studies of Uralic and Altaic language and peoples.

The Johns Hopkins University, B.A., Classics.
1984 - 1986
Inducted into Phi Beta Kappa, May, 1986. Graduated with Departmental and university honors.

LANGUAGES

Fluent speaker of Russian. Reading knowledge of several European languages.

SLecture 1.1c Salaries of reporting professionals* by area of job assignment
Library Journal Oct. 2012, 2011 numbers. Full-time placements

ASSIGNMENT	No.	% of Total	Low Salary	High Salary	Average Salary	Median Salary
Access Services	19	1.10%	34K	53K	42K	40K
Acquisitions	15	0.8%	19K	46K	32K	32K
Administration	77	4.3%	26K	150K	49K	41K
Adult Services	38	2.1%	12K	50K	38K	39K
Archives	107	6.0%	21K	57K	39K	38K
Automation/Systems	12	0.7%	28K	90K	55K	51K
Cataloging & Classification	74	**4.1%**	19K	70K	39K	40K
Children's Services	74	4.1%	18K	53K	37K	38K
Circulation	86	4.8%	17K	60K	35K	36K
Collection Development	17	0.9%	18K	80K	44K	40K
Database Management	22	1.2%	28K	84K	49K	41K
Electronic or Digital Services	65	3.6%	13K	88K	44K	45K
Government Documents	4	0.2%	23K	55K	42K	45K
Information Architecture	11	0.6%	41K	84K	63K	65K
Info Technology	66	3.7%	21K	110K	56K	52K
Instruction	63	3.5%	25K	80K	45K	44K
Interlibrary Loans/ Doc. Del.	20	1.1%	22K	48K	36K	39K
Knowledge Management	13	0.7%	30K	68K	50K	48K
Metadata	7	0.4%	38K	59K	49K	52K
Other	299	16.6%	12K	180K	46K	42K
Public Services	35	1.9%	21K	61K	41K	42K
Records Management	17	0.9%	34K	70K	46K	43K
Reference/Info Services	265	**14.7%**	15K	129K	46K	44K
Research	13	0.7%	30K	75K	47K	43K
School Library Media Spec.	178	**9.9%**	12K	71K	45K	44K
Solo Librarian	57	3.2%	15K	70K	40K	40K
Technical Services	36	2.0%	16K	60K	40K	41K
Usability/Usability Testing	46	2.6%	30K	100K	57K	70K
Web Design/Development	6	0.3%	38K	95K	69K	70K
Youth Services	553	**3.1%**	10K	73K	38K	39K
TOTAL	**1797**	**100.00**	**10K**	**180K**	**45K**	**42K**

Library Jobs by Level, ALA survey 2008. Average salary
2008 *ALA-APA Salary Survey: Librarian – Public and Academic* (*Librarian Salary Survey*)

Job title	Public	Academic
Director/Dean/Chief Officer	100K	98K
Deputy/Associative/Assistant Director	78K	82K
Dept Head/Branch Mgr/Coordinator/Senior Mgr	66K	65K
Manager/Supervisor of Support Staff	55K	57K
Librarian Who Does Not Supervise	53K	56K
Beginning Librarian	49K	47K

http://ala-apa.org/newsletter/2010/11/01/salary-survey-librarian-pay-increased-3-percent-despite-2010-economic-woes/ (Tables 4 and 5)

Some jobs in other environments (original numbers from www.payscale.com, swz.salary.com, cbsalary.com, and 2003 compilation by Roberta Shaffer and amended using the Inflation Calculator from http://www.bls.gov/data/inflation_calculator.htm)

Job title	From	To	Source
Chief Knowledge Officer	82K	163K	payscale
Chief Information Officer	113K	191K	payscale
Information Technology (IT) Manager	60K	115K	payscale
Chief Information Security Officer	159K	230K	salary
Information Architect	49K	129K	payscale
Ontologist	78K	105K	payscale
Senior content specialist	66K	93K	salary
Information Analyst	58K	158K	cbsalary
Consumer Information Director	69K	148K	cbsalary
Archivist	48K	65K	payscale
Strategic Information Planner	71K	94K	RS 2003
Business Intelligence Manager	69K	113K	RS 2003
Manager, Campus Technology and Academic Computing	78K	169K	RS 2003
Legal Information Specialist	63K	100K	RS 2003
Sarbanes-Oxley Compliance Manager, IT	111K	121K	payscale

* Note that numbers from payscale.com are the **Median** Salary by Years Experience charts

Supplement for Lecture 1.2

SLecture 1.2.1 Example 3. Medline, a bibliographic information system

Medline, from the National Library of Medicine, is the premier bibliographic system in medicine

Purpose	Find documents on a given subject Answer the question: what Documents X <*dealsWith*> a given Subject?

System for simple subject search

Two types of facts
A title facts **B indexing facts** (index terms)

Question	What documents deal with Hearing tests? Document X <*dealsWith*> Hearing tests
Facts	A1 Document 1 <*hasTitle*> Measurement of acoustic impedance in the ear canal B1 Document 1 <*dealsWith*> Acoustic impedance tests B2 Document 1 <*dealsWith*> Computer simulation B3 Document 1 <*dealsWith*> Hearing--physiology A2 Document 2 <*hasTitle*> Optimization of automated hearing test algorithms B4 Document 2 <*dealsWith*> Algorithms B5 Document 2 <*dealsWith*> Auditory threshold B6 Document 2 <*dealsWith*> Computer simulation B7 Document 2 <*dealsWith*> Hearing tests A3 Document 3 <*hasTitle*> Expert systems for medical diagnosis B8 Document 3 <*dealsWith*> Diagnosis B9 Document 3 <*dealsWith*> Expert systems B10 Document 3 <*dealsWith*> Neural networks (computer) A4 Document 4 <*hasTitle*> New standard enhances efforts in hearing conservation. B11 Document 4 <*dealsWith*> Audiometry B12 Document 4 <*dealsWith*> Data interpretation, statistical B13 Document 4 <*dealsWith*> Ear protective devices--standards B14 Document 4 <*dealsWith*> Hearing loss, noise-induced--prevention and control
Rules	None
Answer	Document 2 (due to fact B7 document 2 <*dealsWith*> Hearing tests)

But things are not so simple. There are actually more documents on the topic, but they deal with specific hearing tests rather than hearing tests in general. To find these documents the system needs additional knowledge, an additional type of facts, namely hierarchical relationships between concepts. Such facts are available in the Medical Subject Headings published by National Library of Medicine. The database Medline uses the inclusive searching method discussed below.

System for more complete subject search exploiting knowledge of hierarchy (fact type C)

Three types of facts
A title facts
B indexing facts (index terms)
C concept hierarchy facts

Question	What documents deal with Hearing tests? Document X *<dealsInclusivelyWith>* Hearing tests
Facts	C1 Hearing tests *<hasNarrowerTerm>* Audiometry C2 Hearing tests *<hasNarrowerTerm>* Acoustic impedance tests
Rules	Document X *<dealsInclusivelyWith>* Subject Y IF Document X *<dealsWith>* Subject Y Document X *<dealsInclusivelyWith>* Subject Y IF Subject Y *<hasNarrowerTerm>* Subject Z AND Document X *<dealsWith>* Subject Z
Answer	Document 1 (due to fact B1 document 1 *<dealsWith>* Acoustic impedance tests) Document 2 (due to fact B7 document 2 *<dealsWith>* Hearing tests) Document 4 (due to fact B11 document 4 *<dealsWith>* Audiometry)

The human reader can assimilate hierarchy facts better in a linear arrangements as shown below:

Hierarchy excerpt from Medical Subject Headings

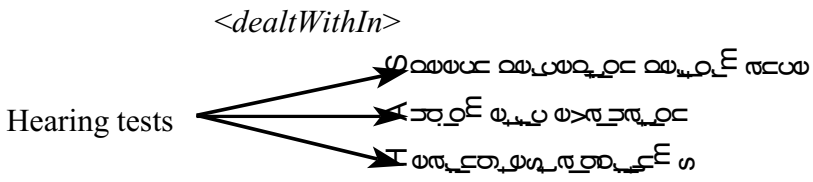
E1	**Diagnosis**
E1.276	. Diagnosis, otorhinolaryngologic
E1.276.299	. . Diagnosis, ear
E1.276.299.375	. . . **Hearing tests**
E1.276.299.375.100	. . . . **Acoustic impedance tests**
E1.276.299.375.297	. . . . **Audiometry**
E1.276.299.375.297.45	. . . . . Audiometry, evoked response
E1.276.299.375.297.92	. . . . . Audiometry, pure-tone
E1.276.299.375.297.105	. . . . . Audiometry, speech
E1.276.299.375.297.105.89	. . . . . . Speech discrimination tests
E1.276.299.375.297.105.902	. . . . . . Speech reception threshold test
E1.276.299.375.330	. . . . Dichotic listening tests
E1.276.299.375.570	. . . . Recruitment detection (audiology)
E1.276.299.816	. . . Vestibular function tests
E1.276.299.816.250	. . . . Caloric tests
E1.276.299.816.435	. . . . Electronystagmography
E1.276.591	. . Laryngoscopy
E1.276.660	. . Nasal provocation tests

Note: The term numbers (also called codes or notations) make the connection between an alphabetical index and the hierarchy listing.

SLecture 1.2.2 Types of information systems from simple to complex (and more useful)

simple: System just finds data, user does all the processing needed to find an answer, often adding knowledge from other sources

to complex: System does a lot of processing for the user, provides final answers / problem solutions

<table>
<tr><td>Information systems by extent of processing</td><td colspan="2">Information systems differ in the extensiveness of their knowledge base (or database) and the intensity of information processing to find or create an answer. The more extensive the knowledge base and the more intensive information processing, the more useful are the answers the system can give and the easier is interaction with the system.
Plain retrieval systems vs.
knowledge-based systems, intelligent information systems, expert systems
The term decision support system is also used, particularly in connection with systems that use simulation and modeling to support business decisions.</td></tr>
<tr><td>Types of information processing</td><td>• inferential reasoning
• mathematical computations
• statistical analysis</td><td>• simulation and modeling
• neural networks
• genetic algorithms</td></tr>
<tr><td>Plain information retrieval or database system</td><td colspan="2">A plain information retrieval or database system finds answers from statements that exist ready-made in the database. Another way of saying this: A plain IR system uses one-step linkages.
Example: bibliographic IR system
Question: Find documents dealing with Hearing tests
Query: Document X <dealsWith> Hearing tests
Answers
Speech perception performance <dealsWith> Hearing tests
<dealsWith> Hearing tests
<dealsWith> Hearing tests
 

Could also use links from words in the text or from person who is author.</td></tr>
</table>

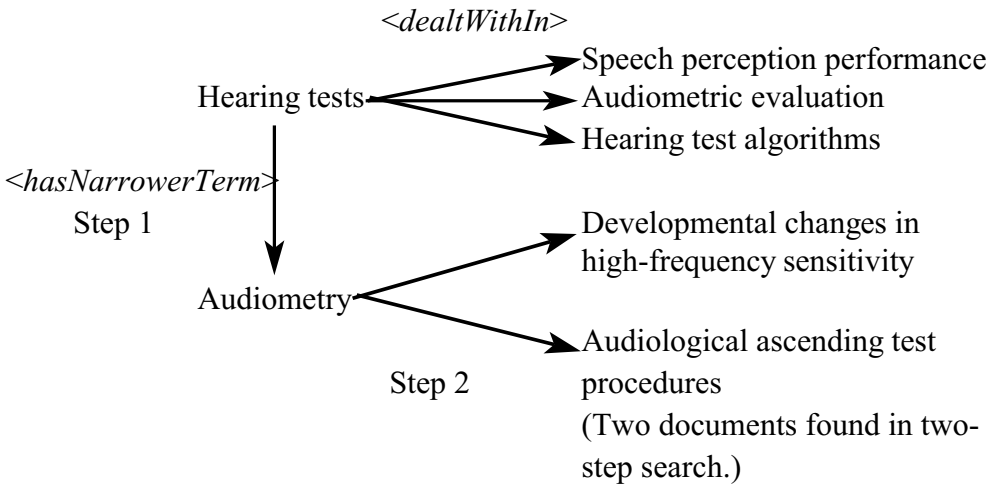
Expert system

An **expert system** uses a chain of inferences relying on many types of data concerning many types of objects/entities, for example:

- Prescription of drugs is based on data about the illness to be treated, the effectiveness of drugs against certain illnesses, contra-indications of drugs, and other conditions of the patient.

- Expert system for college choice. Such a system starts by simply comparing the criteria entered by the user with the corresponding data about the colleges – simple retrieval. But such a system would also consider user characteristics (such as grades and test scores) and compare them with the admissions standards of the college – qualified by subject applied for and other relevant factors – and thus arrive at a probability of admission. Or it would use data about alumni who are relatives of the user - if these data are available.

- Inclusive (explode) searching in MEDLINE uses data on the hierarchical relationships between descriptors in addition to the data about document-descriptor linkages. So it does combine two types of data to arrive at retrieval results and could therefore be called an expert system. But inclusive searching is a borderline case, and MEDLINE is not commonly seen as an expert system (even though it mimics an expert librarian).



- There is no sharp boundary between ordinary information systems and expert systems (also called knowledge-based systems). The more different types of facts are in the system and the more inference (combination of different types of facts) is used in deriving answers, the more expert the system is. Medline would not normally be considered an expert system, but it is capable of inclusive searching, thus it uses knowledge about concept relationship just as a knowledgeable reference

Characteristics of a good information system

- Adapts to the special needs of the user and the specific situation.
- Interprets requests (including understanding natural language) and asks user for clarification when needed. Engages users in a dialog to clarify requests.
- Processes raw data and gives answers that are directed toward the solution of the user's problem or a solution itself, saving the user the considerable effort required for assimilating and processing raw data.
- Asks for more information if it is needed to derive a good answer.
- Gives answers in easily-understood format.
- Gives reasons for suggested problem solutions, explains its reasoning.
- Assists in knowledge acquisition, for example by extracting facts from text.
- Learns.

Advanced ideas to ponder

Interrelatedness of knowledge
- Inference relationships
- Contradictory knowledge

More input/output

Understanding graphical representation, receiving instrument-generated data
Generating language and graphics.

Expert system examples (under construction)

Expert systems can give us ways to build solutions to real problems. Examples of things that an expert system might do:

- Diagnosis and advice (medical diagnosis and advice, automotive diagnosis and advice, skin care and cosmetics, color combinations, ...).
  http://easydiagnosis.com/
  OSHA eTools and: www.osha.gov/dts/osta/oshasoft/index.html

- Troubleshooting techniques for machinery (cars, phones, household appliances etc), a variation on 1.

- Identifying plants, fish, insects etc.

- Selecting foods for particular occasions.

- Support for making a decision or choice, for example choosing a music CD based on ones you enjoy or hate.

- Working out the best way to do some task (for example, what is the best way to get from Kings Meadows to Invermay on a Friday night?)

- Making a decision on a mortgage product (consumer) or on approving a mortgage (bank) www.bankrate.com/brm/mortgage-advisers/home.asp

- Making a decision on what school to apply to (student) or what students to admit (university/college)
  http://ieeexplore.ieee.org/iel5/8934/28293/01265222.pdf?arnumber=1265222
  Related: Choice of major
  http://findarticles.com/p/articles/mi_m0FCR/is_4_36/ai_96619963

- Configuring a computer or other machinery

- Applying cataloging rules

(From www.education.tas.gov.au/itproject/topics/expertsystems/expertsystems.htm)

For more information: www.aaai.org/AITopics/html/expert.html
www.generation5.org/content/2005/Expert_System.asp

Supplement for Lecture 2.1 More on categories

SLecture 2.1 2.2 Objectivist vs. organism-centered view of categories.

A balanced view. Elaboration

Scientists believe that visual information is processed along two pathways in the brain, one which specializes in spatial information and coordinating vision with action, and a second which identifies objects (Anderson, 2005, p. 41). The first pathway ties directly into the emotional center of the brain which controls the flight or fight response. This pathway is much faster than the second, such that all visual information is filtered through an affective response before being evaluated cognitively (Barry, 2005, p. 45-62). The brain attempts to match what has been seen with previous templates, patterns, or features in order to identify what was viewed and assign an appropriate response (Anderson, 2005, p. 48-58)/|Because of this, human visual perception does. not duplicate reality.

> ... what our eye register is not a picture of reality as it is. Rather our brains combine information from our eyes with data from our other sense, synthesize it, and draw on our past experience to give us a workable image of our world. This image orients us, allows us to comprehend our situation, and helps us to recognize significant factors within it....The visual world, then, is an interpretation of reality but not reality itself. It is an image created in the brain, formed by an integration of immediate multi-sensory information, prior experience, and cultural learning (Barry, 1997, p. 15).

Because the human brain processes visual information based on a variety of unique factors, meaning often varies among individuals and even over time for the same individual.

Barry, M. (1997). Image Intelligence: Perception, Image, and Manipulation in Image Communication. New York: State University of New York Press.

From Rachael Bradley dissertation, p. 26

SLecture 2.1 2.4 Basic level categories (Eleanor Rosch)

Further Quotes here are from
Rosch, Eleanor.
Classification of real-world objects: Origins and representation of cognition.
Johnson-Laird and Wason, eds. *Thinking*. 1977)

Importance - some applications in information systems:

Classification for children's collections
Easiest level of specificity in indexing
Book at right level of specificity for reader
Medical concepts known to health consumers (?)

"In so far as categorization occurs to reduce the infinite differences between stimuli to behaviorally and cognitively usable proportions, two opposing principles of categorization are operative:

(a) On the one hand, it is to the organism's advantage to have each classification as rich in information as possible. This means having as many properties as possible predictable from knowing any one property (which, for humans, includes the category name), a principle which would lead to formation of large numbers of categories with the finest possible discriminations between categories.

(b) On the other hand, for the sake of reducing cognitive load, it is to the organism's advantage to have as few classifications as possible, a principle which would lead to the smallest number of and most abstract categories possible.

We believe that the basic level of classification, the primary level at which 'cuts' are made in the environment, is a compromise between these two levels; it is the most general and inclusive level at which categories are still able to delineate real-world correlational structures." (p. 213)

How basic level categories apply in information architecture (slides):

http://www.kapsgroup.com/presentations/Semantic%20Technology%20-%20Basic%20Categories.ppt

Supplement for Lecture 2.2. Approaches to knowledge representation

0 Forming categories in a set of entities to create a more efficient data structure using hierarchical inheritance.

Introductory exercise with three examples.

Example 3, Food products, semantic network representation.

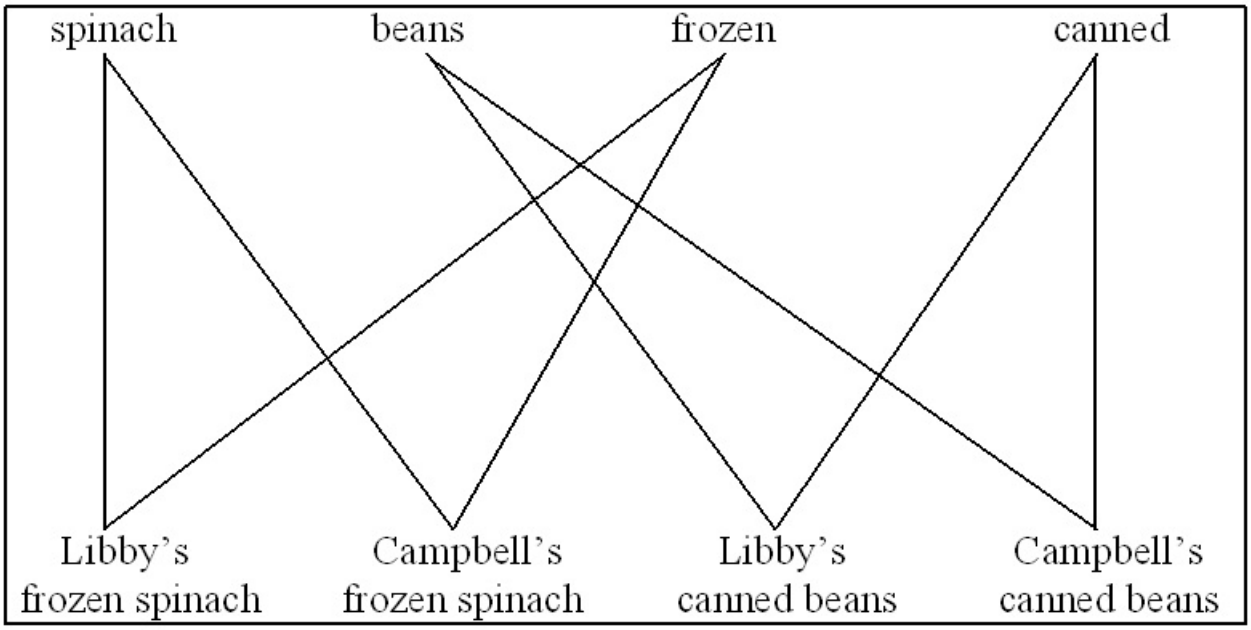
Hierarchical inheritance was first formally described in the context of semantic networks, and semantic network data were presented as graphs (as in the optional Lindsay reading). However, it turns out that the graphical representation is not as suitable for discovering information in common to several items as is the representation as database records or, for a small number of items, the feature comparison table modeled after Consumer Reports product comparisons. For those interested in looking at the graphical representation, it is given here for Example 3 and for Assignment 4.

In the graphical representation, the *common record* from the database representation is represented as a *new node* that captures the information in common to all items in a group.

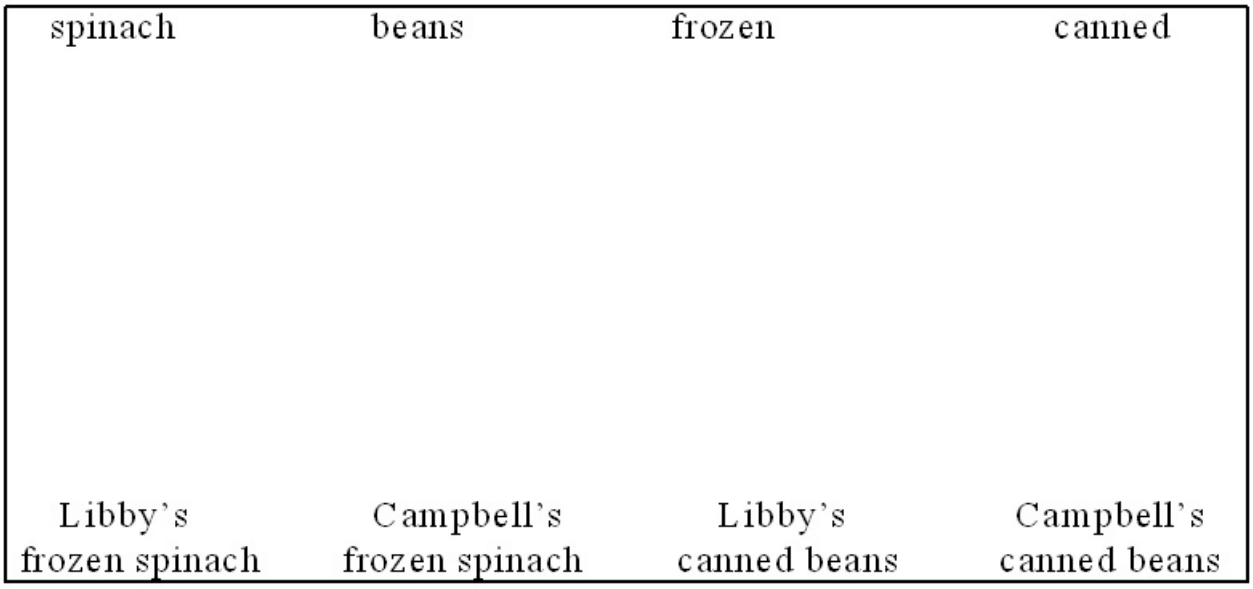
Example 3a. Food products. Semantic network representation

Exactly the same data about food products are now represented graphically as a semantic network (Lindsey reading). Corresponding to the common record, create for each group a *common node* that captures what is in common to the food products in the group. Place the common node on an invisible horizontal line through the middle of the box. Link the common node to the applicable nodes on the top and the specific food products in the group on the bottom.

Original semantic network

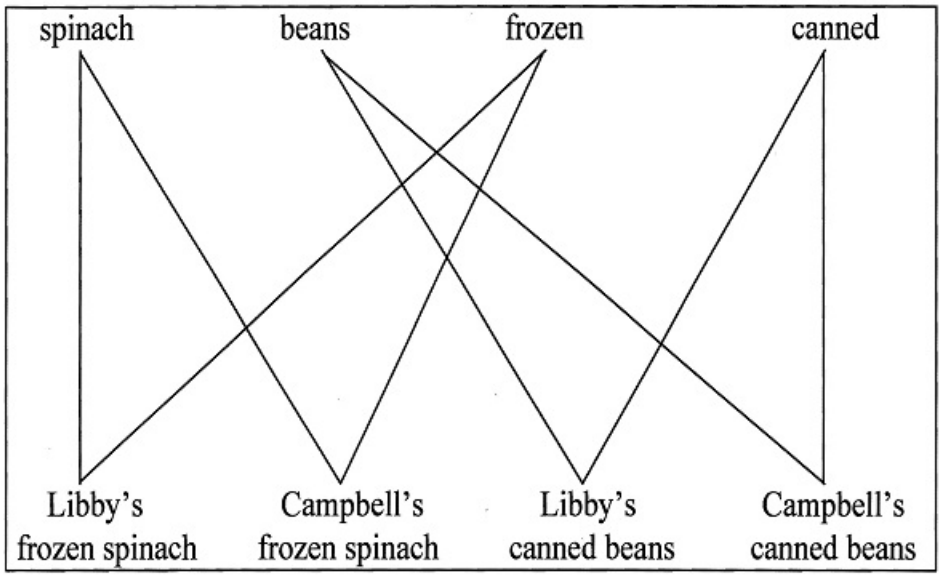


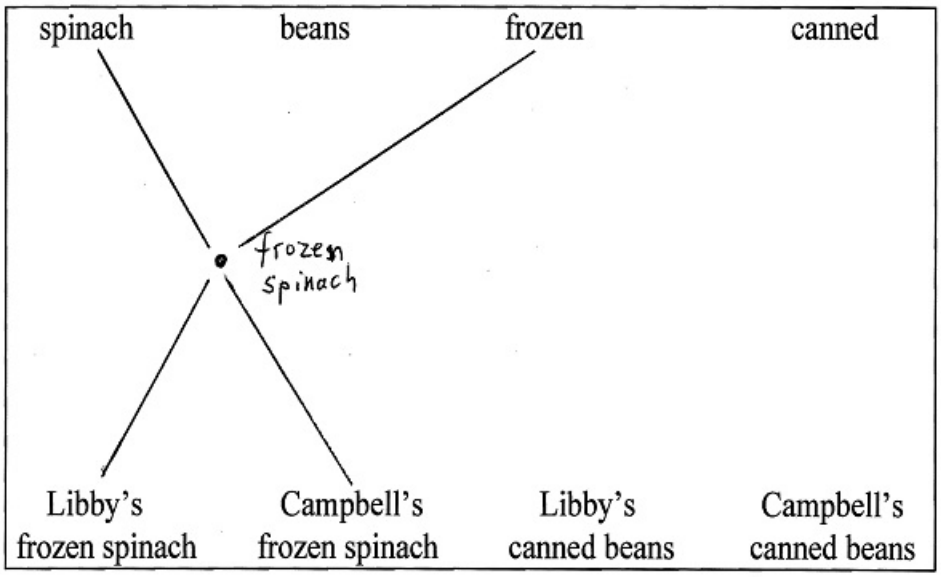
Restructured semantic network



Example 3a. Food products - Semantic Network Answer partial
See whether you can now complete the answer for the second group

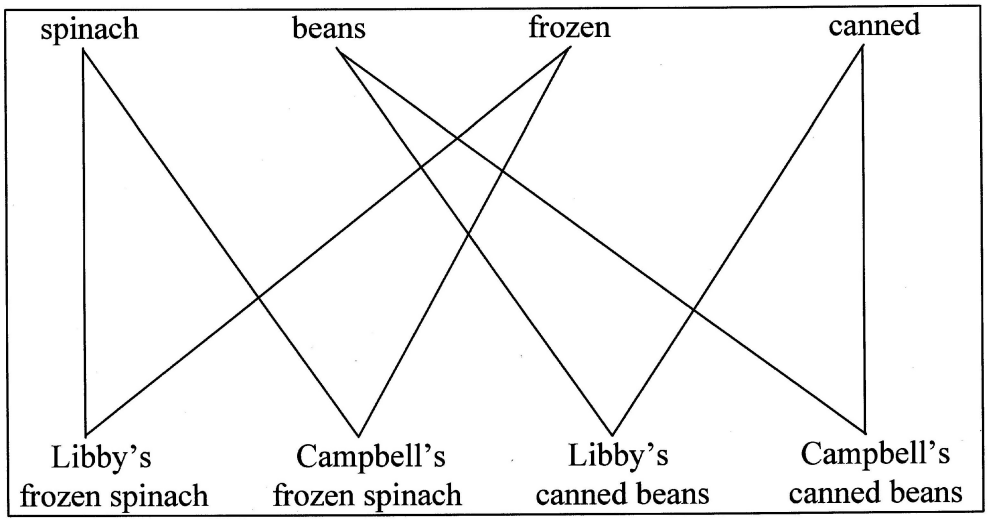
Look at the semantic network below. How can it be restructured for more efficient storage? Complete the second copy of the network to show your restructuring.

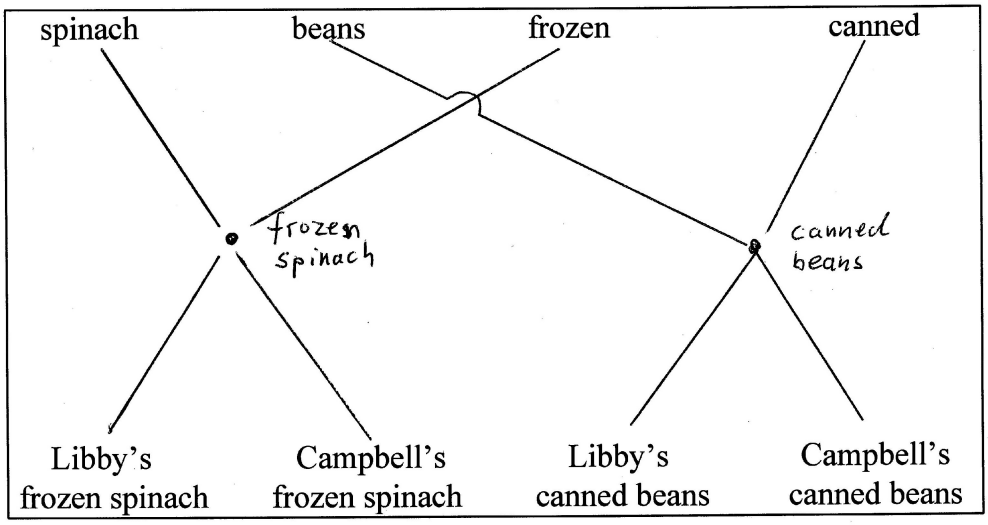




Example 3a. Food products - Semantic Network Answer

Look at the semantic network below. How can it be restructured for more efficient storage? Complete the second copy of the network to show your restructuring.





- captures what is in common to a group of food products
  if the database includes 50 frozen spinach products this saves many connecting lines

SLecture 2.2 Describing and Evaluating Knowledge Representation

Advanced objective

3 Solidify the understanding of the approaches to knowledge representation as a basis for evaluating knowledge representation schemes

4 Some criteria for describing and evaluating knowledge representations

Completeness, expressiveness, detail (subdivided by type of knowledge)

Extensibility - can easily add new types of knowledge

Parsimony of syntax and of vocabulary - use small number of syntactic constructs and of entity and relationship types

Modularity

In a modular system, small pieces of knowledge can be added to the knowledge base without changing what is already there

Compactness / redundancy

In a compact system, knowledge that can be inferred or derived is not stored but produced on the fly as needed, which may take time. In a redundant system, inferable knowledge is stored explicitly; this may save time but does take up space. An additional problem is that when knowledge changes stored inferred knowledge may no longer be true; the system has to watch out for that (truth maintenance).

Ease of processing by people or by computer programs

Ease of producing a knowledge base

Ease of writing knowledge items

Support for knowledge elicitation, support for association

Consistency checks

Plausibility checks

Ease of retrieval

Ease of reading

Ease of reasoning, drawing inferences by deduction and induction

These criteria can be applied

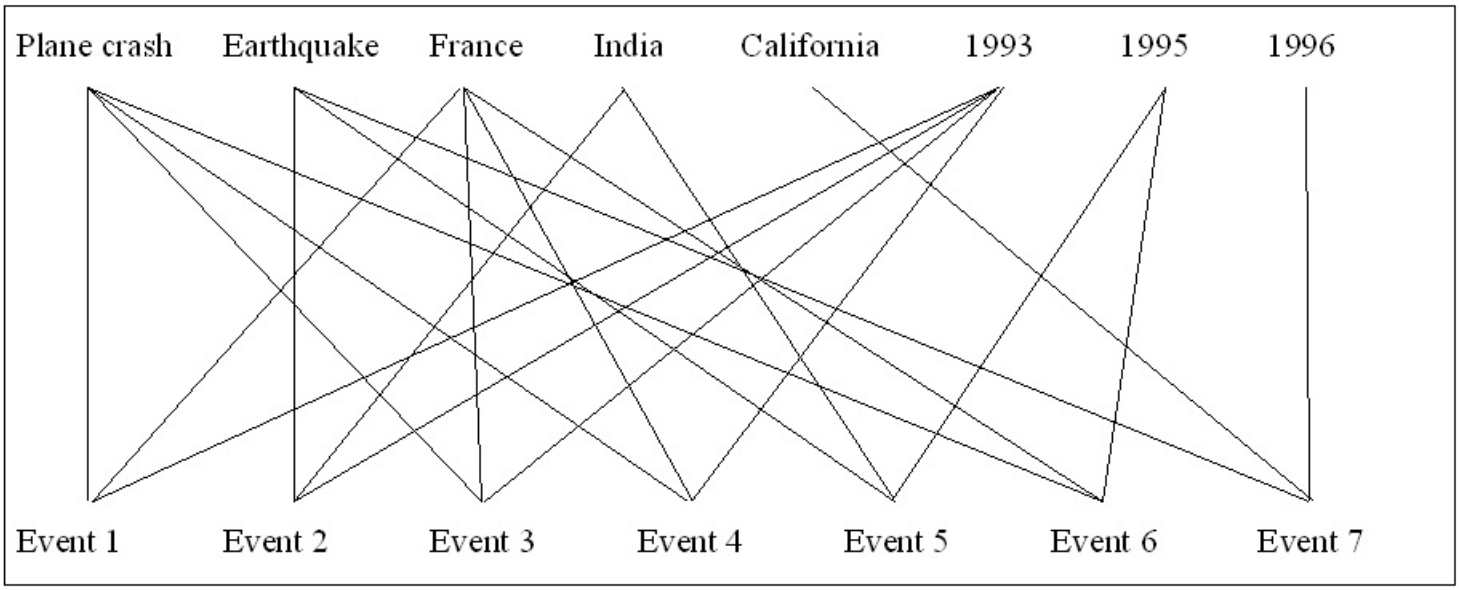
- to the syntax (the format of knowledge representation);

- to the conceptual data schema (entity types and relationship types);
- to the vocabulary (entity values).

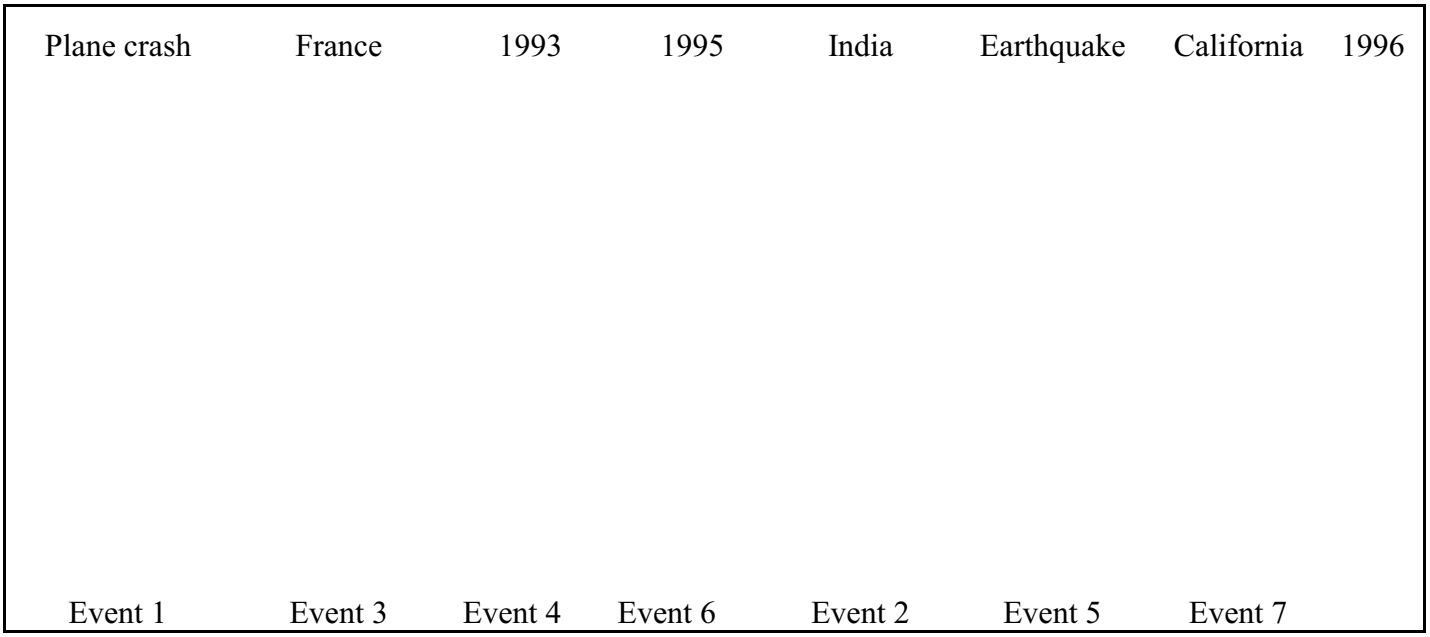
Distinguish between domain-independent vocabulary and domain-dependent vocabulary. For example, in medicine such terms as **asthma** and **prednisone** are domain-dependent (domain-specific) while such terms as **cost-benefit analysis** and **triage** are domain-independent (general)

The material at the end of Lecture 1.2 is related.

Original semantic network



Restructured semantic network



Note 1: The original semantic network is drawn from the original database by drawing from each event a line to its *Type*, *Place*, and *Time*. So the database shown on the next page and the semantic network represent exactly the same data.

Note 2: In the outline for the restructured semantic network the events have been reordered to make drawing the restructured network easier.

SLecture 3.1 Information system description examples
Also useful for Assignment 5 Analytical description of an information system

Darling High School Library
by Ruth Milin

1. Purpose Profile of DHS (included here for illustration, not required in the assignment)

Some priority is set for (1a), provide references relevant to search requests (or tools to get such references), whether or not the corresponding documents are actually stored within the system. The emphasis in purpose at DHS is (1b), physically store books and other documents and make them available. Purpose (2), providing tailor-made packages of substantive data, is given minimum priority at DHS, except to a small degree in the area of extracting and providing immediate data. There is practically no priority given to analysis of documents or preparation of state-of-the-art reports.

2.A.

Process	Organizational unit
Identification of user needs, general	central system library office and head librarian
Identification of user needs, specific	head librarian, assistant librarian
Acquisition of documents	selection-head librarian, assistant librarian with suggestions from teachers, students, and administration buying-system library office - acquisitions dept.
Formulating search requests in terms of descriptors	system library office-cataloging dept.
Indexing	system library office-cataloging dept.
Feeding into storage	system library office-cataloging dept. at school - clerk, student assistants with head or assistant librarian checking the work.
Comparison, match	system office - cataloging office school - head librarian, assistant librarian
Check whether satisfactory	system office - cataloging office school - head librarian, assistant librarian
Make material available to user	clerk, student assistants
Public relations librarian	head librarian, assistant

FILE	LOCATION
People having problems and info arising from there	school-wide-students, teachers, administrators
Documents, data, knowledge "floating around"	"floating around"
Information on user needs in general	curriculum lists and school interests in general cabinet file in librarian's work room
Interest profile/search request, as acquired	Current Interest File in librarian's Current Interest File in librarian's office
Documents, data, as required	books kept at central acquisitions office until cataloged. Sent to school where kept in storage room at rear of library until cards etc. pasted in books and ready to be put out on the shelves.
Store 1 - Interest profile/request formulations in terms of descriptors	Current Interest File in librarian's office
Store 2 - Document representations in terms of descriptors	file in cataloging dept. of center office. Entered in Sears Catalog in library workroom. Index file - card catalog file at center of school library. File of documents - books on shelves in library at school second floor, rear of school.
Documents, data selected (retrieved)	school library
Material processed, according to user needs	school library card catalog, Sears Cat.
Output: potential users know about services available	announcements over p.a., bulletin boards and display cases around school publicity records kept in file cabinet in workroom of school library.
Output: User has information available	school, school library

Note: Information about the conceptual data schema / system rules is missing from this example.

The structure of information systems in our public schools
By Karen Levitan

Introduction

Observations were made at two Montgomery county school media centers, East Silver Spring Elementary School and Montgomery School and Montgomery Blair Senior High School, both in Silver Spring, Maryland. These two schools were chosen for the following reasons: 1) they provided a basis for a gross comparison between information systems for young children and those for adolescents; 2) they represented "average" media centers in terms of available funds, age of school, background of community; 3) they were conveniently located for the researcher. The objective behind this gross comparison was to ascertain a general description of information systems in our public schools, point out similarities and differences between elementary and senior high media centers, and identify areas where research and further study and development are needed.

Purpose profiles

Again, this is included here only for illustration of the functions (Text, Section 2.6). This section is no longer required in the assignment.

Montgomery Blair Sr. High Media Center (2100 students, 118 teachers, 3 librarians, 3 aides). Montgomery Blair's primary purpose is to provide access to materials. At least 90% of the center's activities deals with storage and availability of print and nonprint media. Approximately 10% of its work is to provide references relevant to search requests. No attempt is made to analyze data, to extract data for bibliographic or state-of-the-art reports or to repackage data in any way.

Each Silver Spring[1] Elementary School Media Center (500 students, 18 teachers, 1 librarian working ½ day, 1 aide). The main purpose of the ESS media center is to store and make available print and nonprint media. Approximately 92% of its activities deal with this function, while about 8% of its work revolves around reference services. There is no attempt to analyze or repackage data for users.

[1] Hereafter cited as ESS.

Searching line		
Process and File	**Montgomery Blair**	**ESS**
Identification of user need, general	Information of needs of students and teachers in general derived from curriculum guides published by Montgomery County and from general curriculum goals of the individual schools.	
Identification of user need, specific	No interest profiles are kept on teachers or students. Some attempt is made to learn of interests of students and teachers through personal contact on a totally informal basis. Otherwise, user needs are learned at time of search requests.	No interest profiles. Little rapport between teachers and librarians.
Search requests in terms of descriptors	All search requests are translated by user or librarian into descriptors appropriated to Dewey notation.	

Storage line		
Process and File	**Montgomery Blair**	**ESS**
Acquisition	Head librarian is 75% responsible for acquisitions, with 20-25% input from teaching staff.	95% of acquisitions made by librarian, with 5% input from staff.
Documents, data as required	Documents acquired through Montgomery County School System. Time lag between order and possession of data is between 6 mo. and 2 years. In both schools print media is shelved in reading room; nonprint media is shelved in another room across from media center. Montgomery Blair librarians purchase books personally in neighborhood book stores to meet demands of users. Departmental collections are housed throughout the building in appropriate rooms. (This point should actually be mentioned later, D.S.)	

Storage line (continued)		
Process and File	**Montgomery Blair**	**ESS**
Indexing	Print and nonprint media are indexed using Dewey schedule by catalogers of Montgomery County School System. Indexing of data acquired prior to establishment of county system is done by individual librarians using Dewey notation and Sears headings.	
Feeding into storage	Student aides perform this function.	Library aide performs this.
Storage	The files are organized on card catalogues located in the center of the reading room. Easily accessible . All print and non-print documents plus department collections and textbooks are indexed in the file.	The card catalogue is located alongside the circulation desk. Very inconvenient.

Retrieval		
Process and File	**Montgomery Blair**	**ESS**
Documents retrieved	User or librarian retrieve data, which is immediately for use either in media center, in class or at home.	
	Student aides carry on circulation functions, supervised by librarian.	Library aide handles circulation.
Further process of material	None	None
Public relations	Librarians print and distribute to teachers monthly lists of new materials. Sophomores receive a general orientation to the media center in one assembly program. Nothing else.	None

Conclusion

Information systems for children and adolescents based on these two schools show strong similarities. Their purpose is to store and make available documents. They are libraries first, in spite of their fancy new name of "media center." Functions and files on the storage line are far more developed than those on the searching line. The entire area of user needs requires exploration and creativity.

The public relations function is another area worthy of study and improvement, since it barely exists at all in the schools observed for this report.

Space and money will be important to bring changes, but just as crucial to change is the need for creative personnel, not just librarians but teachers, as well, since they are the users and can direct their students to the media center. It would be interesting to study what and how teachers are taught in the use of media centers as part of their teaching strategies.

With such a heavy emphasis on the storage function it is no wonder that children grow up thinking that libraries are static. An emphasis on the searching line is due to change this imbalance.

Charles E. White Chemistry Library

Conceptual data schema

The conceptual data schema, or rules for cataloging, is described in the **Anglo-American Cataloging Rules** (AACR2). The rules indicate how items in the library must be described and how subject headings (access points or descriptors) should be established. The conceptual data schema is extended by the **Library of Congress Subject Headings**, which lists allowable values for subject headings, and the **Library of Congress Classification System**, which gives allowable values for classification numbers. Therefore, these two documents serve as the thesauri for the system.

The AACR2 document specifies the data which must be stored for the entities and, for manual card catalog entries, indicates how the data should be displayed on the catalog card. For the on-line catalog system, the system's computer programs impose further rules for how query and entity data should be input into the system, as well as how it is displayed on the terminal screen.

The **Journal Shelving Authority File** gives rules for the physical arrangement of journals. The physical arrangement of books is based on the **Library of Congress Classification System.**

Files and processes

File 1: People with problems and resulting needs. The primary component of this file is faculty members and undergraduate and graduate students in the chemistry, biochemistry, and microbiology departments. The file could also contain other members of the university community and the public at large. Their needs could be information required for problems in course work, research, or general interest.

Process 1: Identification of needs in general. Setting priorities. The library views the users' needs as the need for access to a comprehensive collection of information and materials in the areas of chemistry, biochemistry, and microbiology.

File 2: File of needs in general. A formal information directory of user needs is not maintained.

Process 2: Acquisition of the needs of specific users. In many cases, the student or faculty member analyzes his problem himself and has a clear idea of the nature of his need. In other cases, when the user is inexperienced or when the problem is complex, one of the librarians assists in the determination of the specific need by conducting a reference interview.

File 3: Query statements as acquired. This consists of a statement of the user's problem in natural language. It is located in the user's and/or librarian's memory, and may also be written on a piece of paper.

Process 3: **Formulating in terms of entities and relationships.** This is performed by the user, sometimes with the assistance of the librarian, and involves the determination of descriptors which will be used in the search. If the need is very specific (i.e. the user needs a specific book or work by a specific author), the title or author's name to be used as an access point is determined. If the user simply needs information on a topic, the relevant subject heading(s) which describe the user's problem are determined.

Process 4: **Feeding into storage.** For a search in the card catalog, the user or librarian writes the descriptor(s) on a piece of paper, or may simply store them in memory.

For a search in the online catalog (OPAC), the user or librarian selects the proper descriptor type (AUT, TIL, or SUB) and then keys in the desired descriptor.

File 4: **Query store.** For a search in the card catalog, this consists of a list of author(s), title(s), or subject heading(s) stored in the user's or librarian's memory or written on a piece of paper.

For a search in the online catalog, this consists of a title, author, or subject heading stored in the memory of the computer at McKeldin. Query 'files' are temporary; since there is no SDI function, no queries are stored in permanent computer files.

File 5: **Information or entities.** This file could be defined as the entire body of the world's recorded knowledge; a more realistic definition might include only published "knowledge".

Process 5: **Selection and acquisition of entities and information.** The library's goal is to maintain a comprehensive collection of 'the best' materials on chemistry, biochemistry, and microbiology.

The selection function is performed by the two librarians at the Chemistry library using the following three methods:

(1) Books which match the library's 'profile' are selected by the Ballen book approval plan.

(2) Faculty and graduate students recommend books and journals.

(3) Librarians select books and serials described in marketing materials received from publishers.

Librarians forward a list of desired materials to the Acquisitions unit at McKeldin. This unit types orders and mails them to the publishers.

Shipments of books are received at McKeldin from the publishers about once a week. Before cataloging is performed, Chemistry librarians examine books selected via the book approval plan and may reject ones which they do not want; rejected books are returned to the publisher.

Serials are received by the chemistry library directly.

File 6: Entities and information as acquired. This consists of uncataloged books at McKeldin library and unprocessed serials at the chemistry library.

Processes 6 and 7 must be considered together as a 4-step process.

Process 6: Indexing, intellectual process. Establishing relationships between entities. Classification assignment and preliminary cataloging. This is performed by the Catalog Management unit at McKeldin, who assign the call number and decide upon the form of the author and title under which the book will be cataloged. The call number is pasted on the book.

Process 7: Feeding into storage. Building and maintaining the database. This is performed by the Library Technical Assistant at Chemistry library, with the help of student assistants, in the following steps:

(1) Make preliminary entry (call number, author, title, CSN number only) into the online catalog, so that books may circulate.

(2) Make entry in New Acquisitions Card Catalog.

(3) Shelve book on New Book Shelf for a week.

(4) After a week, shelve book in stacks.

Process 6: Indexing, intellectual process. Establishing relationships between entities. Final cataloging. This is performed by the Catalog Management unit at McKeldin. The MARC catalog record is received from OCLC and revised if necessary to produce a final catalog entry. In some cases, original cataloging must be done.

Process 7: Feeding into storage. Building and maintaining the database. This is performed by Catalog Management unit at McKeldin. The preliminary entry in the online catalog system is overlaid with a final catalog entry.

File 7: Entity store, database.

(1) Book Stacks. This contains books shelved in order by classification no.

(2) New Book Shelf. This contains books acquired during the past week. They are not shelved in any particular order and may not circulate.

(3) Current Serials Shelves. This contains serial issues published during the past two months. They are shelved in alphabetical order by authority title.

(4) Bound Serials Stacks. This contains serial issues published three or

more months ago. Usually, all issues for a volume are bound together. The shelved in alphabetical order by authority title.

(5) Online Catalog. This is a computer file maintained on a computer in the basement of McKeldin library. It contains bibliographic records (entities) which are representations of materials added to UMCP libraries since 1983 (including items in 1-4 above). Entities are indexed by author, title, and subject heading.

(6) Card Catalog. This is a file of 3x5 cards, maintained in the Chemistry library, which contains bibliographic records which are representations of materials acquired by the Chemistry Library before 01/01/87. Entities are indexed by author, title, and subject heading.

(7) New Acquisitions Catalog. This is a file of 3x5 index cards, which contains bibliographic records which are representation of materials acquired by the chemistry library since 01/01/87. Entities are indexed by title only.

Process 7 also includes the reshelving of materials which were previously retrieved from the database. This is performed by student assistants.

Process 8: **Comparison-match, display.** If the query is processed using the card catalog, this is performed by the user or librarian. He searches for card(s) containing the proper descriptor (author, title, subject heading) in the catalog. The card(s) found will display bibliographic data (descriptive data and access point) about the entities which have been indexed with the descriptor.

If the query is processed using the online catalog, the computer program performs the match and display. The system attempts to match the descriptor entered on the screen (Query Store file) with index terms in the bibliographic record file. If it finds a match, it displays the bibliographic data about the entity on the screen.

The user, occasionally with the assistance of the librarian, then uses the information displayed on the card or screen to locate the materials on the shelves.

File 9: **Entities or information retrieved.** This consists of materials retrieved from the Book Stacks, New Book Shelf, Current Serial Shelves, or Bound Serials Stacks.

Process 9: **Check whether satisfactory.** The user examines the materials retrieved to determine whether they are relevant to his problem.

Process 10: **Further processing (data analysis, extracting substantive data).** This function is not performed by the library.

File 10: **Materials processed according to user needs.** This file is not used by the library.

Process 11: **Making material available to user.** Users may check out many of the books found in the book stacks. This circulation process is performed at the service desk by any of the 14 employees. Other materials may not be checked out and are thus photocopied by the user if he wishes to take a copy out of the library. Alternatively, he may simply read the materials at the library.

File 11: **Output: User has information or entities available.** This consists of a student, faculty member, or other person who has the information he needs.

Process 12: **Public relations.** The library publishes a brochure describing its services. The librarians also make presentation to new graduate students in chemistry, biochemistry, and microbiology, as well as some undergraduate classes, about the library's services.

File 12: **Output: Potential users know about services available.** This consists of the part of the academic community which has at least a basic knowledge of the library's services.

SLecture 5.1 RDF, linked data, SPARQL

RDF schema definition for the food database

Name space declarations
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dsfoods: <www.dsoergel.com/foodschema#> .
@prefix dsfoodd: <www.dsoergel.com/fooddata#> .

 # Class (entity type) definitions

dsfoods:FoodProduct
rdf:*type* rdfs:Class .

dsfoods:Organism
rdf:*type* rdfs:Class .

dsfoods:OrganismPart
rdf:*type* rdfs:Class .

dsfoods:Substance
rdf:*type* rdfs:Class .

dsfoods:Purpose
rdf:*type* rdfs:Class .

dsfoods:Process
rdf:*type* rdfs:Class .

dsfoods:ProcessIntensity
rdf:*type* rdfs:Class .

dsfoods:Form
rdf:*type* rdfs:Class .

dsfoods:Container
rdf:*type* rdfs:Class .

 # Property (relationship type) definitions

dsfoods:hasName
rdf:*type* rdf:Property ;
rdfs:*domain* dsfoods:FoodProduct ;
rdfs:*range* rdfs:String .

dsfoods:isA
rdf:*type* rdf:Property ;
rdfs:*domain* dsfoods:FoodProduct ;
rdfs:*range* rdfs:FoodProduct .

dsfoods:comesFromSource
rdf:*type* rdf:Property ;

rdfs:*domain* dsfoods:FoodProduct ;
rdfs:*range* rdfs:Organism .

dsfoods:comesFromPart
rdf:*type* rdf:Property ;
rdfs:*domain* dsfoods:FoodProduct ;
rdfs:*range* rdfs:OrganismPart .

dsfoods:isExtractedSubstance
rdf:*type* rdf:Property ;
rdfs:*domain* dsfoods:FoodProduct ;
rdfs:*range* rdfs:Substance .

dsfoods:isMadeFrom
rdf:*type* rdf:Property ;
rdfs:*domain* dsfoods:FoodProduct ;
rdfs:*range* rdfs:FoodProduct .

dsfoods:hasIngredient
rdf:*type* rdf:Property ;
rdfs:*domain* dsfoods:FoodProduct ;
rdfs:*range* rdfs:FoodProduct .

dsfoods:processedBy
rdf:*type* rdf:Property ;
rdfs:*domain* dsfoods:FoodProduct ;
rdfs:*range* rdfs:Process .

dsfoods:hasForm
rdf:*type* rdf:Property ;
rdfs:*domain* dsfoods:FoodProduct ;
rdfs:*range* rdfs:Form .

dsfoods:packedIn
rdf:*type* rdf:Property ;
rdfs:*domain* dsfoods:FoodProduct ;
rdfs:*range* rdfs:Container .

dsfoods:eats
rdf:*type* rdf:Property ;
rdfs:*domain* dsfoods:Organism ;
rdfs:*range* rdfs:FoodProduct .

Assume this stored at

File found at www.dsoergel.com/FoodSchema

```
@prefix rdf:        <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dsfoods:    <http://www.clis.umd.edu/faculty/soergel/FoodSchema> .
@prefix dsfoodd:    <http://www.clis.umd.edu/faculty/soergel/FoodSchema> .
```

 # Entity values assigned to entity types

```
dsfoodd:CarrotPlant
        rdf:type        dsfoods:Organism .

dsfoodd:Root
        rdf:type        dsfoods:OrganismPart .

dsfoodd:Diced
        rdf:type        dsfoods:Form .
```

 # Statements on Food Products

```
dsfoodd:FP0
        rdf:type                dsfoods:FoodProduct ;
        dsfoods:hasName         "Food Product" .

dsfoodd:FP1
        rdf:type                dsfoods:FoodProduct ;
        dsfoods:hasName         "Vegetable product" ;
        dsfoods:isA             dsfoodd:FP0 .

dsfoodd:FP11
        rdf:type                        dsfoods:FoodProduct ;
        dsfoods:hasName                 "Diced carrots" ;
        dsfoods:isA                     dsfoodd:FP1 ;
        dsfoods:comesFromSource         dsfoodd:CarrotPlant ;
        dsfoods:comesFromPart           dsfoodd:Root ;
        dsfoods:hasForm                 dsfoodd:Diced .
```

Supplement for Lecture 5.2. **Further elaboration of data structures**

SLecture 5.2 4.2 Further elaboration of data structures (Advanced, ➜LIS 506 Information Technology)

Relational databases: Storage in tables (example: University Database in Chapter 3)

Each table contains all the statements that use the same relationship type; statements pertaining to one entity value are distributed over many tables. In retrieval, data can be combined in many ways. The system gives equal consideration to the user who wants to know everything about a document, including the person who authored it, and to the user who wants to know everything about a person, including the documents he authored.

"Flat file" databases: Storage in records

As discussed in Lecture 4.2 (Organizing Information, Section 9.2), a record assembles the information about one entity value - the various statements that pertain to that entity value. Records are needed for eliciting input and for presenting output. Often, storage is also based on records. With storage records, statements pertaining to one entity value are all in one place, while statements using the same relationship type are distributed over many records. Storage by records introduces a perspective or focus: If data are assembled in document records, the data structure gives more consideration to the user who wants to know everything about a document; the linkage between a document and the person who authored it is stored in the document record. If, on the other hand, data are assembled in person records, the data structure gives more consideration to the user wanting to know everything about a person; the linkage between a document and the person who authored it is stored in the person record. By storing the same information twice, both users can be accommodated.

See also the example on bibliographic data in MARC records (flat file) and as a collection of

Object-oriented databases are based on frames with hierarchical inheritance (see Lecture 2.2). They are closer to the record model than to the relational model.

Searching printed indexes vs. searching by computer.

Division of labor between system and user: Degree of order and amount of information presented in search output (See example 13 from *Design of an integrated information structure interface. Prologue*.)

Supplement for Lecture 6.1-6.2. Document function, structure, analysis, and design.
SLecture 6.1a Elaboration of text types adapted from
Beaugrande *Text, discourse, and process*, VII.1.8

Descriptive	The text revolves around **object and situation concepts**, about which statements are made through links in multiple directions. The link types of *state, attribute, instance, and specification* are frequent. The surface text reflects a corresponding density of *modifier* dependencies. The most commonly applied global knowledge pattern is *the frame*.
Argumenta -tive	The text revolves around **entire propositions** which are assigned values of truthfulness and give reasons for considering beliefs as facts; often there is an opposition between propositions with conflicting value and truth assignment. The link types of *value, significance, cognition, volition, and reason* are frequent. The surface text contains a density of evaluative expressions. The most commonly applied global knowledge pattern is the plan whose goal state is the inducement of shared beliefs.
Didactic	The text revolves around a **topic or theme about which the receiver is to learn something**, that is, integrate new objects and relationships into her memory. The text must present the subject via a process of gradual integration, because the receiver does not yet have the matchable knowledge spaces that a scientific text would require. Therefore, the linkages of estab-lished facts are problematized (put into question) and then de-problematized.
Narrative	The text revolves around the main **event and action concepts** which are arranged in an *ordered directionality* of linkage. The link types of *cause, reason, enablement, purpose,* and *time proximity* are frequent). The surface text reflects a corresponding density of *subordinative* dependencies. The most commonly applied global knowledge pattern is the *schema*. (Freedle and Hale (1979) show that a narrative schema, once learned, can easily be transferred to the processing of a descriptive text on the same topic.)
Conversa- tional	The text has an especially **episodic and diverse range of sources for admissible knowledge**. Less emphasis on expanding current knowledge of the participants than for the other text types. The surface organization assumes a characteristic mode because of the changes of speaking turn.

Literary	The text revolves around **alternatives to matchable patterns of knowledge about the accepted real world**. The intention is to motivate, via contrasts and rearrangements, some new insights into the organization of the real world. From the standpoint of processing, the linkages within real-world events and situations is PROBLEMATIZED, that is, made subject to potential failure, because the text-world events and situations may (though they need not) be organized with different linkages. (*Problematize* = put into question, consider as uncertain, therefore problematic.) The effect is an increased *motivation* for linkage on the side of the text producer and increased focus for linkage on the side of the receiver. This problematized focus sets even "realistic" literature (reaching extremes in "documentary" art) apart from a simple report of the situations or events involved: the producer intends to portray events and situations as *exemplary* elements in a framework of *possible alternatives.* In **poetic texts**, the alternativity principle is extended to the *interlevel mapping of options,* e.g. sounds, syntax, concepts/ relations, plans, and so on. In this fashion, both the organization of the real world and the organization of discourse about that world are problematized, and the resulting insights can be correspondingly richer.
Scientific	The text revolves around an **optimal match with the accepted real world** unless there are explicit signals to the contrary (e.g., a disproven theory). Rather than alternative organization of the world (as in literary text, see above), a more exact and detailed insight into the established organization of the real world is intended. In effect, the linkages of events and situations are eventually *de-problematized* via statements of causal necessity and order.

Lecture 6.1a (20 min)

Document design (information design)
Formatting documents for understanding by people
External representation of information

Further elaboration of these principles through a series of

Document design examples

1 Two formats for salary data

2a Alphabetical vs. meaningful display (Art and Architecture Thesaurus)

2b Alphabetical vs. meaningful display (Art and Architecture Thesaurus)

3 Examples from the Longman Lexicon of the English Language

4 Display of information on buildings on a site in Perseus

5 Two displays of the same hierarchy

6 Two displays of a catalog record in a public library OPAC (Online Public Access Catalog)

In the Supplement, SLecture 5.2d

7 Winners and losers in the forecasting game (from Tufte 1983)

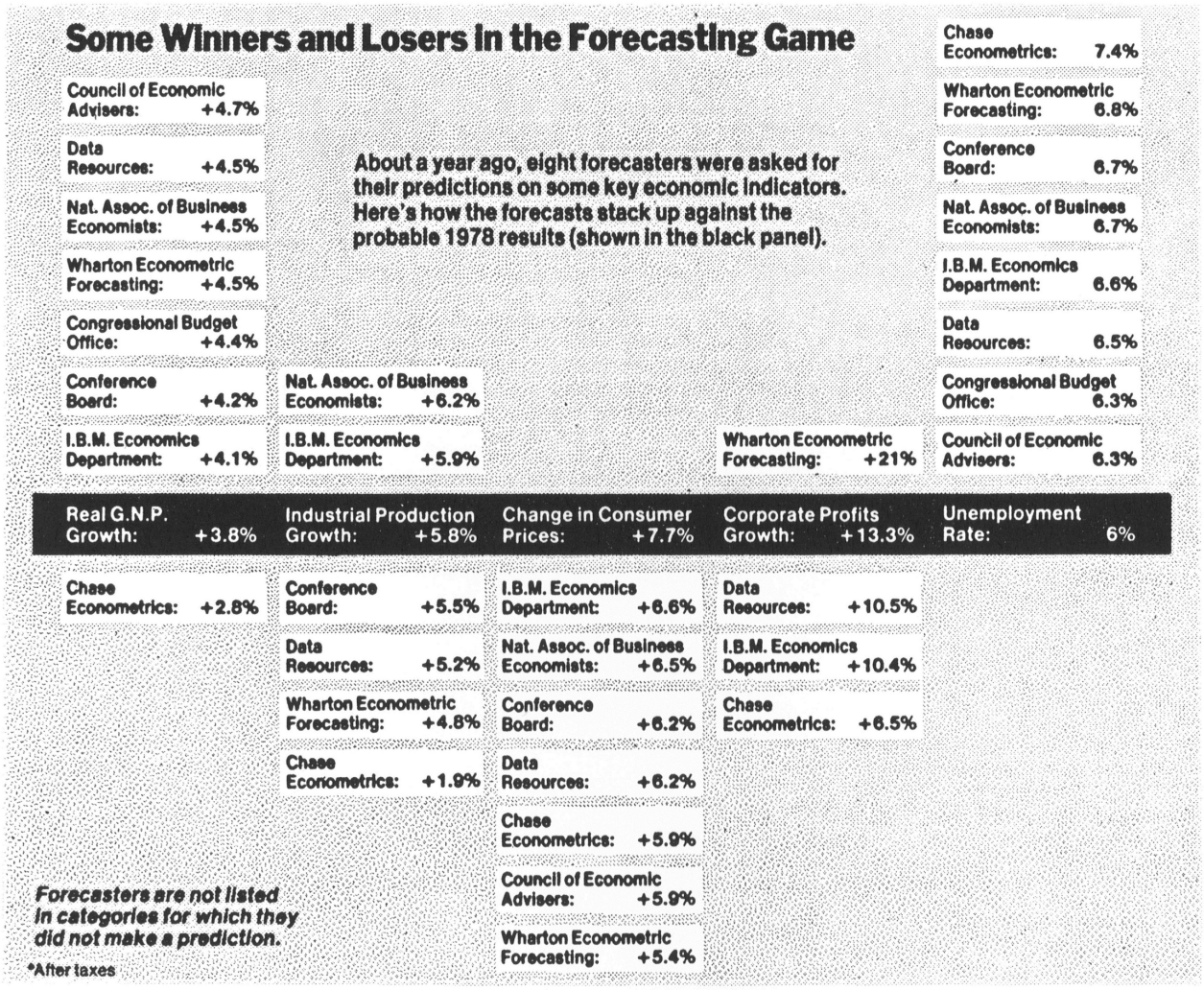
8 Thermal conductivity of tungsten: Arrangement of labels to facilitate interpretation (from Tufte 1983)

9 Napoleon's campaign to Russia (from Tufte1983)

10 Classified arrangement of descriptors in a document record for indexing test (Alcohol and Other Drug Thesaurus)

11 Contents page from *Alcohol Research*

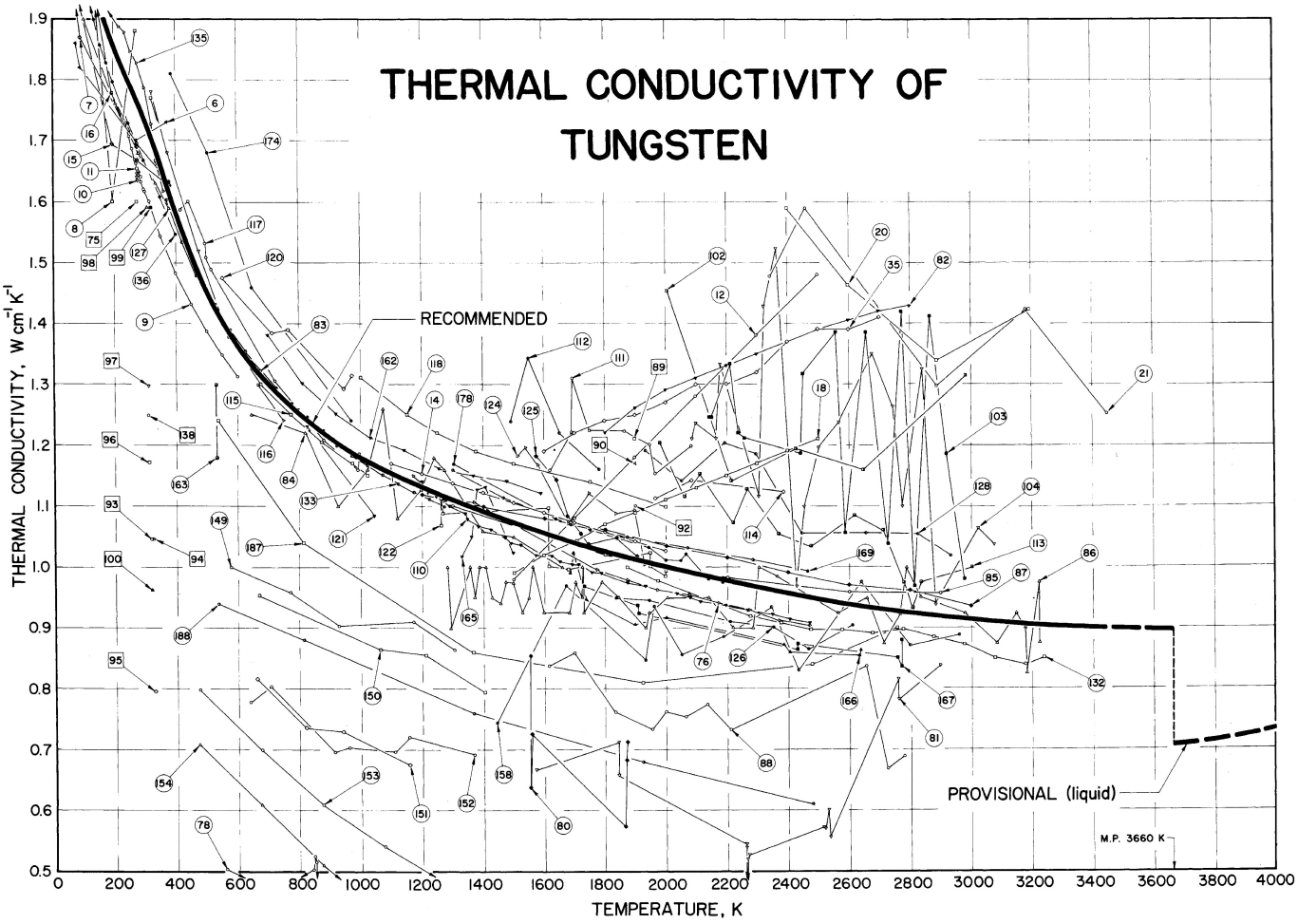
The syllabus and lecture notes are an example of document design, using boxes, labels, comparative columns, tables showing a concept space that has two dimensions (such as the table in this lecture) and color and striving for consistent format. For example, first pages of lectures follow a common format, so do first pages of assignments. Also running heads as guide posts for orientation in the document.



7 Winners and losers in the forecasting game (from Tufte 1983)

Question. If we could use color, how could we make it easier to see how each organization did on each of the predictions.

8 Thermal conductivity of tungsten: Arrangement of labels to facilitate interpretation



(from Tufte 1983)

Double-Functioning Labels

Numbers can double-function when used both to name things (like an identification number) and to reflect an ordering. In this graphic (in which the circled numbers fail to double-function), each number identifies a particular study of the thermal conductivity of tungsten, ordered alphabetically by the last name of the first author. If that list were ordered by date of publication instead, then the code would also indicate the time order in which the various conductivity determinations were made. Thus, "1" would indicate the earliest study, and so on; or, alternatively, "61c" would be the third study published in 1961. Such information has interest, since we could see which of the early studies got the right answer. In addition, the movement of the studies toward the "correct" recommended values could be tracked. This extra information requires no additional ink. (Tufte 1983, p. 149 - 150)

Tufte Russia see next page

9 **Napoleon's campaign to Russia** (from Tufte1983)

An especially effective device for enhance the explanatory power of time-series displays is to add spatial dimensions to the design of the graphic, so that the data are moving over space (in two or three dimensions) as well as over time. Three excellent space-time- story graphics illustrate here how multivariate complexity can be subtly integrated into graphical architecture, integrated so gently and unobtrusively that viewers are hardly aware that they are looking into a world of four or five dimensions. Occasionally graphics are belligerently multivariate, advertising the technique rather than the data. But not these three.

The first is the classic of Charles Joseph Minard (1781-1870), the French engineer, which shows the terrible fate of Napoleon's army in Russia. Described by E. J. Marey as seeming to defy the pen of the historian by its brutal eloquence,[12] this combination of data map and time-series, drawn in 1861, portrays the devastating losses suffered in Napoleon's Russian campaign of 1812. Beginning at the left on the Polish-Russian border near the Niemen River, the thick band shows the size of the army (422,000 men) as it invaded Russia in June 1812. The width of the band indicates the size of the army at each place on the map. In September, the army reached Moscow, which was by then sacked and deserted, with 100,000 men. The path of Napoleon's retreat from Moscow is depicted by the darker, lower band, which is linked to a temperature scale and dates at the bottom of the chart. It was a bitterly cold winter, and many froze on the march out of Russia. As the graphic shows, the crossing of the Berezina River was a disaster, and the army finally struggled back into Poland with only 10,000 men remaining. Also shown are the movements of auxiliary troops, as they sought to protect the rear and the flank of the advancing army. Minard's graphic tells a rich, coherent story with its multivariate data, far more enlightening than just a single number bouncing along over time. *Six* variables are plotted: the size of the army, its location on a two-dimensional surface, direction of the army's movement, and temperature on various dates during the retreat from Moscow.
It may well be the best statistical graphic ever drawn.

[12] E. J. Marey, *La Methode Graphique* (Paris, 1885), p. 73. For more on Minard see Arthur H. Robinso. The Thematic Maps of Charles Joseph Minard. *Imago Nundi*, 21 (1967), 95-108

Example 10 Alcohol and Other Drugs Thesaurus (AOD Thesaurus)

Many examples of meaningful sequence (Baldy 14A or http://etoh.niaaa.nih.gov/aodvol1/aodthome.htm).

Sample document record from AOD Thesaurus indexing test (next page)

To test the AOD Thesaurus, 20 indexers indexed 25 documents. A cumulative list of the descriptors assigned to each document was then printed. Each descriptor is followed by a list of symbols identifying the indexers who assigned this descriptor. The list is arranged in classified order, facilitating analysis. For example, if the indexers among them assigned several related descriptors, it is easy to see that most indexers covered the basic concept but chose slightly different descriptors; then one can select the best descriptor from those assigned by the various indexers. See the bolded groups at JP8 treatment and MO24.2 public policy on AOD for an illustration. With an alphabetic arrangement of descriptors, this analysis would be much more difficult.

Legend:

***C**orrect*
***B**road* (assigned descriptor is too broad, above the correct descriptor)
***N**arrow* (assigned descriptor is too narrow, below the correct descriptor)
***R**elated* (assigned descriptor is related to the correct descriptor)
***E**xhaustive* (minor point in document)
***T**hesaurus problem* (for example, missing scope note)
***W**rong*

CTRL002 **Substance-abusing chronically mentally ill client: Prevalence, assessment, treatment, and policy concerns**

The bolded groups show assignment of related terms by different indexers.

AB	AODD (ARG) ***B**road*
AB2	AOD abuse (CSRJ, CSRT, CSRP, MAR, BCP) ***B**road*
AM	prevention, diagnosis, and treatment of AODU (CSRJ) ***T**hesaurus problem*
BA	AOD substances of abuse (CSRJ) ***C**orrect*
EC10.10	alcohol interactions (CSRD) ***E**xhaustive*
EC10.8	adverse drug interaction (CSRD) ***E**xhaustive*
FV20.8	assessment (CSRP, BCP) ***B**road*
GA2.12.4	mental dysfunction (BCP) ***N**arrow*
GA2.14.6	dual diagnosis (CSRJ, CSRK, CSRS, CSRA, CSRT, CSRD, CSRP, SHS, MAR, BCP, ...) ***C**orrect*
GA6.10.4.4	chronic disease (CSRJ) ***C**orrect*
GD4	alcohol use disorder (CSRD) ***N**arrow*
GD4.2	alcohol abuse (CSRJ) ***N**arrow*
GD6.2	alcohol related mental disorders (CSRJ) ***N**arrow*
GE2	other drug use disorder (CSRD) ***N**arrow*
GE2.2	other drug abuse (CSRJ) ***N**arrow*
GE4.2	other drug related mental disorders (CSRJ) ***N**arrow*
GY	behavioral and mental disorders (CSRJ, CSRP, MAR) ***N**arrow*
GY2.2.6	other chronic organic psychotic conditions (RIA) ***N**arrow*
HA	screening and diagnostic methods (ARFL) ***B**road*
HB	AODU screening, identification, and diagnostic methods (CSRD) ***C**orrect*
HH2.2	patient AODU history (CAS) ***E**xhaustive*
HK	treatment methods (CSRK, CSRA, CSRD, SHS, ARG) ***C**orrect*

HN10	combined modality therapy (BCP) ***N**arrow*
HX	psychosocial treatment approaches (CSRA, CSRD) ***N**arrow*
HX4.18	cognitive techniques of affect and behavior change (CAS) ***N**arrow*
JK	intervention and treatment (CSRD, ARFL) ***C**orrect*
JM	identification and screening (RIA) ***B**road*
JM2.2	identification and screening for AOD use (SHS, CAS) ***N**arrow*
JP4	patient assessment (CSRJ, CSRK, CSRS, CSRA, CSRD MAR, ARG, CAS, DINF) ***C**orrect*
JP4.4	self report (MAR, RIA) ***N**arrow*
JP8	**treatment** (CSRT, CSRP, BCP, RIA, DINF) ***C**orrect*
JP8.10	**treatment issues** (MAR) ***N**arrow*
JP8.16	**treatment factors** (MAR) ***N**arrow*
JP8.16.2	**patient treatment factors** (CSRK, CSRS) ***N**arrow*
JP8.18.4	**mental health care** (BCP) ***N**arrow*
JT6	mental health services (BCP) ***E**xhaustive*
JV8	health records (RIA) ***E**xhaustive*
MO24.2	**public policy on AOD** (CSRJ, CSRT, BCP, ARFL) ***R**elated*
MO24.2.6	**public policy on other drugs** (ARFL) ***R**elated*
MO24.2.8	**AOD public policy strategies** (DINF) ***N**arrow*
MO24.6	**public policy on health** (CSRK, CSRS, RIA) ***C**orrect*
MT12	employee related issues (CSRT) ***C**orrect*
NM56	literature review (CSRK, CSRP, SHS) ***C**orrect*
OF2	alcoholic beverages (CSRJ) ***E**xhaustive*
PL2.2	incidence and prevalence of AODU (CSRK, CSRT) ***E**xhaustive*
PL2.6	prevalence (CSRS, CSRP, SHS, BCP, ARG, DINF) ***E**xhaustive*
PL4	comorbidity (SHS, ARG) ***E**xhaustive*
PT2.4.6	state wide areas (CSRK) ***C**orrect*
RB	research and evaluation methods (MAR) ***B**road*
RC6.2	survey of research (MAR) ***W**rong*
RM10	assessment of variables and methods (CSRT) ***C**orrect*
RM10.2	reliability (research methods) (RIA) ***N**arrow*
RM10.4	validity (research methods) (RIA) ***N**arrow*
RP	data collection (CSRD) ***C**orrect*
RP10.6.4	interview (RIA) ***E**xhaustive*
SG8.2	social work (field) (CSRP, SHS) ***B**road*
TK4.4.6.2	mentally ill (CSRK, CSRA, CSRT) ***C**orrect*
TL2	AOD user (CSRA, CSRD) ***C**orrect*
TT14.2	social worker (CSRS, DINF) ***C**orrect*

11 Contents page from *Alcohol Research*. A month of the most important results at a glance

Editorial

page 5

Epidemiology

page 7

Medicine

page 16

Physiology

page 25

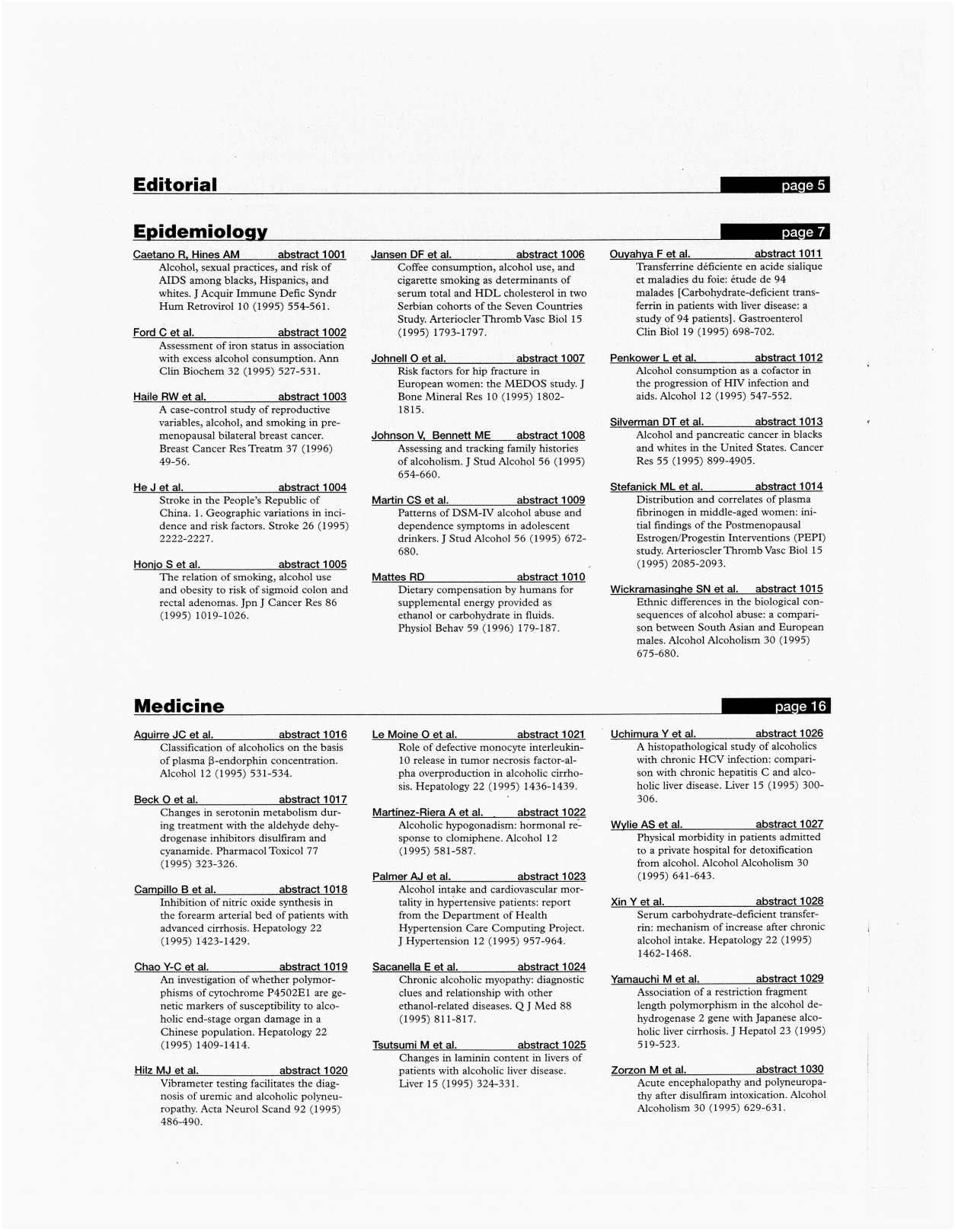
New books

Psychology

page 43

Image of the actual looks, but hard to read

Example 2. A more complex document system (World Bank)



Editorial page 5

Epidemiology page 7

Caetano R, Hines AM abstract 1001
Alcohol, sexual practices, and risk of AIDS among blacks, Hispanics, and whites. J Acquir Immune Defic Syndr Hum Retrovirol 10 (1995) 554-561.

Ford C et al. abstract 1002
Assessment of iron status in association with excess alcohol consumption. Ann Clin Biochem 32 (1995) 527-531.

Haile RW et al. abstract 1003
A case-control study of reproductive variables, alcohol, and smoking in premenopausal bilateral breast cancer. Breast Cancer Res Treatm 37 (1996) 49-56.

He J et al. abstract 1004
Stroke in the People's Republic of China. 1. Geographic variations in incidence and risk factors. Stroke 26 (1995) 2222-2227.

Honjo S et al. abstract 1005
The relation of smoking, alcohol use and obesity to risk of sigmoid colon and rectal adenomas. Jpn J Cancer Res 86 (1995) 1019-1026.

Jansen DF et al. abstract 1006
Coffee consumption, alcohol use, and cigarette smoking as determinants of serum total and HDL cholesterol in two Serbian cohorts of the Seven Countries Study. Arteriocler Thromb Vasc Biol 15 (1995) 1793-1797.

Johnell O et al. abstract 1007
Risk factors for hip fracture in European women: the MEDOS study. J Bone Mineral Res 10 (1995) 1802-1815.

Johnson V, Bennett ME abstract 1008
Assessing and tracking family histories of alcoholism. J Stud Alcohol 56 (1995) 654-660.

Martin CS et al. abstract 1009
Patterns of DSM-IV alcohol abuse and dependence symptoms in adolescent drinkers. J Stud Alcohol 56 (1995) 672-680.

Mattes RD abstract 1010
Dietary compensation by humans for supplemental energy provided as ethanol or carbohydrate in fluids. Physiol Behav 59 (1996) 179-187.

Ouyahya F et al. abstract 1011
Transferrine déficiente en acide sialique et maladies du foie: étude de 94 malades [Carbohydrate-deficient transferrin in patients with liver disease: a study of 94 patients]. Gastroenterol Clin Biol 19 (1995) 698-702.

Penkower L et al. abstract 1012
Alcohol consumption as a cofactor in the progression of HIV infection and aids. Alcohol 12 (1995) 547-552.

Silverman DT et al. abstract 1013
Alcohol and pancreatic cancer in blacks and whites in the United States. Cancer Res 55 (1995) 899-4905.

Stefanick ML et al. abstract 1014
Distribution and correlates of plasma fibrinogen in middle-aged women: initial findings of the Postmenopausal Estrogen/Progestin Interventions (PEPI) study. Arterioscler Thromb Vasc Biol 15 (1995) 2085-2093.

Wickramasinghe SN et al. abstract 1015
Ethnic differences in the biological consequences of alcohol abuse: a comparison between South Asian and European males. Alcohol Alcoholism 30 (1995) 675-680.

Medicine page 16

Aguirre JC et al. abstract 1016
Classification of alcoholics on the basis of plasma β-endorphin concentration. Alcohol 12 (1995) 531-534.

Beck O et al. abstract 1017
Changes in serotonin metabolism during treatment with the aldehyde dehydrogenase inhibitors disulfiram and cyanamide. Pharmacol Toxicol 77 (1995) 323-326.

Campillo B et al. abstract 1018
Inhibition of nitric oxide synthesis in the forearm arterial bed of patients with advanced cirrhosis. Hepatology 22 (1995) 1423-1429.

Chao Y-C et al. abstract 1019
An investigation of whether polymorphisms of cytochrome P4502E1 are genetic markers of susceptibility to alcoholic end-stage organ damage in a Chinese population. Hepatology 22 (1995) 1409-1414.

Hilz MJ et al. abstract 1020
Vibrameter testing facilitates the diagnosis of uremic and alcoholic polyneuropathy. Acta Neurol Scand 92 (1995) 486-490.

Le Moine O et al. abstract 1021
Role of defective monocyte interleukin-10 release in tumor necrosis factor-alpha overproduction in alcoholic cirrhosis. Hepatology 22 (1995) 1436-1439.

Martínez-Riera A et al. abstract 1022
Alcoholic hypogonadism: hormonal response to clomiphene. Alcohol 12 (1995) 581-587.

Palmer AJ et al. abstract 1023
Alcohol intake and cardiovascular mortality in hypertensive patients: report from the Department of Health Hypertension Care Computing Project. J Hypertension 12 (1995) 957-964.

Sacanella E et al. abstract 1024
Chronic alcoholic myopathy: diagnostic clues and relationship with other ethanol-related diseases. Q J Med 88 (1995) 811-817.

Tsutsumi M et al. abstract 1025
Changes in laminin content in livers of patients with alcoholic liver disease. Liver 15 (1995) 324-331.

Uchimura Y et al. abstract 1026
A histopathological study of alcoholics with chronic HCV infection: comparison with chronic hepatitis C and alcoholic liver disease. Liver 15 (1995) 300-306.

Wylie AS et al. abstract 1027
Physical morbidity in patients admitted to a private hospital for detoxification from alcohol. Alcohol Alcoholism 30 (1995) 641-643.

Xin Y et al. abstract 1028
Serum carbohydrate-deficient transferrin: mechanism of increase after chronic alcohol intake. Hepatology 22 (1995) 1462-1468.

Yamauchi M et al. abstract 1029
Association of a restriction fragment length polymorphism in the alcohol dehydrogenase 2 gene with Japanese alcoholic liver cirrhosis. J Hepatol 23 (1995) 519-523.

Zorzon M et al. abstract 1030
Acute encephalopathy and polyneuropathy after disulfiram intoxication. Alcohol Alcoholism 30 (1995) 629-631.

A frame/object hierarchy of document templates and documents

A document template is a frame with a slot for each part of the document (a part can be a single line or part of a line). Many slots have a procedure attached; the procedure obtains the information from a database, if it is available, or displays a menu of possible values, or asks the user a question. The document templates are arranged in a hierarchy, so that the slots in common to all documents of a class, such as meeting announcements, need to be specified only once; these slots than inherit down to all descendants of the class.

Figure WB1. **Frame/object hierarchy of document templates and documents**

Template for generic meeting announcement (Figure WB2)

- Template for announcement of any meeting of group A
  - Announcement of a specific meeting of the group (Figure WB3)
- Template for announcement of any meeting of group B
- etc.

Template for generic meeting agenda

- Template for the agenda of a meeting of a specific group (Figure WB4)

Template for an internal report

- Template for a progress report
  - Template for a market study progress report
- Template for market study report
  - Template for a market study progress report
- Template for a technical development report

Note: This figure presents the "deep structure", the slots with a brief description of their attached procedures. The next figure presents an example with filled-in values.

Figure WB2. **Meeting announcement template**

Author: Fill in on sign-on (*Most other instructions are to carried out by the system*)

Parent document (class hierarchy): Announcement

Child document (class hierarchy): Country Review Committee meeting announcement

Group that is meeting: Possible values: Get list of groups for which the author calls meetings

Note: "Get" is an instruction to the system to get this information from database. A complete template includes the database query to be used. Possible values are automatically displayed in a pop-up menu with multiple selection.

Business function served: Possible values: Get list of business functions in which this group is involved.

Note: The meeting must deal with one or more of these business functions

Purpose: Possible values: Get list of purposes served by this group

Receiver: Possible values: Get list of group members, of regular guests, and of others who should receive agendas of this group's meetings

Date and time: Fixed date and time (on the 15th of each month) or relative date and time (e.g., every second Tuesday at 10 am) or

Get schedules of all participants and determine the date(s) and time(s) at which everybody can make it.

Note: System fills in time or presents pop-up menu if there are several possibilities.

Room: Default room. If not available, select a big enough room that is available at the meeting time from room scheduling database.

Deadline:

Time needed for preparation:

Starting date for preparation:

Status:

Includes document: Agenda

Attached document:

Figure WB3. **Country Review Committee meeting announcement — template filled in**

Author: R. Singh (*in this more specific template, filled in by system on sign-on*)

Parent document (class hierarchy): Meeting announcement

Child document (class hierarchy): Specific Country Review Committee meeting announcements

Group that is meeting:

Country Review Committee
(*filled in by system since this is the only group for which R. Singh calls meetings*)

Business function served:

(*System retrieves business functions served by this group from database and displays pop-up menu. Author selects the business function(s) applicable to this meeting.)*

2.1 Develop country operations strategy

2.3 Approve projects

2.4 Supervise projects through completion

Purpose:

Receiver:

Members: R. Singh, B. Smith, J. Dubois
Guests: D. Suarez
(*All filled in by system from database information*)

Date and time:

Monday, November 30, 1992, 10 am
(*Determined by system based on schedules of participants and general instruction: End of every month*.)

Room: F1057 (*Determined by system*)

(continued on next page)

Figure WB3. **Country Review Committee meeting announcement — continued**

Deadline:

Monday, November 2, 1992
(4 weeks before meeting date)

Time needed for preparation:

3 days (elapsed time)

Starting date for preparation:

Thursday, October 26, 1992

Status:

In process

Includes document:

Agenda for Country Review Committee

Attached document:

Determined based on agenda

Figure WB4. **Agenda for Country Review Committee — template**

Information needed:

Status of country operations strategy

From: Country desk

If decisions needed and all necessary documents are ready, put on agenda

Documents needed for deliberation (attachments to meeting announcement)

Status of projects in the appraisal process

From: Project management database

Get projects for which the appraisal is completed. Put on agenda

Appraisal report as attachment

Status of operating projects

From: Project management database

Get projects for which a review is due. Put on agenda

Project progress report as attachment

Note: Once the agenda is complete, it can be used to automatically generate deadlines for documents needed at the meeting (or a specified time prior to the meeting) and send appropriate messages to the authors of these documents. Such a message, in conjunction with the template for the requested document, can in turn be used to automatically update the work plan of the recipient.

SLecture 6.1b Supplement. Hypertext

Citation relationships These are used in a citation index, such as SCI (Science Citation Index) but without differentiating types of citation relationships	• Giving the source of data and ideas in order to enable checking (authenticating), call on an authority, or give credit. • Referring to documents that describe methodology, equipment etc. • Providing background reading; citing whole sections from another document so as to avoid rephrasing an idea already formulated elsewhere but needed for background (avoiding redundancy). • Providing pointers to further reading, including forthcoming work. • Criticizing or correcting previous work (one's own or others).

Notes	1 In hypermedia systems the line between within-document relationships defining the document macrostructure and inter-document relationships becomes blurred. 2 Citation relationships and relationships (links) in hypermedia systems are often untyped, leaving the reader to guess what the relationship is. In the context of the World Wide Web, there are efforts to allow for the specification of link types.

Discussion questions	1 How can we design hypermedia systems that support the user in constructing coherent documents? 2 When should sequence be in the writer's hands, and when should it be in the reader's hands?

gold **Assignment 7** ***Assigned: Feb. 25***
Lecture 6.1b ***Due: Mar. 4***

Applying linguistic techniques to retrieval problems

Objectives	Understand, through exploration, the possible improvements in free-text retrieval that can be achieved through **linguistic techniques** from Lecture 6.1b such as (for complete list see Lecture 6.1b) (P2.3.3,1) 1. Using all terms that designate a query concept (all **synonyms** of the query term). (P2.3.8,2#) 2. **Word Sense Disambiguation (WSD)** by syntactic analysis to determine part of speech (POS) and/or noun phrases (NP) and by semantic interpretation (from the multiple meanings of a homonym or polyseme, pick out the one that applies in the context. PXXX 3. **Resolution of anaphoric references** (what do *it*, *she*, *they*, *the machine*, ... refer to). PXXX
Materials: Explanation of the query.	The proximity operator **WS** requires that the two words occur **within the same sentence**. Thus the query formulation *forest* **WS** *fire* retrieves all passages in which the two **words** occur in the same sentence. This is the operator used in the baseline query formulation in the assignment. Most IR systems will take this query quite literally and look for the words (and that is how you need to analyze retrieval performance in Task 1. But the user is interested in the close mention of two concepts. That is where linguistic techniques come in.
Tasks	Explore possible improvements in free-text retrieval through **linguistic techniques**, using the examples in Table 1, which give some short passages of text and a query to be applied to this "collection". (1) analyze retrieval performance of a query using the **WS** operator and (2) (main task) suggest linguistic techniques that could be added to the retrieval system to improve retrieval. See the next page for more detailed instructions. You should still adhere to the requirement that the two concepts must be mentioned somehow in the same sentence.
Deliverables	The filled-out Tables 1 -3 with some analysis of Table 3.
Time	2 hours

over

LIS 571 Soergel Spring 2015 Assignment 7 Name
gold

Task 1 **Prelude**:

In Table 1 (facing page) for all passages that are relevant to the user's need as expressed in the query, put Y in the *Relevant* column; for all other passages put N.

Then for all passages that are retrieved by the query formulation, put Y in the *Retrieved* columns; for all other passages, put N

Fill in the 3x3 grid in Table 2 and compute performance measures: recall, discrimination, precision.

Task 2 **Main point:**
What **linguistic techniques** could be used to improve free-text retrieval performance? (Adding index terms to the passages is not an option.)
In Table 2, analyze each passage in turn; check for each the applicable linguistic technique(s).
In Table 3 summarize retrieval effects. For each technique, list all affected passages and indicate the effect: If the passage is now correctly retr

Query statement (description of information need / topic): **Forest fires**

Query formulation: forest **WS (W**ithin same **S**entence) fire* (fire* finds fire or fires)

Take out Table 1 (on next page) for passages to be retrieved and do Task (1), then fill in Table 2.

Then do Task (2).

In Table 1, check for each passage the linguistic technique(s) that would improve retrieval.
Then summarize the effects for

Table 1. Passages to be retrieved **Query formulation**: forest **WS** fire* (fire* finds fire or fires)

• For each passage, in col.1 put Y if it is relevant, N if it is not • Then for each passage, in col.2 put Y if it is retrieved, N if it is not • For passages that are *relevant and retrieved* or *not relevant and not retrieved* no action is needed, no linguistic technique needs to be used. • For passages that are *relevant but not retrieved* or *not relevant but retrieved*, check linguistic techniques that would solve the problem (but preferably not introduce new problems). Check all techniques that apply. **Passage**		Linguistic technique to use								
	Relevant	Retrieved	None needed	Synonyms	Noun Phrase	Word Sense Dis	Pt Of Speech	Anaphora res	
P1	Forest fires in Indonesia cause serious air pollution in South East Asia.									
P2	The fire in Yellowstone Park destroyed 25% of the forest.									
P3	The fire station is located behind the city forest.									
P4	With fire in her eyes she chased him through the forest.									
P5	The soldiers opened fire into the forest.									
P6	The fire went out of control. It reached the forest and destroyed many acres.									
P7	The animal got scared by the fire burning in the field. It ran into the forest.									
P8	He asked whether he should fire the forest workers.									
P9	Many square miles of forest in the West are burning.									
P10	The dry wooded area went up in flames.									

Table 2. Recall, discrimination, precision

	Relevant	Not relevant	All
Retrieved			
Not retrieved			
All			

Recall: _____ **Discrimination:** _____ **Precision:** _____

$$\frac{\textit{relevant correctly retrieved}}{\textit{all relevant}} \qquad \frac{\textit{irrelevant correctly rejected}}{\textit{all irrelevant}} \qquad \frac{\textit{relevant correctly retrieved}}{\textit{all retrieved}}$$

Table 3. Linguistic techniques effect on individual passages

In the following table, enter only passages whose retrieval status changed by applying the technique. The row for synonym expansion is already filled in

	Passage relevant		**Passage not relevant**	
	Good change	Bad change	Good change	Bad change
	Was not retrieved Now retrieved	Was retrieved Now not retrieved	Was retrieved Now not retrieved	Was not retrieved Now retrieved
Synonym expansion	P9, P10	none	none	none
Noun phrase				
WSD				
Part of speech				
Anaphora resolution				

Table 4. Linguistic techniques effect summary.

	Effect on recall	**Effect on discrimination**
Synonym expansion	Always increase	No effect in sample, but could decrease (if an added synonym has other meanings)
Noun phrase		
WSD		
Part of speech		
Anaphora resolution		

SLecture 6.2a Elaboration on XML

Materials included for XML

- **An example of a Document Type Definition for a self-assessment memo**.
  DTD's were used before XML Schema was available, and is still used to some extent today (not to speak of the many legacy DTD's still in everyday use).

- **A long example of a database (the food data from Lecture 2.2) stored as an XML document**, including
  - The schema definition
  - The actual data
  - An XSLT style sheet that produces a report from the food database document consisting of a table of contents, a full food product listing, and an index, all produced by arranging the data in different ways
  - The report produced.

- **More explanation and elaboration of the principles behind XML**

SLecture 6.2a Document Type Definition: Self assessment memo

To: Sue Feldman, CIO
From: Bob Boiko, content management specialist
Subject: Self assessment for year 2000
Date: February 7, 2001
Keywords: Content management; planning; XML; intranet; Web site
URI: www.jasca.com/bboiko/memo20010207-07

Accomplishments in year 2000:

Developed a content management master plan. . . .

Goals for year 2001:

Begin implementation of the content management master plan. . . .

Training needs:

. . .

SGML/XML document type definition (DTD) for self assessment memo

```
<!ENTITY % doctype "selfAssessmentMemo" - document type generic identifier >
<!--        ELEMENTS          MIN   CONTENT (EXCEPTIONS)                  -->
<!ELEMENT   selfAssessmentMemo --    (metadata, memoBody)                  >
<!ELEMENT   metadata          --    (to, from, subject, date, keywords, URL)>
<!ELEMENT   to                -O    (#PCDATA)                             >
<!ELEMENT   from              -O    (#PCDATA)                             >
<!ELEMENT   subject           -O    (#PCDATA)                             >
<!ELEMENT   date              -O    (#PCDATA)                             >
<!ELEMENT   keywords          -O    (#PCDATA)                             >
<!ELEMENT   URL               -O    (#PCDATA)                             >
<!ELEMENT   memoBody          -O    (accomplishments, goals  trainingNeeds)>
<!ELEMENT   accomplishments   -O    (#PCDATA)                             >
<!ELEMENT   goals             -O    (#PCDATA)                             >
<!ELEMENT   trainingNeeds     -O    (#PCDATA)                             >

<!--        ELEMENTS          NAME  VALUE                  DEFAULT-->
<!ATTLIST   selfAssessmentMemo STATUS (confidential | public) confidential>
```

A DTD defines a document structure and identifies each element of the structure by a tag. This DTD creates a **selfAssessmentMemo class**. The documents in the memo class must contain two elements, *metadata* and *memBody*. These, in turn, consist of other elements, as listed in ( ). The elements at the bottom of this tree have a data type, in the examples always #PCDATA, which means a character string. ELements can be required or optional; there sequence can be fixed (as in the example) or fixed. This example does not the various syntactic means to specify these options. The memo also has a **status attribute**, whose default value is *confidential*. Alternatively, the status can be *public*.

Large XML example: A database of foods

Food database schema

```
<?xml version="1.0"?>
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema">

    <xsd:element name="foodDatabase" type="foodDatabaseType"/>

    <xsd:complexType name="foodDatabaseType">
        <xsd:sequence>
            <xsd:element name="foodProduct" type="foodProductType"
            maxOccurs="unbounded"/>
        </xsd:sequence>
    </xsd:complexType>

    <xsd:complexType name="foodProductType">
        <xsd:sequence>
            <xsd:element name="name" type="xsd:string"/>
            <xsd:element name="isa" type="IDREF"/>
            <xsd:element name="origin" type="originType" minOccurs="0"/>
            <xsd:element name="form" type="xsd:string" minOccurs="0"/>
            <xsd:element name="processedBy" type="processedByType"
            minOccurs="0" maxOccurs="unbounded"/>
            <xsd:element name="packedIn" type="xsd:string"minOccurs="0"/>
        </xsd:sequence>
        <xsd:attribute name="foodID" type="ID"  use="required"/>
    </xsd:complexType>

    <xsd:complexType name="originType">
        <xsd:choice>
            <xsd:element name="foodSource" type="xsd:string"/>
            <xsd:element name="part" type="xsd:string" maxOccurs="unbounded"/>
            <xsd:element name="extractedSubstance" type="xsd:string"
            minOccurs="0" maxOccurs="unbounded"/>
            <xsd:element name="madeFrom" type="xsd:string" minOccurs="0"
            maxOccurs="unbounded"/>
            <xsd:element name="ingredient"
            type="ingredientType"maxOccurs="unbounded"/>
        </xsd:choice>
    </xsd:complexType>

    <xsd:complexType name="ingredientType">
        <xsd:attribute name="intensity" type="xsd:string" use="optional"/>
        <xsd:attribute name="purpose" type="xsd:string" use="optional"/>
    </xsd:complexType>

    <xsd:complexType name="processedByType"  mixed="true">
        <xsd:attribute name="purpose" type="xsd:string" use="optional"/>
    </xsd:complexType>
```

Not shown in this example are two options (think back to frames):

(1) Declaring restrictions on an element or attribute (by form, e.g. min and max length of a string, min and max values of numbers, or by an enumerated authority list of allowed values, see text, Section 9.1.1).

(2) Declaring a default value.

See SLecture 5.1 for the RDF schema; it can specify the entity types allowed with a relationship type.

The food data from Lecture 2.2

```
<?xml version="1.0"?>
<?xml:stylesheet type=”text/XSLT”
        xlink:href=”http://www.afw.com/it/database.xslt”?>
<!-- File Name: foodDatabase.xml -->
<!-- Element content shown in italics for clarity -->

<foodDatabase xmlns=”http://www.afw.com/it/database.xsd”>
        <foodProduct foodID=”FP0">
                <name>Food product</name>
        </foodProduct>

        <foodProduct foodID=”FP1">
                <name>Vegetable product</name>
                <isa>FP0</isa>
                <origin>
                        <foodSource>Plant</foodSource>
                </origin>
        </foodProduct>

        <foodProduct foodID=”FP2">
                <name>Meatproduct</name>
                <isa>FP0</isa>
                <origin>
                        <foodSource>Animal</foodSource>
                        <part>Carcass</part>
                </origin>
        </foodProduct>

        <foodProduct foodID=”FP3">
                <name>Egg product</name>
                <isa>FP0</isa>
                <origin>
                        <foodSource>Bird</foodSource>
                        <part>Egg</part>
                </origin>
        </foodProduct>

        <foodProduct foodID=”FP4">
                <name>Prepared food</name>
                <isa>FP0</isa>
                <processedBy>Process</ProcessedBy>
        </foodProduct>
```

```
<foodProduct foodID=”FP5">
        <name>Soup</name>
        <isa>FP4</isa>
        <processedBy>Process</ProcessedBy>
        <form>Liquid or semiliquid</form>
</foodProduct>

<foodProduct  foodID=”FP11">
        <name>Diced carrots</name>
        <isa>FP1</isa>
        <origin>
                <foodSource>Carrot plant</foodSource>
                <part>Root</part>
        </origin>
        <form>Diced</form>
</foodProduct>

<foodProduct foodID=”FP12">
        <name>Cut green beans</name>
        <isa>FP1</isa>
        <origin>
                <foodSource>Bean plant</foodSource>
                <part>Immature fruit</part>
        </origin>
        <form>Cut</form>
</foodProduct>

<foodProduct foodID=”FP13">
        <name>Chicken broth</name>
        <isa>FP2</isa>
        <origin>
                <foodSource>Chicken</foodSource>
                <part>Meat</part>
                <part>Bones</part>
        </origin>
        <extractedSubstance>Fat</extractedSubstance>
        <extractedSubstance>Protein</extractedSubstance>
        <extractedSubstance>Flavor</extractedSubstance>
        <processedBy>Cooking</processedBy>
        <form>Liquid</form>
</foodProduct>

<foodProduct foodID=”FP14">
        <name>Cubed cooked chicken</name>
        <isa>FP2</isa>
        <origin>
                <foodSource>Chicken</foodSource>
                <part>Skeletal meat</part>
        </origin>
        <processedBy>Cooking</processedBy>
        <form>Cubed</form>
</foodProduct>
```

```
<foodProduct  foodID=”FP15">
        <isa>FP3</isa>
        <name>Eggs</name>
        <origin>
                <foodSource>Chicken</foodSource>
                <part>Egg</part>
        </origin>
</foodProduct>

<foodProduct  foodID=”FP16">
        <isa>FP1</isa>
        <name>Durum wheat flower</name>
        <origin>
                <foodSource>Durum wheat</foodSource>
                <part>Seed, kernel</part>
        </origin>
        <form>Ground</form>
</foodProduct>

<foodProduct foodID=”FP17">
        <isa>FP4</isa>
        <name>Noodles</name>
        <origin>
                <ingredient>FP16</ingredient>
                <ingredient>FP15 </ingredient>
        </origin>
        <processedBy>Mixing</processedBy>
        <processedBy>Extruding</processedBy>
        <processedBy>Drying</processedBy>
</foodProduct>

<foodProduct  foodID=”FP18">
        <name>Flavoring</name>
        <isa>FP0</isa>
</foodProduct>

<foodProduct  foodID=”FP19">
        <name>BHT</name>
        <isa>FP0</isa>
</foodProduct>

<foodProduct foodID=”FP20">
        <name>Chicken noodle soup</name>
        <isa>FP5</isa>
        <origin>
                <ingredient>FP13<ingredient>
                <ingredient>FP14<ingredient>
                <ingredient>FP11<ingredient>
                <ingredient>FP12<ingredient>
                <ingredient>FP17<ingredient>
                <ingredient>FP18<ingredient>
                <ingredient purpose=”Preservation“>FP19<ingredient>
        </origin>
```

```
        <processedBy purpose=”Make edible” purpose=” Preservation”>Sterilizing by heat
        </processedBy>
        <form>Liquid with Solid Pieces</form>
</foodProduct>

<foodProduct  foodID=”FP21">
        <name>Diced parsley</name>
        <isa>FP1</isa>
</foodProduct>

<foodProduct foodID=”FP22”>
        <name>Campbell's chicken noodle soup</name>
        <isa>FP20</isa>
        <origin>
                <ingredient> FP13“<ingredient>
                <ingredient> FP14 <ingredient>
                <ingredient> FP11 <ingredient>
                <ingredient> FP12 <ingredient>
                <ingredient> FP22 <ingredient>
                <ingredient> FP17 <ingredient>
                <ingredient> FP18 <ingredient>
                <ingredient purpose=”Preservation“> FP19<ingredient>
        </origin>
        <processedBy purpose=”Make edible” purpose=” Preservation”>Sterilizing by heat
        </processedBy>
        <form>Liquid with Solid Pieces</form>
        <packedIn>Steel can</packedIn>
</foodProduct>

<foodProduct foodID=”FP23”>
        <name>Frozen cut green beans</name>
        <isa>FP12</isa>
        <origin>
                <foodSource>Bean plant</foodSource>
                <part>Immature fruit</part>
        </origin>
        <form>Cut</form>
        <processedBy>Freezing</processedBy>
        <packedIn>Carton</packedIn>
</foodProduct>
```

Style sheet for displaying food data

```
<xsl:stylesheet
  xmlns:xsl=”http://www.w3.org/TR/WD-XSLT”
  xmlns=”http://www.w3.org/TR/REC-html40”>
  <xsl:template match=”/”>
     <HTML>
        <HEAD>
        <TITLE>Food database listing</TITLE>
           <META NAME="Author" CONTENT="Association of Food Wholesalers"/>
           <META NAME="keywords"
              CONTENT="food products; formulation; ingredients"/>
           <META NAME="GENERATOR" CONTENT=""/>
        </HEAD>
        <BODY>
           <H1><Center>Association of Food Wholesalers Product List</Center></H1>
           <H1><Center>Table of Contents</Center></H1>
           <xsl:for-each select=”foodDatabase/foodProduct”>
              <xsl:value-of select=”@foodID“/> 
              <xsl:value-of select=”name“/><BR/>
           </xsl:for-each>

           <H1><Center>Full Food Product Listing<Center></H1>
           <xsl:for-each select=”foodDatabase/foodProduct”>
              <B><xsl:value-of select=”@foodID“/> 
              <xsl:value-of select=”name“/> </B><BR><I>isa <I> 
              <xsl:value-of select=”isa”/><BR/>
              <I>Food source or ingredients:<I> 
              <xsl:value-of select=”origin/foodSource“/>
              <xsl:for-each select=”origin//part”>
                 <xsl:value-of select=”.“/>
              </xsl:for-each>
              <xsl:for-each select=”origin/ingredient“>
                 <xsl:value-of select=”.“/>
                 <xsl:value-of select=”@purpose“/>, 
              </xsl:for-each><BR/>
              <I>Extracted substance:<I> 
              <xsl:for-each select=”origin/extractedSubstance“>
                 <xsl:value-of select=”.“/>, 
              </xsl:for-each><BR/>
              <I>Processsed by:<I> 
              <xsl:for-each select=”processsedBy“>
                 <xsl:value-of select=”.“/>
                 <xsl:value-of select=”@purpose“/>
              </xsl:for-each><BR/>
              <I>Form:<I> 
              <xsl:value-of select=”form“/><BR/>
              <I>Packed in:<I> 
              <xsl:value-of select=”packedIn“/>
              <BR/><BR/>
```

```
            </xsl:for-each><BR/>
```

```
            <H1><Center>Food Naome Index</Center></H1>
            <xsl:for-each select=”foodDatabase/foodProduct” order-by’”+Name”>
                <xsl:value-of select=”name“/> 
                <xsl:value-of select=”@foodID“/><BR>
            </xsl:for-each>
        </BODY>
    </HTML>
</xsl:template>
```

Display of the HTML document produced by the style sheet

Association of Food Wholesalers Product List

Table of Contents

Full Food Product Listing

FP0 Food product

FP1 Vegetable product
isa: FP0
*Food source or ingredients:*Plant

FP2 Meat product
isa: FP0
Food source or ingredients: Animal Carcass

FP3 Egg product
isa: FP0
Food source or ingredients: Bird Egg

FP4 Prepared food
isa: FP0
Processed by: Process

FP5 Soup
isa: FP0
Processed by: Process
Form: Liquid or semiliquid

FP11 Diced carrots
isa: FP1
Food source or ingredients: Carrot plant Root
Extracted substance:
Processed by:
Form: Diced
Packed in:

FP12 Cut green beans
isa: FP1
Food source or ingredients: Bean plant Immature fruit
Extracted substance:
Processed by:
Form: Cut
Packed in:

FP13 Chicken broth
isa: FP2
Food source or ingredients: Chicken Meat Bones
Extracted substance: Fat, Protein, Flavor,
Processed by: Cooking
Form: Liquid
Packed in:

FP14 Cubed cooked chicken
isa: FP2
Food source or ingredients: Chicken Skeletal meat
Extracted substance:
Processed by: Cooking
Form: Cubed
Packed in:

FP15 Eggs
isa: FP3
Food source or ingredients: Chicken Egg
Extracted substance:
Processed by:
Form:
Packed in:

FP16 Durum wheat flower
isa: FP1
Food source or ingredients: Durum Wheat Seed, kernel
Extracted substance:
Processed by:
Form: Ground
Packed in:

FP17 Noodles
isa: FP4
Food source or ingredients: FP16, FP15,
Extracted substance:
Processed by: Mixing, Extruding, Drying
Form:
Packed in:

FP18 Flavoring
isa: FP0
Food source or ingredients:
Form:
Extracted substance:
Processed by:
Packed in:

FP19 BHT
isa: FP0
Food source or ingredients:
Form:
Extracted substance:
Processed by:
Packed in:

FP20 Chicken noodle soup
isa: FP5
Food source or ingredients: FP13, FP14, FP11, FP12, FP17, FP112, FP113 Preservation,
Extracted substance:
Processed by: Sterilizing by heat Make edible, Preservation
Form:
Packed in: Steel can

FP21 Diced parsley
isa: FP1
Food source or ingredients:
Form:
Extracted substance:
Processed by:
Packed in:

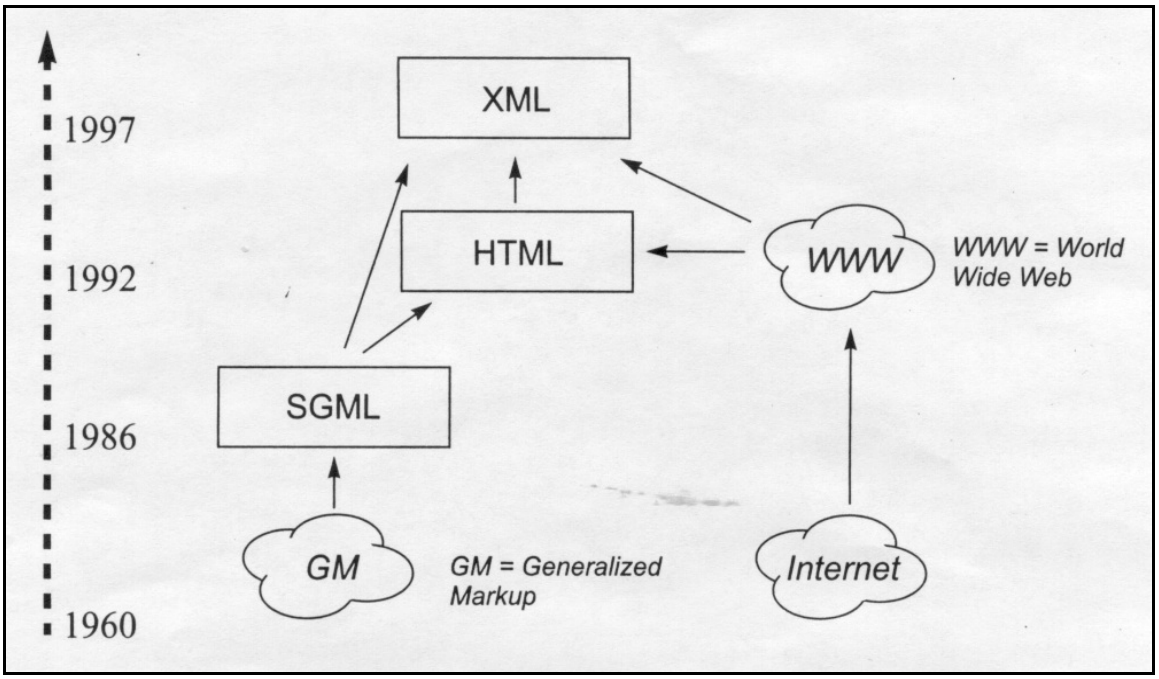
FP22 Campbell's chicken noodle soup
isa: FP20
Food source or ingredients: FP13, FP14, FP11, FP12, FP22, FP17, FP18, FP19 Preservation,
Extracted substance:
Processed by: Sterilizing by heat Make edible, Preservation
Form:
Packed in: Steel can

FP23 Frozen cut green beans
isa: FP12
Food source or ingredients: Bean plant Immature fruit
Extracted substance:
Processed by: Freezing
Form: Cut
Packed in: Carton

Food Name Index

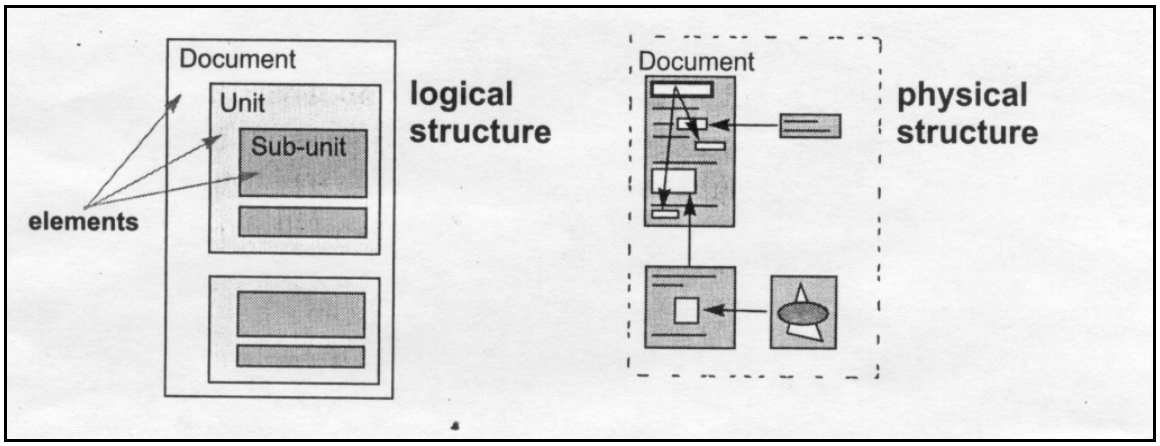
XML in a larger context

History

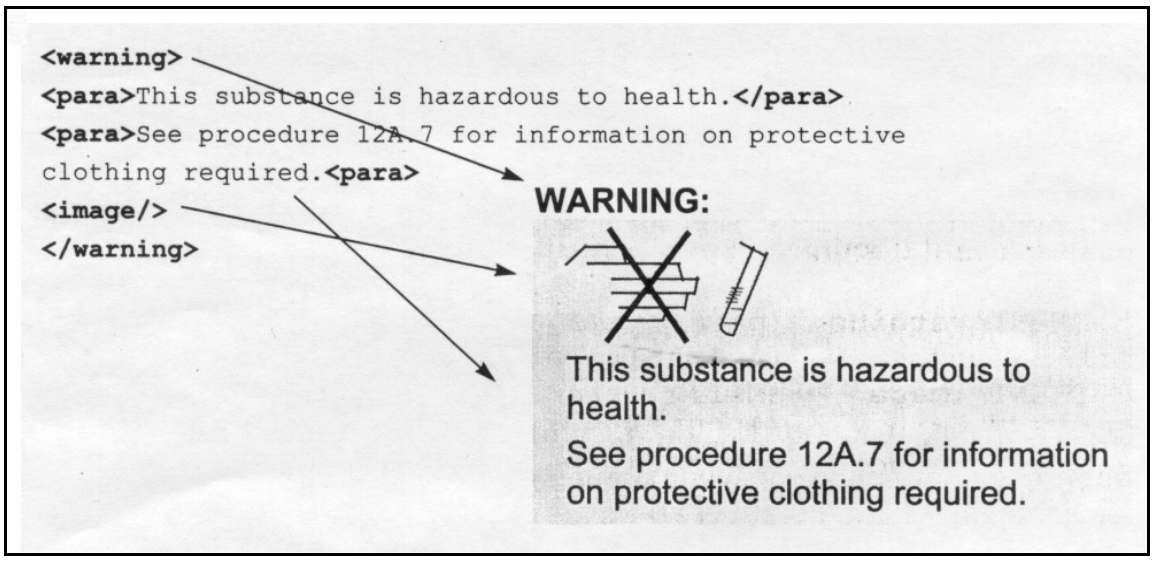


XML: 80% of the functionality, 20% of the complexity of SGML

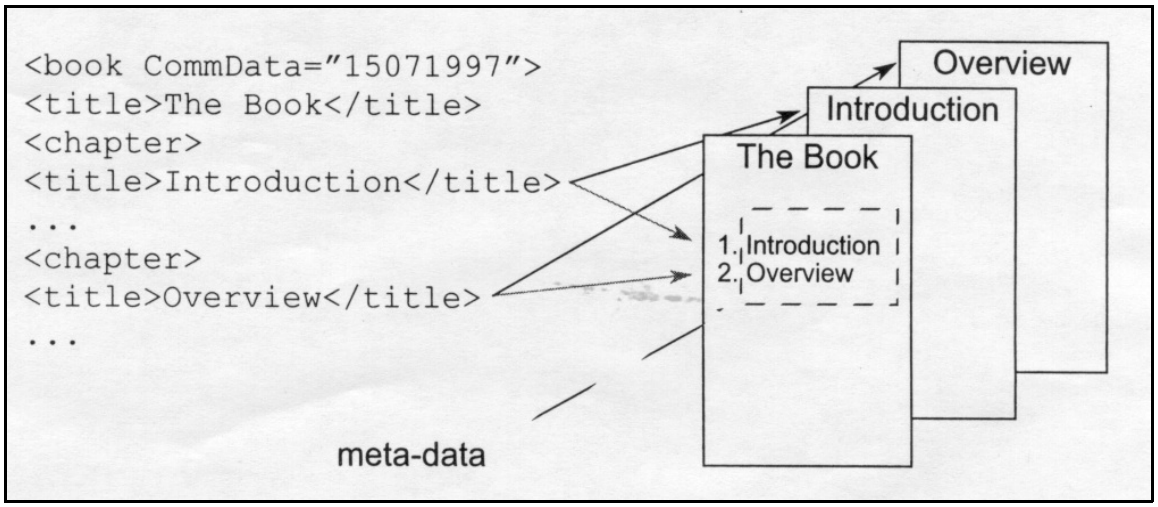
Document management: Logical and physical structure



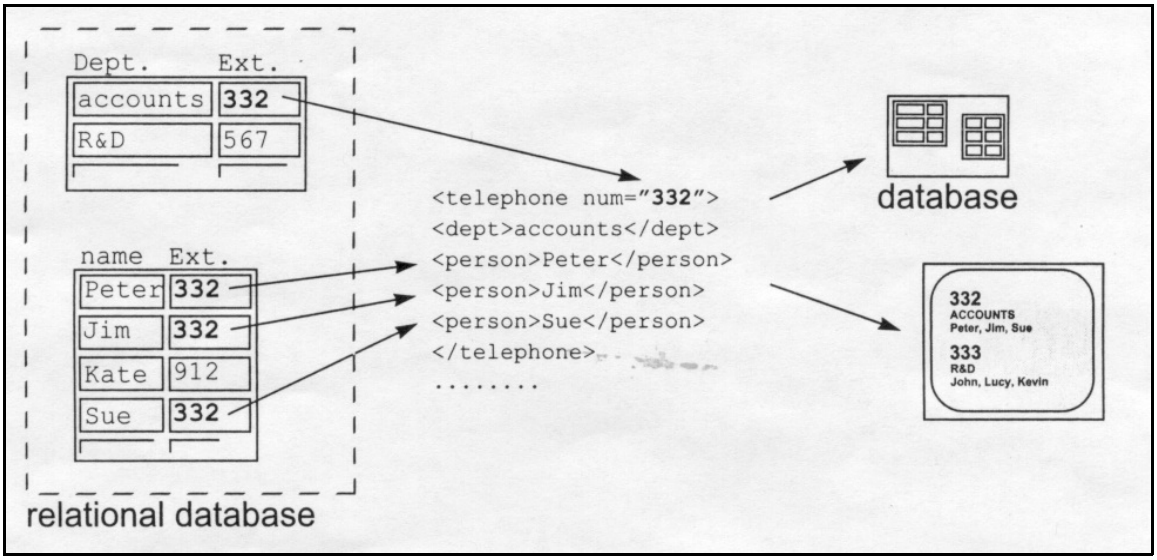
From XML to display (style sheets are used to govern this mapping)



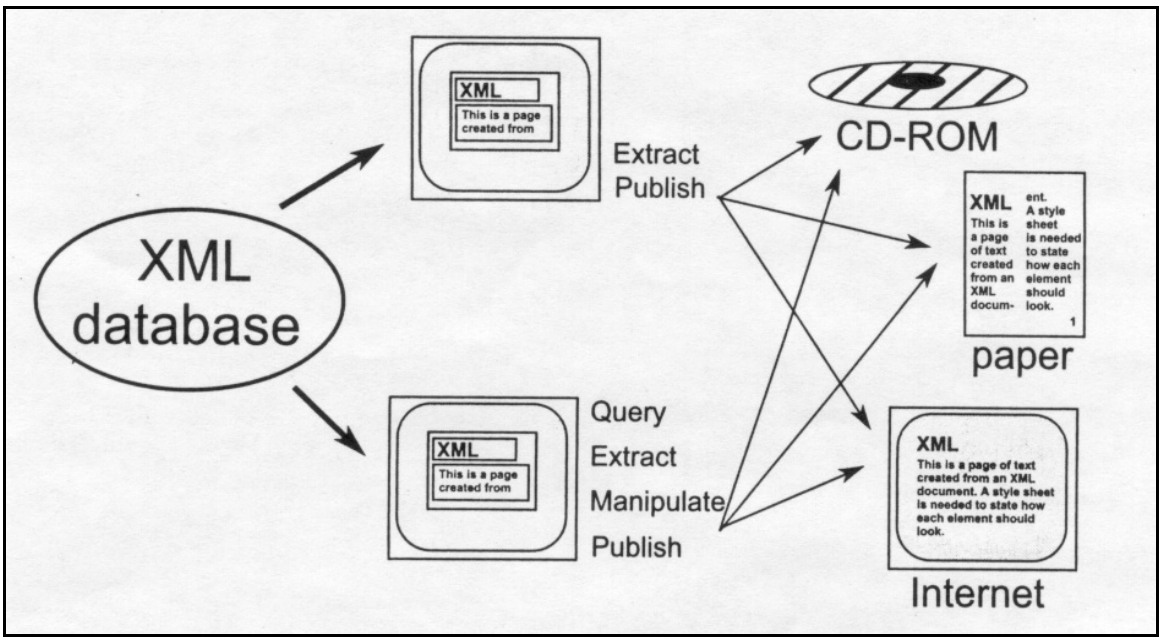
From XML to display: Multiple use of data (also: extraction of cataloging data)



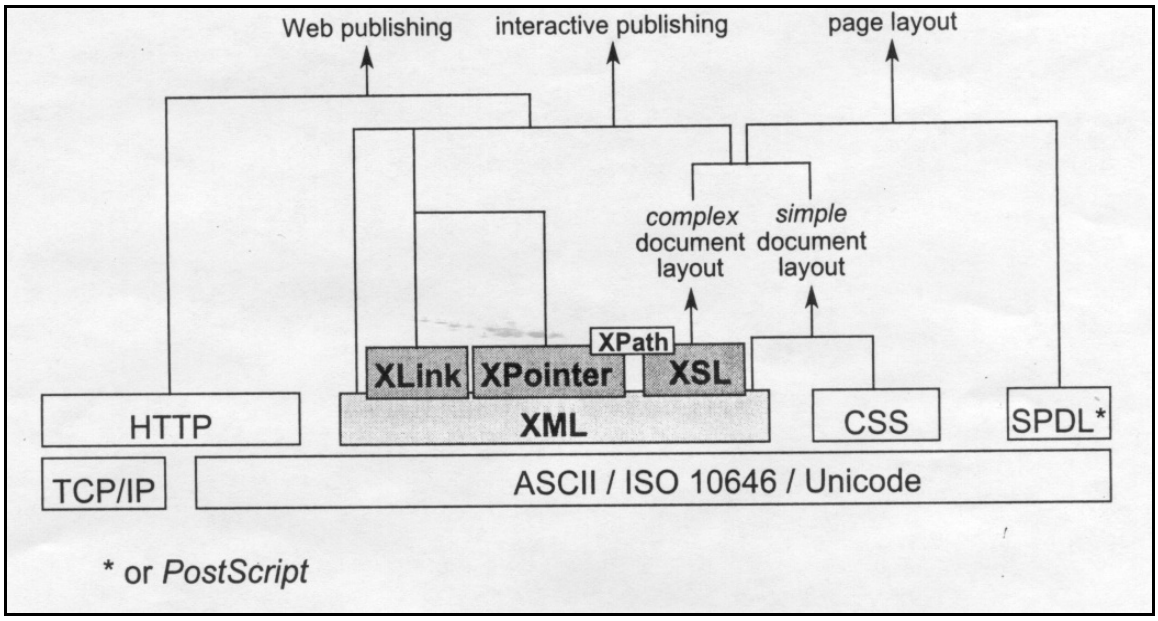
From database to XML, from XML to display or a different database

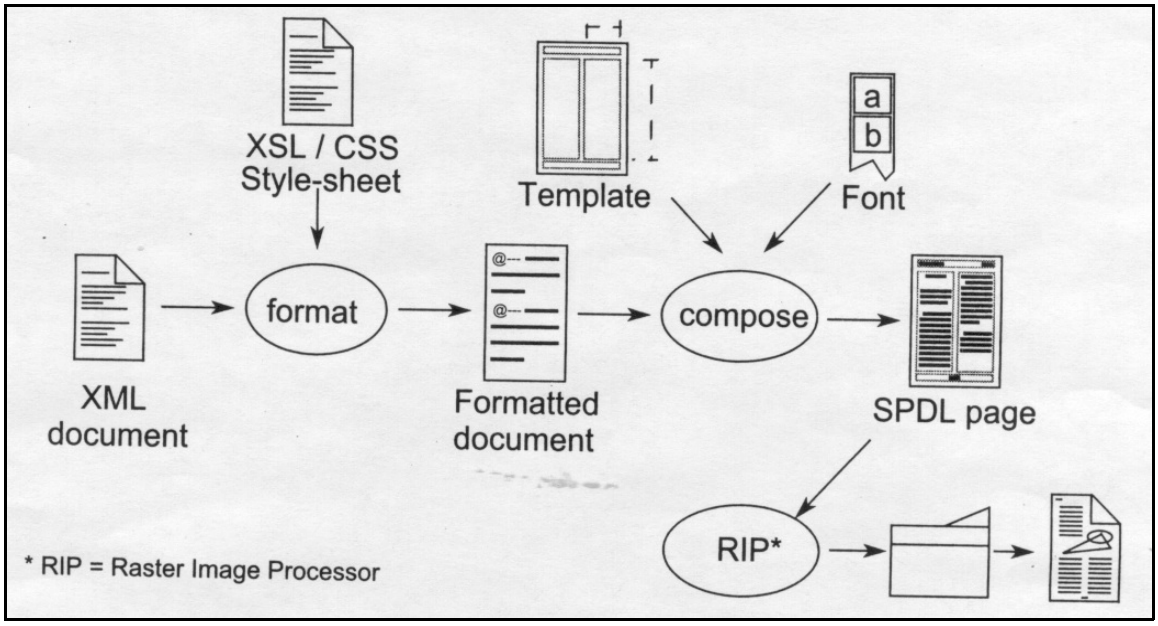


XML database for multiple uses through simple extraction or more complex manipulation



Alphabet soup: Many standards working together





Transfer protocols

TCP/IP	Transfer Control Protocol / Internet Protocol
HTTP	Hypertext Transfer Protocol

Character code standards

ASCII	American Standard Code for Information Interchange common characters used in English. 7 bit (extended ASCII 8 bit). Almost identical to ISO/IEC 646:1991 (7 bit, 128 characters) ISO International Standards Organization IEC International Electrotechnical Commission
ISO/IEC8859/1	8 bit, 256 characters. Superset of ASCII / ISO/IEC646:1991. Adds accented characters and such. Used in HTML and Windows. One of family of 8 bit codes
Unicode	16-bit superset of ISO/IEC8859/1. Extended character sets for many languages. 64,000 characters
ISO/IEC 10646	32-bit superset of Unicode, 4 billion characters

Formatting standards

SGML	Standard Generalized Markup Language
HTML	Hypertext Markup Language, an SGML application
XML	eXtensible Markup Language
MathML	Mathematical Markup Language, an XML application
SMIL	Synchronized Multimedia Integration Language, an XML application
CSS	Cascading Style Sheets (for use with HTML)
XSL	XML Stylesheet Language
XSLT	XSL Transformation. A standard for transforming any XML document into a specific XML format suitable for output according to a style sheet
SPDL	Standard Page Description Language
PostScript	ADOBE's proprietary but widely used page description language
XLink	XML Linking Language (more complex linking than HTML), an XML application
XPointer	Complementary to XLink, allows linking to a specific position within a document

XML related query language standards

XQL	eXtensible Query Language - a general query language for XML documents and structured data
XPath	An expression language for querying XML documents and data structures
SPARQL	SPARQL Protocol And RDF Query Language), a query language for RDF triples, Linked Data

Image format standards

JPEG	Joint Photographic Experts Group 24-bit raster format for image compression, preferred when high resolution is needed
GIF	Graphic Interchange Format, 8-bit raster format, preferred on the Web for simple images
PNG	Portable Network Graphics. Developed for the Internet, consistent appearance across platforms, my replace GIF.
MPEG	Moving Picture Experts Group. Developed a number of standards for the physical storage and logical description of moving images

General Models and frameworks

RDF	Resource Description Framework. A framework for the representation of metadata, uses an entity-relationship approach
DOM	Document Object Model. A software interface standard for accessing and manipulating document elements as objects

Lecture 6.2b Text analysis overview and examples. Supplement

SLecture 6.2b In-class exercises and examples illustrating the importance of text analysis through several linguistic techniques

2 Extracting data through slot-filling in frames: examples

Extracting data from pesticide reports

Pesticide frame

Slot	Instructions: What to look for to find slot fillers
Substance	a term that designates a substance
Pest fought	the name of an organism that can be a pest or the name of a disease
Crop or livestock	the name of a useful plant or animal
When applied	the name of a season or a term indicating weather condition
Dosage	a symbol for mass, such as *pound*, *g*, *kg* and the number preceding. Also look for *per* or *for each*
Route of administration	a term such as *spray*, *work into the soil*

3 Exacting data from text, especially importance of **resolving anaphoric references**

Contact Dermatitis-Irritant and Allergic

Contact dermatitis may result from irritants or substances to which an individual has become allergic. Depending upon the source of irritation, the duration or frequency of exposure, and other variables, different uncomfortable changes in the skin occur.

Irritant contact dermatitis occurs when the skin is exposed to a mild irritant-such as detergents or solvents-repeatedly over a long period of time or to a strong irritant, such as acid or alkali, which can cause immediate damage to the skin.

This disorder is an "occupational hazard" for housewives, chemical workers, doctors and dentists, restaurant workers, and others whose work brings them into regular or prolonged contact with **soaps, detergents, chemicals, and abrasives**. *These substances* either erode the protective oily barrier of the skin or injure its surface.

Allergic dermatitis occurs when skin which has been sensitized to a specific substance comes in contact with that substance again. With the exception of poison ivy and poison oak, to which about 70 percent of people become sensitized after first contact, most contact allergies produce sensitivity in only a few people. The most common of these allergies are nickel and other metals, rubber and elasticized garments, dyes, cosmetics (especially nail polish), and leather. But anyone can become sensitized to almost anything, so the search for the offending substance is often tedious and success is sometimes elusive.

In **irritant dermatitis** the **skin** becomes stiff, dry, and tight-feeling. *It* may crack, blister, or become ulcerated. Some itching may accompany mild inflammation, but the fissures and ulcers will be painful, not itchy. Mild irritants cause a progression from reddening and blistering to drying and cracking, while strong irritants cause blistering on contact and then erosion and ulcers.

Allergic dermatitis appears as reddening, followed by blistering and oozing. *In severe cases* there may be swelling of the face, eyes, and genital area. The rash will appear wherever the allergen has touched the skin, either directly or by transference from the hands. However, the palms, soles, and scalp seldom show any reaction. Fluid from the blisters will not spread the disease to other parts of the body or to other people.

There are no tests to determine the cause of **irritant dermatitis**. Finding the source may require persistent and creative detective work on the part of both doctor and patient. Patch tests can often determine or point the way to the allergens responsible for the reaction in **allergic dermatitis**. It may, however, take some sleuthing to find the specific product or products which contain the offending substance.

Preventive measures for irritant dermatitis are easy to define and difficult to carry out. The disease is usually the direct result of the working environment, and adequate protective measures are often impractical, if not impossible, to achieve. To the extent possible, then, it is recommended that the patient take the following precautions:

1. Wear cotton gloves under rubber gloves for all wet work. If gloves are impractical, use a barrier cream to protect the skin. Reapply the cream 2 or 3 times per day and after each handwashing.

Finally, consider this text:

Leukemia

Acute Lymphocytic Leukemia (ALL) and **Chronic Myelogenous Leukemia** (CML) occur in different populations with different symptoms. *The former* primarily affects children under age 5, who often show signs of anemia, fatigue, fever, and bleeding, indicating a depressed functioning of the bone marrow. *The latter* occurs primarily in men between 20 and 50, with symptoms varying from none at all to anemia and general malaise to weight loss, night sweats, fatigue, and an enlarged spleen that may cause discomfort on the left side of the abdomen. *The disease* can develop gradually, almost insidiously. The number of granulocytes is markedly increased, . . .

Application to searching (advanced exploration)

Try searching for some of the noun phrases from example 2 in Google. Just type them in without using quotes. In all cases, a large proportion of the top 100 documents (Web sites, but in Google Scholar also articles) found have the noun phrase in them. So Google must have some mechanism for searching phrases; it may be as simple as giving a document a higher score if the search words are close together.

Sequence also seems to matter. *library school* gets results about evenly divided between library schools and school library (school at all levels, not just K-12, the meaning of school in the phrase *school libraries*)

Try *peer pressure*, *pressure by peers*, and *pressured by peers*
The first two find very similar Web sites, the last finds additional relevant sites

Try *social pressure*
Look-ahead note: While all of the *peer pressure* Web sites are relevant, only a few are found

A system could use noun phrases to disambiguate homonymous and polysemous words, so it would know whether *pressure* means *physical pressure* (as in *vapor pressure*, *water pressure*, *barometric pressure*) and when it means "*mental pressure"* (as in *peer pressure*, *parental pressure*, *social pressure*). Then the user could search for these general concepts, whereas in Google a search for *pressure* returns everything.

SLecture 6.2b Natural language processing (NLP) achieves the purposes listed in *Practical significance* through several techniques

Identifying noun phrases	(in all their variant forms) in document texts and in query statement texts as good indexing terms and search terms, respectively. (Some search engines look for noun phrases in the string of words entered into the query box and rank documents with the noun phrase higher than documents that just have the individual words.) Note the difficulty posed by situations like *information retrieval*, *retrieval of information*, *retrieval of legal information*; looking simply for the string *information retrieval* will give incomplete results. But that is what the above-mentioned search engines most likely do, because the alternatives are (1) still costly syntactic processing of all Web page texts or (2) using proximity operators, which is less precise.
Complete or partial sentence parsing	Note: Emphasis is not so much on the role of a parser identifying a string of words as a well-formed sentence. What really matters is: • identifying the role of each word or group of words in the sentence, which is the basis for determining part of speech of a word (is man used as a noun or a verb?), • identifying noun phrases, • semantic parsing For purposes of simply "understanding" the text, it is even useful if the system can deal with sentences that are not well-formed; in this context, checking for grammaticality is important only insofar as it supports understanding, especially through disambiguation.
Semantic parsing	Disambiguating homonyms, word sense disambiguation (WSD)
Statistical NLP methods	Increasingly used for several functions, replacing or working in combination with formal syntax methods • part-of-speech tagging • summarization • automatic translation • automatic speech recognition

Statistical and formal methods	As we discussed, both statistical analysis and syntactic analysis are used for NLP. Systems differ in the degree to which they rely on these two approaches. All of the purposes listed below are amenable to either approach; automatic summarization of single documents is usually done statistically, multi-document summarization systems and information extraction systems often use at least some syntactic and semantic processing.
Multiple languages	The methods discussed can be applied to any language; of course, each language needs its own syntax and semantics knowledge base. Statistical systems may process a multilingual collection; syntactic-semantic systems usually deal with one language at a time. One could put together many such systems into one package, with a program that can recognize the language of a document sending incoming documents to the appropriate language-specific program.

Examples of statistics-based and NLP-based summarizers
Overview: http://itt.nissat.tripod.com/itt0202/ruoi0202.htm www.copernic.com/en/products/summarizer/ The MS Word AutoSummarize function on the Tools menu http://domino.research.ibm.com/cambridge/research.nsf/0/74c0a77cbfad5ae585256bf80054b036?OpenDocument

`Example NLP tools, including parsers
This site has many links to NLP tools, nicely classified http://www-a2k.is.tokushima-u.ac.jp/member/kita/NLP/nlp_tools.html

This lecture uses **transition network diagrams** as an example to illustrate parsing. These diagrams are intended as the blueprint for a computer program that could process a document one sentence at a time. Inter-sentence relationships, such as anaphoric reference, would have to be detected in a second phase. We will start with the analysis of noun phrases and then move to simple sentences. A full parsing system would be orders of magnitude more complex.

In-class exercise in parsing: Identification of noun phrases for indexing

Note: The examples are not meant to present the latest approaches to syntactic and semantic parsing nor should the reader memorize specific steps.
`

P. ~41- 49 The parsing game (take these pages out of your binder)

P. ~51 - 67 More detail about the syntactical analysis (look at together with the parsing game)

Note: Audio for this is in preparation, ask if you are interested

571 Soergel

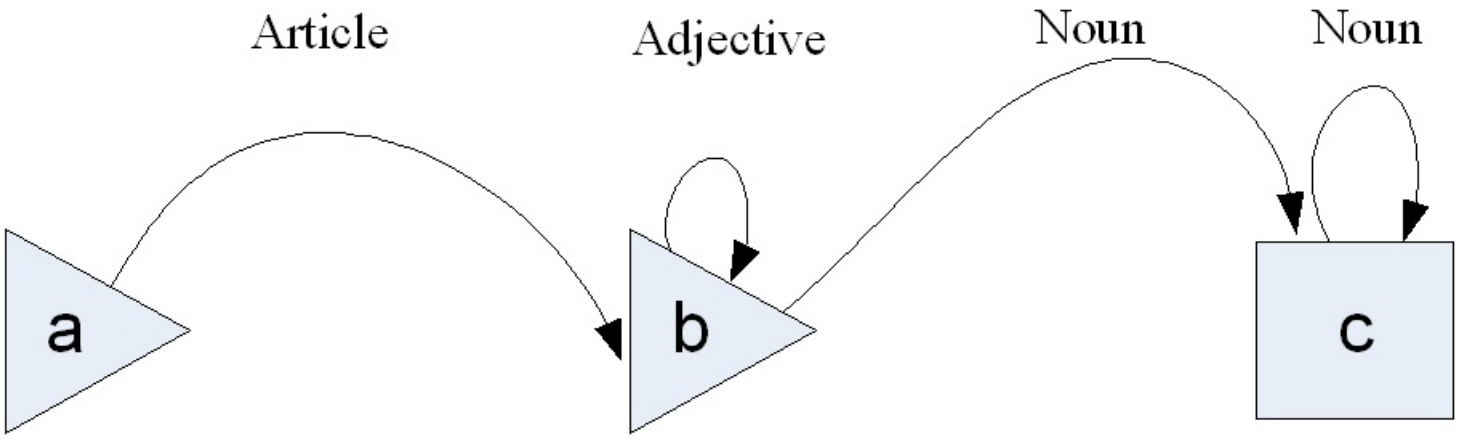
The parsing game

To start, put game piece on a triangle.

Move game piece along the arc corresponding to the next work in the string of words, cross off the word
If you cannot move and there are still words left, you loose.

If you arrive at a square and no words are left, you win.

Simple transition network diagram: noun phrase



Sample noun phrases (by general linguistic convention, * means syntactically incorrect)

1 $_0$ the $_1$ dishwasher $_2$

2 $_0$ the $_1$ jolly $_2$ dishwasher $_3$

3 $_0$ the $_1$ jolly $_2$ white $_3$ dishwasher $_4$

4 $_0$ bones $_1$

5 $_0$ regular $_1$ daily $_2$ consumption $_3$

6 *$_0$ daily $_1$ consumption $_2$ regular $_3$

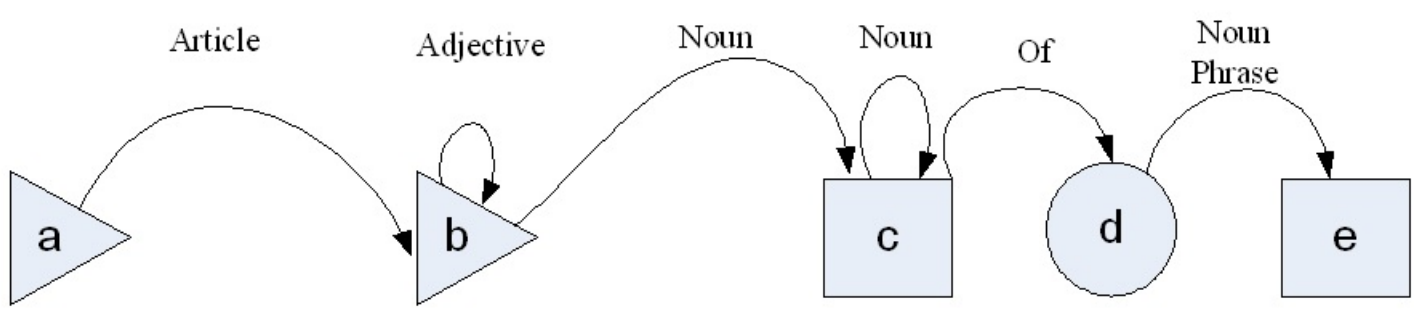
7 $_0$ bone $_1$ mass $_2$

8 $_0$ the $_1$ calcium $_2$ supply $_3$

9 *$_0$ supply $_1$ calcium $_2$

10 $_0$ a $_1$ deficient $_2$ calcium $_3$ supply $_4$

Noun phrase network (1)



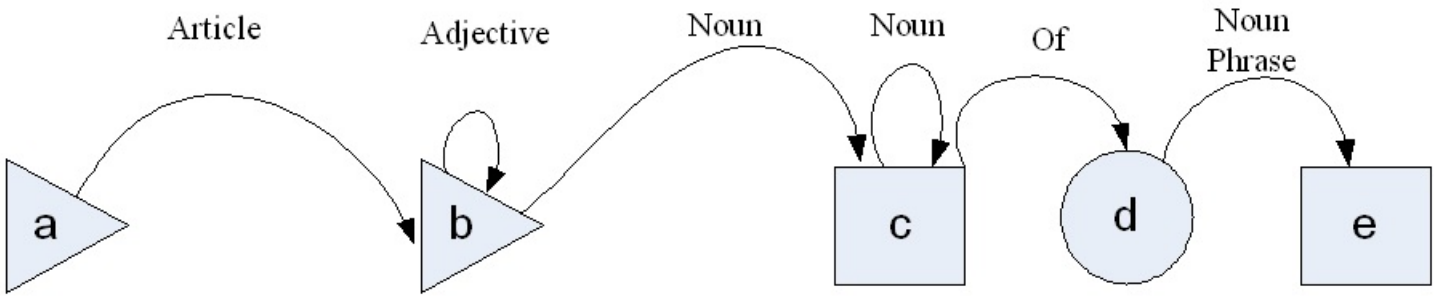
Sample noun phrases (by general linguistic convention, * means syntactically incorrect)

1 $_0$ the $_1$ main $_2$ source $_3$ of $_4$ calcium $_5$

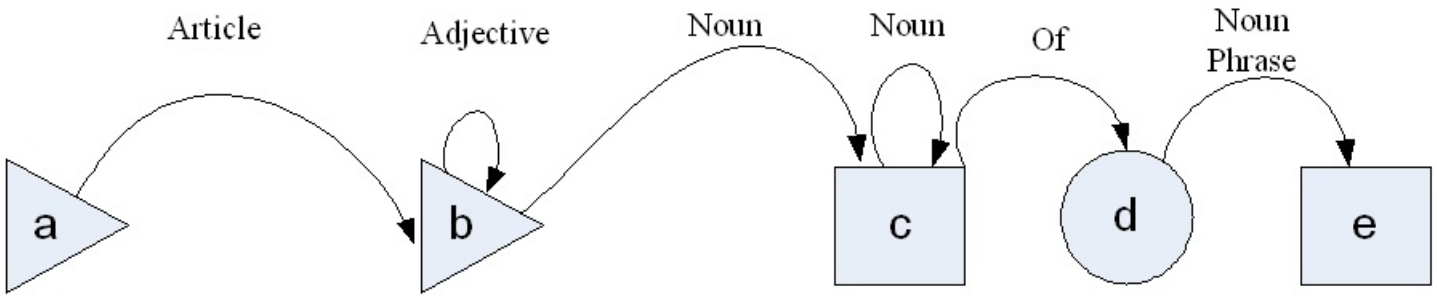
2 $_0$ the $_1$ growing $_2$ skeleton $_3$ parts $_4$ of $_5$ healthy $_6$ small $_7$ children $_8$

3 $_0$ the $_1$ growing $_2$ skeleton $_3$ parts $_4$ of $_5$ healthy $_6$ small $_7$ children $_8$ of $_9$ healthy $_{10}$ parents $_{11}$

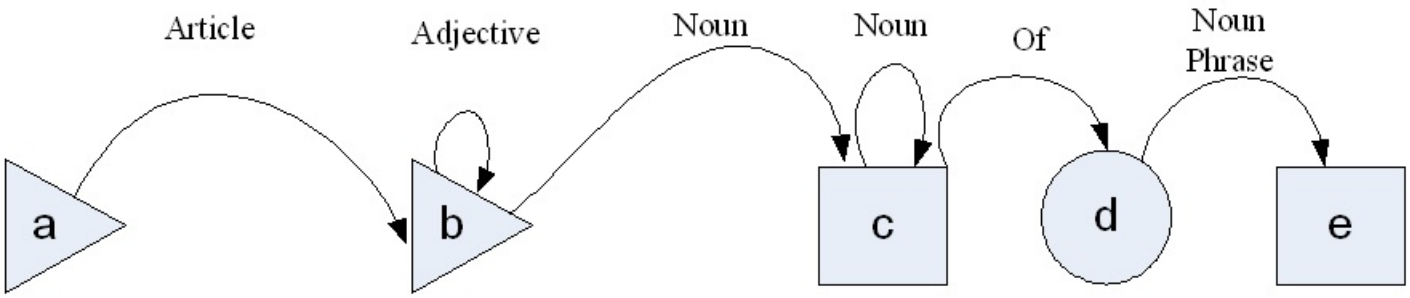
Noun phrase network (2)



Noun phrase network (3)



Noun phrase network (1)



OSTEOPOROSIS

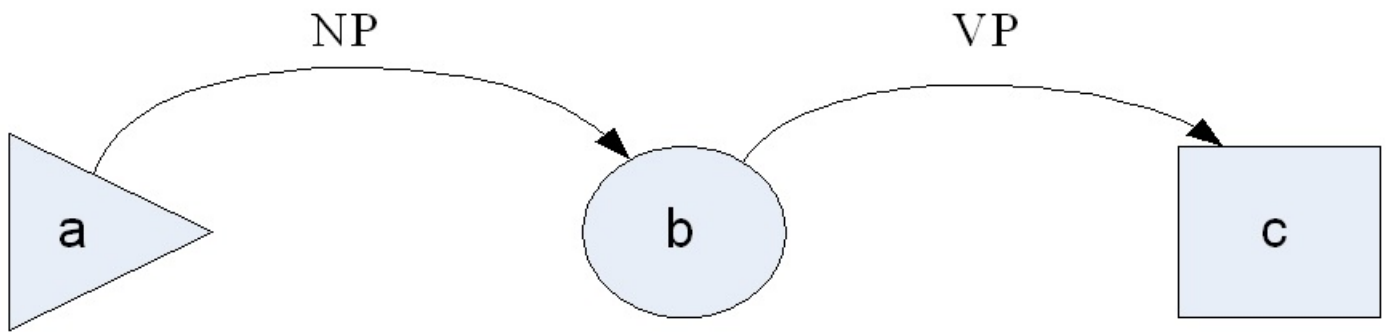
BONES NEED CALCIUM to maintain their strength, hardness, and to stay healthy. Milk, the main source of calcium in the diet, is important for the growing skeletons of children and adolescents as well as the bone-forming cells of adults. Regular daily consumption of at least 1 cup of skim or low-fat milk is essential for adults who want to keep their bones strong and to help prevent osteoporosis, a disease in which the body's bone mass decreases and bones become thin and brittle. Bones weakened by osteoporosis, a disease common to postmenopausal women, are prone to fracture if a person falls.

When calcium enters the body, it is absorbed into the bloodstream. If there is any excess, it is deposited in the end of the bone shafts where it is stored until the body needs to tap this reserve. (Some is also excreted via the kidneys.) When the calcium supply is deficient, the blood must take it back from the bones. If calcium intake remains inadequate over a long period of time, the bones eventually become porous and weak.

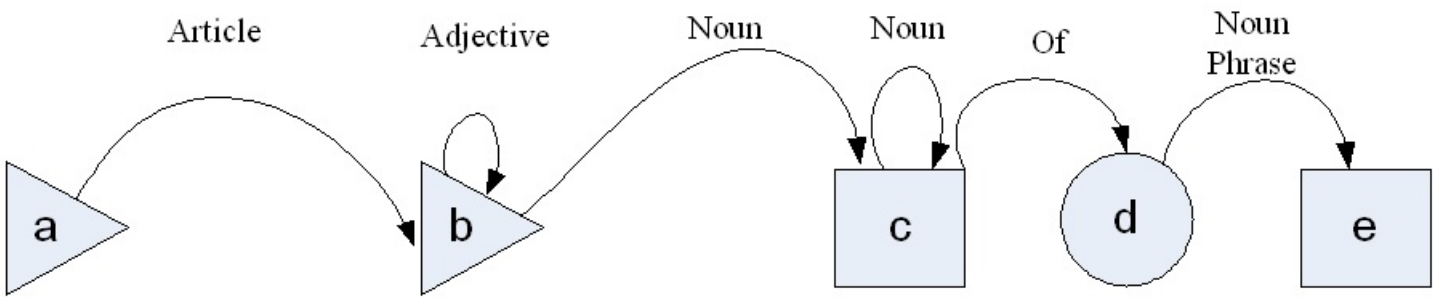
It is not known why calcium loss occurs. That postmenopausal women tend to get osteoporosis points in the direction of a hormonal disorder as estrogen in women of this age falls off sharply. Estrogen therapy is one treatment but its ability to decrease calcium loss may last only several years. Increased calcium intake and exercise are other therapies. The links between lack of exercise and osteoporosis are becoming firmer as research into the causes of this disease progresses.

The disease most frequently affects the spinal column, causing backaches and rounded shoulders. in severe cases, the bone becomes as porous as a sponge and can collapse as a result. Collapsing **vertebrae**, which can cause sudden and sharp backaches, is one reason why elderly people tend to get shorter.

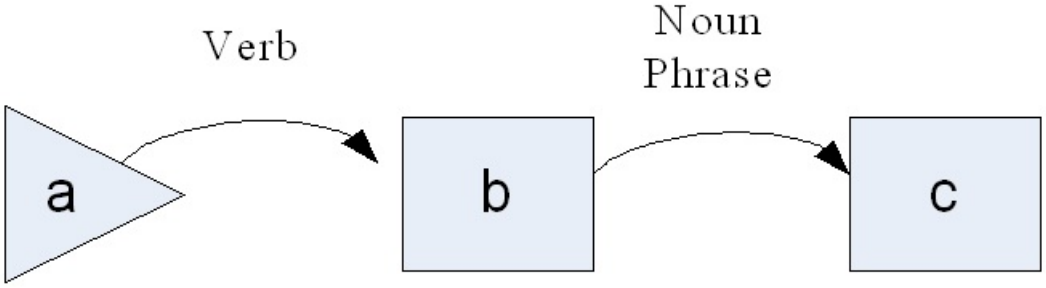
Sentence network



Noun phrase network (1)



Verb phrase network



1 **$_0$ The $_1$ green $_2$ vegetables $_3$ supply $_4$ calcium $_5$.**

2 The green vegetables supply calcium to the body. [Not recognized by our simplistic parser.]

3 The green vegetables supply digestible calcium.

4 The green vegetables supply determines sufficiency of calcium.

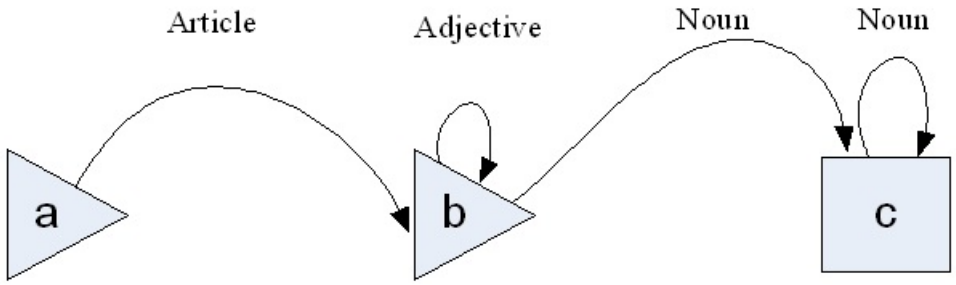
End of parsing game

Begin of more formal treatment of syntax and parsing

Look at together with the parsing game

Simple transition network diagram: noun phrase

NP



▷ possible start state □ possible end state

(Note that there are two possible starting points; if the first does not work, try the second)

Dictionary

a ART	dishwasher N
bone N	jolly ADJ
bones N	mass N
calcium N	regular ADJ
consumption N	supply N
daily ADJ	the ART
deficient ADJ	white ADJ

Sample noun phrases (by general linguistic convention, * means syntactically incorrect)

1 $_0$ the $_1$ dishwasher $_2$

2 $_0$ the $_1$ jolly $_2$ dishwasher $_3$

3 $_0$ the $_1$ jolly $_2$ white $_3$ dishwasher $_4$

4 $_0$ bones $_1$

5 $_0$ regular $_1$ daily $_2$ consumption $_3$

6 *$_0$ daily $_1$ consumption $_2$ regular $_3$

7 $_0$ bone $_1$ mass $_2$

8 $_0$ the $_1$ calcium $_2$ supply $_3$

9 *$_0$ supply $_1$ calcium $_2$

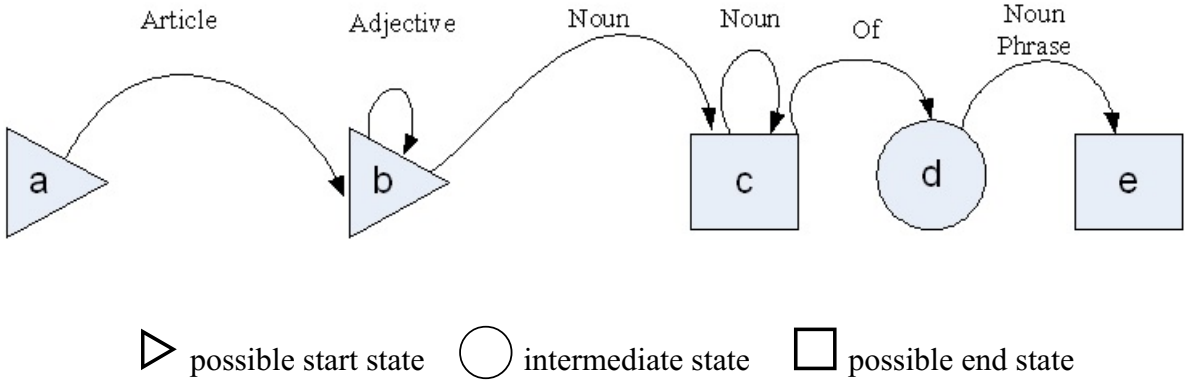
10 $_0$ a $_1$ deficient $_2$ calcium $_3$ supply

Step-by-step trace of the parsing process

From pos	From state	Arc tried	segment (word) processed	To state	To pos	Comment
	$_0$ the $_1$ dishwasher $_2$					
0	a	ART	the	b	1	
1	b	NOUN	dishwasher	**c**	**2**	end state, all words used = **success**
	$_0$ the $_1$ jolly $_2$ dishwasher $_3$					
0	a	ART	the	b	1	
1	b	ADJ	jolly	b	2	
2	b	NOUN	dishwasher	**c**	**3**	**success**
	$_0$ the $_1$ jolly $_2$ white $_3$ dishwasher $_4$					
0	a	ART	the	b	1	
1	b	ADJ	jolly	b	2	
2	b	ADJ	white	b	3	
3	b	NOUN	dishwasher	**c**	**4**	**success**
	$_0$ regular $_1$ daily $_2$ consumption $_3$					
0	a	ART	~~regular~~	a	0	Try next possible start state, namely b.
0	b	ADJ	regular	b	1	
1	b	ADJ	daily	b	2	
2	b	NOUN	consumption	**c**	**3**	**success**
	***$_0$ daily $_1$ consumption $_2$ regular $_3$**					Not a good noun phrase
0	a	ART	~~daily~~	a	0	No arc to follow
0	b	ADJ	daily	b	1	
1	b	NOUN	consumption	c	2	
2	c		regular	**c**	2	No arc to follow, **failure**

Complex transition network diagram: noun phrase

NP



(Note that there are two possible starting points; if the first does not work, try the second)

Dictionary	
a ART	main ADJ
bone N	mass N
bones N	of PREP
calcium N	parents N
children N	parts N
consumption N	regular ADJ
daily ADJ	skeleton N
deficient ADJ	small ADJ
dishwasher N	source N
growing ADJ	supply N
healthy ADJ	the ART
jolly ADJ	white ADJ

Sample noun phrases (by general linguistic convention, * means syntactically incorrect)

1 $_0$ the $_1$ main $_2$ source $_3$ of $_4$ calcium $_5$

2 $_0$ the $_1$ growing $_2$ skeleton $_3$ parts $_4$ of $_5$ healthy $_6$ small $_7$ children $_8$

3 $_0$ the $_1$ growing $_2$ skeleton $_3$ parts $_4$ of $_5$ healthy $_6$ small $_7$ children $_8$ of $_9$ healthy $_{10}$ parents $_{11}$

From pos	From state	Arc tried	segment (word) processed	To state	To pos	comment
	$_0$ the $_1$ main $_2$ source $_3$ of $_4$ calcium $_5$					
0	a	ART	the	b	1	
1	b	ADJ	main	b	2	
2	b	NOUN	source	c	3	
3	c	OF	of	d	4	
4	d	NP	calcium	**e**	**5**	NP network called again, single noun is a noun phrase **success**

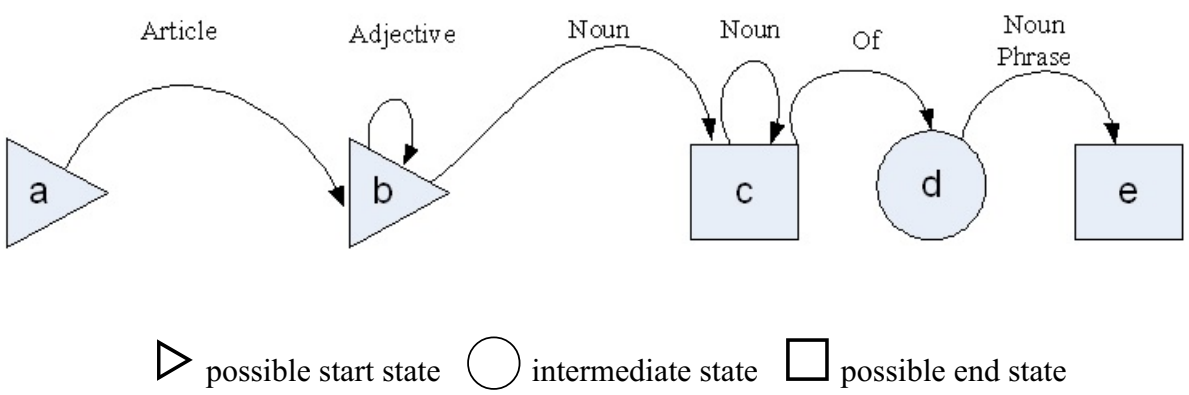
From pos	From state	Arc tried	segment (word) processed	To state	To pos	comment
	$_0$ the $_1$ growing $_2$ skeleton $_3$ parts $_4$ of $_5$ healthy $_6$ small $_7$ children $_8$					
0	a	ART	the	b	1	
1	b	ADJ	growing	b	2	
2	b	NOUN	skeleton	c	3	
3	c	NOUN	parts	c	4	
4	c	OF	of	d	5	
5	d	NP	healthy small children	**e**	**8**	NP network called again, this sequence is a noun phrase **success**

Note: These two examples give a first inkling of nesting transition network diagrams. Here we use the NP diagram to process a sequence of words inside a noun phrase that is itself being analyzed with a NP diagram. Here this nesting is treated very informally; examples to follow will demonstrate the exact process.

Identification of noun phrases for indexing, continued

Complex transition network: noun phrase

NP



Dictionary

a ART
adolescents N
adults N
blood N
bloodstream N
body N
bone N
bone-forming ADJ
bones N
brittle ADJ
calcium N
children N
common ADJ
consumption N
cup N
daily ADJ
deficient ADJ
diet N
disease N
dishwasher N
essential ADJ
excess N
fracture N
growing ADJ
hardness N
healthy ADJ
important ADJ
inadequate ADJ
intake N
jolly ADJ
kidneys N
low-fat ADJ
main ADJ
mass N
milk N
need V N
osteoporosis N
person N
postmenopausal ADJ
prone ADJ
regular ADJ
reserve V N
shafts N
skeletons N
source N
strength N
strong ADJ
supply V N
the ART
thin ADJ
weakened ADJ
white ADJ
women N

Apply the complex transition network and the enlarged dictionary to the identification of noun phrases in the following text.

OSTEOPOROSIS

BONES NEED CALCIUM to maintain their strength, hardness, and to stay healthy. Milk, the main source of calcium in the diet, is important for the growing skeletons of children and adolescents as well as the bone-forming cells of adults. Regular daily consumption of at least 1 cup of skim or low-fat milk is essential for adults who want to keep their bones strong and to help prevent osteoporosis, a disease in which the body's bone mass decreases and bones become thin and brittle. Bones weakened by osteoporosis, a disease common to postmenopausal women, are prone to fracture if a person falls.

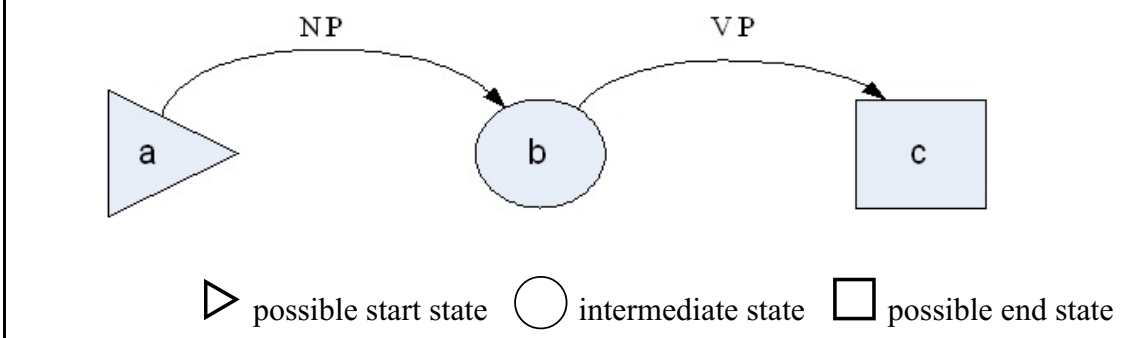
When calcium enters the body, it is absorbed into the bloodstream. If there is any excess, it is deposited in the end of the bone shafts where it is stored until the body needs to tap this reserve. (Some is also excreted via the kidneys.) When the calcium supply is deficient, the blood must take it back from the bones. If calcium intake remains inadequate over a long period of time, the bones eventually become porous and weak.

It is not known why calcium loss occurs. That postmenopausal women tend to get osteoporosis points in the direction of a hormonal disorder as estrogen in women of this age falls off sharply. Estrogen therapy is one treatment but its ability to decrease calcium loss may last only several years. Increased calcium intake and exercise are other therapies. The links between lack of exercise and osteoporosis are becoming firmer as research into the causes of this disease progresses.

The disease most frequently affects the spinal column, causing backaches and rounded shoulders. in severe cases, the bone becomes as porous as a sponge and can collapse as a result. Collapsing **vertebrae**, which can cause sudden and sharp backaches, is one reason why elderly people tend to get shorter.

Parsing of sentences: The sentence network outlines a grammar for simple sentences.

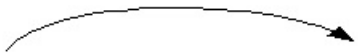
Sentence network



S

NP

means: apply the noun phrase parse transition network



Dictionary	
body N	sufficiency N
calcium N	supply V, N
determines V	the ART
digestible ADJ	to PREP
green ADJ	vegetables N

Sentences

1 $_0$ **The** $_1$ **green** $_2$ **vegetables** $_3$ **supply** $_4$ **calcium** $_5$**.**
2 The green vegetables supply calcium to the body. [Not recognized by our simplistic parser.]
3 The green vegetables supply digestible calcium.
4 The green vegetables supply determines sufficiency of calcium.

Trace of a sentence parse

$_0$ The $_1$ green $_2$ vegetables $_3$ supply $_4$ calcium. $_5$

	From pos	From state	Segment processed	To state	To pos
	0	**S^0 a**	**? (consult NP)**	**?**	**?**
			Magic. Result:		
	0	**S^0 a**	**the green vegetables**	**S^0 b**	**3**
	3	**S^0 b**	**? (consult VP)**	**?**	**?**
			Magic. Result:		
	3	**S^0 b**	**supply calcium**	**S^0 c**	**5**
				Success: End state of S, end of word list	

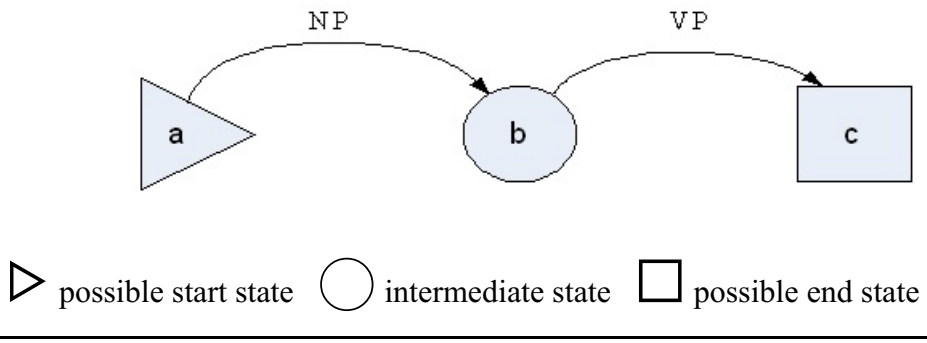
Result: An analysis of the sentence structure, a sentence diagram.

{*S*
[*NP* the green vegetables]
[*VP* supply calcium**]**
}

Parsing of sentences: The three transition networks define a grammar for simple sentences.

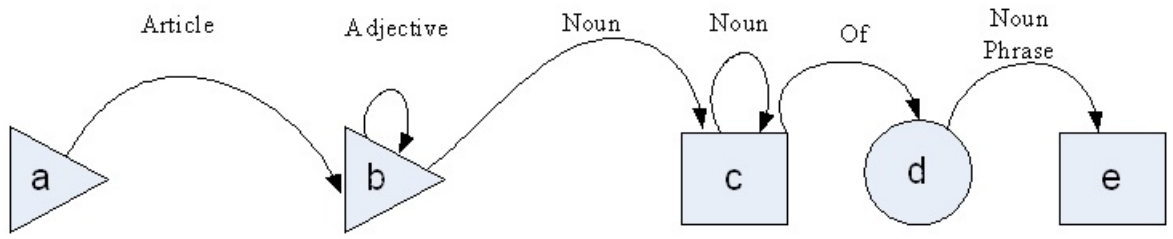
Sentence network

S



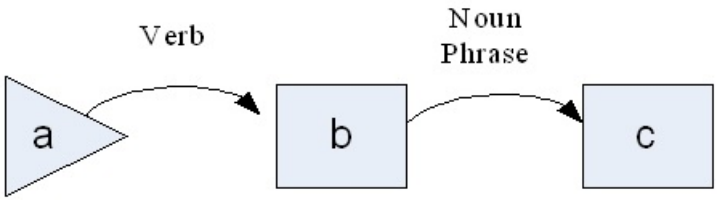
Noun phrase network

NP



Verb phrase network

S



Dictionary	
body N	sufficiency N
calcium N	supply V, N
determines V	the ART
digestible ADJ	to PREP
green ADJ	vegetables N

1 **$_0$ The $_1$ green $_2$ vegetables $_3$ supply $_4$ calcium $_5$.**

2 The green vegetables supply calcium to the body. [Not recognized by our simplistic parser.]

3 The green vegetables supply digestible calcium.

4 The green vegetables supply determines sufficiency of calcium.

Trace of a sentence parse (Arcs from transition network can be inferred)

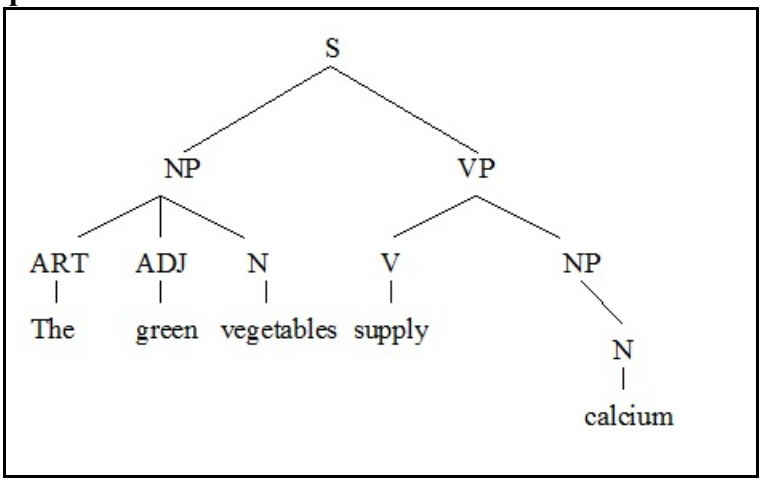
$_0$ The $_1$ green $_2$ vegetables $_3$ supply $_4$ calcium. $_5$

Step	From pos	From state	Segment	To state	To pos
①	**0**	**S^0 a**	**? (consult NP)**	**?**	**?**
②	0	NP^1 a	the	NP^1 b	1
③	1	NP^1 b	green	NP^1 b	2
④	2	NP^1 b	vegetables	NP^1 c	3
⑤	**0**	**S^0 a**	**the green vegetables**	**S^0 b**	**3**
⑥	**3**	**S^0 b**	**? (consult VP)**	**?**	**?**
⑦	3	VP^1 a	supply (V)	VP^1 b	4
⑧	4	VP^1 b	? (consult NP)	?	?
⑨	4	NP^2 a	calcium (*does not work, try starting at b*)	NP^2 a	4
⑩	4	NP^2 b	calcium	NP^2 c	5
1①	4	VP^1 b	calcium	VP^1 c	5
1②	**3**	**S^0 b**	**supply calcium**	**S^0 c**	**5**
1③			**Success:** End state of S, end of word list		

Superscript indicates the nesting depth

Result : {*S*
[*NP* (*ART* the) (*ADJ* green) (*N* vegetables)**]**
[*VP* (*V* supply) (*NP* (*N* calcium)**]**
}

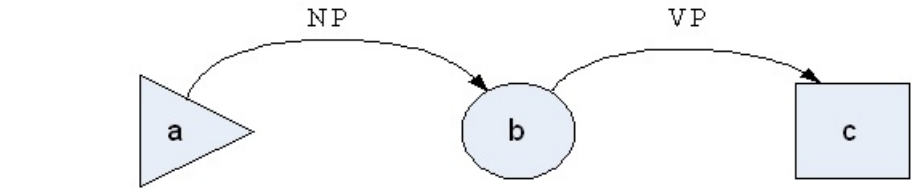
Parse tree representation:



Parsing of sentences: The three transition networks define a grammar for simple sentences.

Sentence network

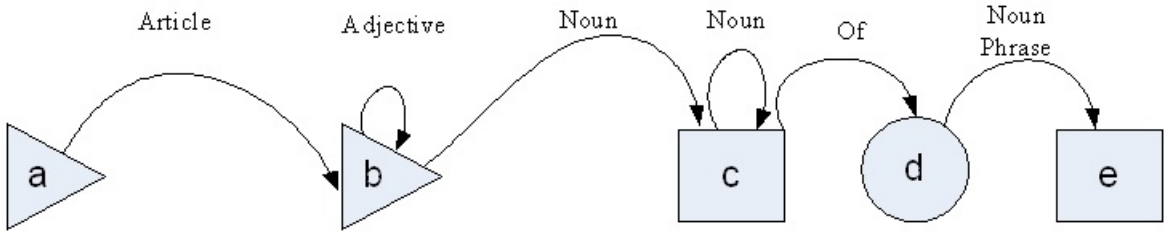
S



possible start state intermediate state possible end state

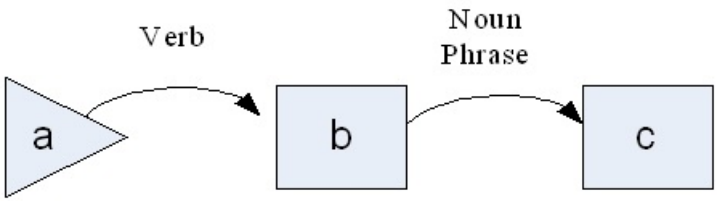
Noun phrase network

NP



Verb phrase network

S



Dictionary	
body N	sufficiency N
calcium N	supply V, N
determines V	the ART
digestible ADJ	to PREP
green ADJ	vegetables N

1 The green vegetables supply calcium

2 The green vegetables supply calcium to the body. [Not recognized by our simplistic parser.]

3 The green vegetables supply digestible calcium.

4 **$_0$ The $_1$ green $_2$ vegetables $_3$ supply $_4$ determines $_5$ sufficiency $_6$ of $_7$ calcium. $_8$**

Trace of a sentence parse with backtracking

$_0$ The $_1$ green $_2$ vegetables $_3$ supply $_4$ determines $_5$ sufficiency $_6$ of $_7$ calcium. $_8$

Step	From pos	From state	Segment processed	To state	To pos
①	**0**	**S^0 a**	**? (consult NP)**	**?**	**?**
②	0	NP^1 a	the	NP^1 b	1
③	1	NP^1 b	green	NP^1 b	2
④	2	NP^1 b	vegetables	NP^1 c	3
⑤	**0**	**S^0 a**	**the green vegetables**	**S^0 b**	**3**
⑥	**3**	**S^0 b**	**? (consult VP)**	**?**	**?**
⑦	3	VP^1 a	supply (V)*	VP^1 b	4
⑧	4	VP^1 b	? (consult NP)	?	?
⑨	4	NP^2 a	determines (*does not work, try starting at b*	NP^2 a	4
⑩	4	NP^2 b	determines (*does not work)*	NP^2 b	4
			Dead end, backtrack to *		
			Dead end, backtrack to *		
1①	3	**Backtrack, continue NP with supply as Noun**			?
	3	NP^1 c	supply (N)	NP^1 c	4
1②	**0**	**S^0 a**	**the green vegetables supply**	**S^0 b**	**4**
1③	**4**	**S^0 b**	**? (consult VP again)**	**?**	**?**
1④	4	VP^1a	determines	VP^1 b	5
1⑤	5	VP^1 b	? (consult NP)	?	?
1⑥	5	NP^2 a	sufficiency (*does not work, try starting at b)*	NP^2 a	5
1⑦	5	NP^2 b	sufficiency	NP^2 c	6
1⑧	6	NP^2 c	of	NP^2 d	7
1⑨	7	NP^2 d	? (consult NP)	?	?
20	7	NP^3 a	calcium (*does not work, try starting at b)*	NP^3 a	7
2①	7	NP^3 b	calcium	NP^3 c	8
2②	7	NP^2 d	calcium	NP^2 e	8
2③	4	VP^1 b	sufficiency of calcium	VP^1 c	8
2④	**4**	**S^0 b**	**determines sufficiency of calcium**	**S^0 c**	**8**
			Success: End state of S, end of word list		

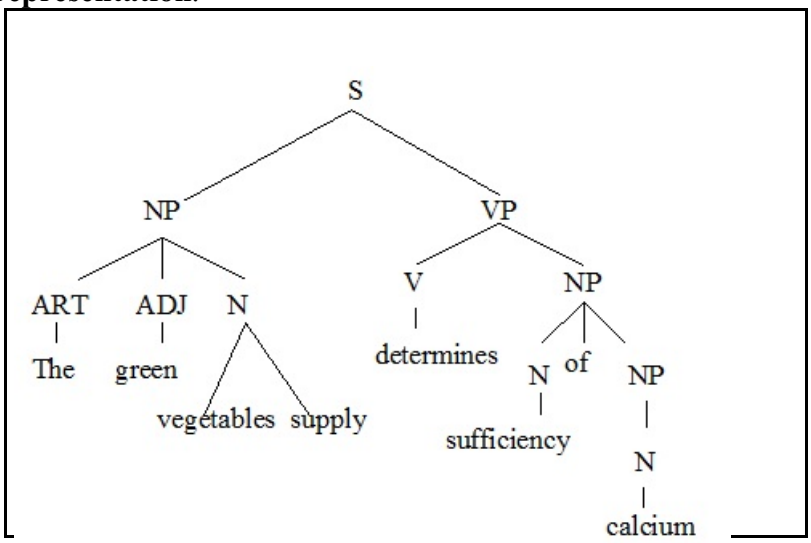
* Backtrack point Superscript indicates the nesting depth

Result: An analysis of the sentence structure, a sentence diagram.

$_0$ The $_1$ green $_2$ vegetables $_3$ supply $_4$ determines $_5$ sufficiency $_6$ of $_7$ calcium. $_8$

{S^0
[NP^1 **(**ART^1 the**) (**ADJ^1 green**) (**N^1 vegetables**) (**N^1 supply**)]**
[VP^1 **(**V^1 determines**) (**NP^2 (N^2 sufficiency) (of^2) (NP^3 (N^3 calcium)))**]**
}

Parse tree representation:



Other example of backtracking:

Compare *The old man cried.* with *The old man the ship.*

Hypothesis: Sentences that do not require backtracking in parsing are easier to read.

Example where backtracking makes reading difficult:

> "Any broadening of the government's role in health risks encouraging employers to give up providing health coverage for employees."
> (Editorial in the Washington Post 1999-7-30)

In a brief search of just the Web I could not find specific research on this. The following lecture materials deal with the issue in general

www.rci.rutgers.edu/~cfs/305_html/Understanding/Understanding_toc.html
(from course Computation and Cognition
www.rci.rutgers.edu/~cfs/472_html/home472.html

The following thesis deals with the problem of if and how people use syntax parsing in understanding sentences. It cites some previous work that found that people take longer in processing syntactically incorrect sentences even if they are not consciously aware of the incorrectness.
http://cognition.iig.uni-freiburg.de/team/members/konieczny/publ/DissLars.pdf

Parser evaluation

Word sequences that will not be recognized as sentences by our very simple parser

The green vegetables supply calcium to the body.	parser wrong in rejecting
*The green vegetables supply calcium strong bones	parser correct in rejecting

Parsing with semantic interpretation

Dictionary with semantic information

dishwasher N
- dishwasher 1
  - *Definition:* A person washing dishes
  - *Category:* Human (therefore animate)
  - *French:* plongeur
  - *German:* Tellerwäscher
- dishwasher 2
  - *Definition:* A machine washing dishes
  - *Category:* Machine (therefore inanimate)
  - *French:* lave-vaisselle
  - *German:* Spülmaschine

jolly ADJ
- *Definition:* Full of merriment and good spirit; fun-loving
- *Modifies:* Human

laughs V
- *Takes subject:* Animate
- *Takes object:*

white ADJ
- *French:* blanc
- *German:* weiss
- white 1
  - *Definition:* A color produced by mixing all rainbow colors, such as in snow.
  - *Modifies:* Non-human (inanimate object or animate object that is not human)
- white 2
  - *Definition:* A race designation used for Caucasian
  - *Modifies:* Human

$_0$ The $_1$ jolly $_2$ dishwasher. $_3$

$_0$ The $_1$ white $_2$ dishwasher $_3$ laughs. $_4$

$_0$ The $_1$ white $_2$ dishwasher $_3$ is $_4$ broken. $_5$

Two traces of semantically augmented parsing

$_0$ The $_1$ jolly $_2$ dishwasher $_3$

Step	From pos	From state	Segment	To state	To pos
①	0	NP a	the	NP b	1
②	1	NP b	jolly *Requires human noun.*	NP b	2
③	2	NP b	dishwasher *Works only if dishwasher is human* *Select* dishwasher 1	NP c	3

[NP (ART the) (ADJ jolly) (N dishwasher 1)**]**

$_0$ The $_1$ white $_2$ dishwasher $_3$ laughs. $_4$

Step	From pos	From state	Segment	To state	To pos
①	**0**	**S^0 a**	**? (consult NP)**	**?**	**?**
②	0	NP^1 a	the	NP^1 b	1
③	1	NP^1 b	white *two meanings:* white 1 *modifies non-human* white 2 *modifies human*	NP^1 b	2
④	2	NP^1 b	dishwasher *two meanings* dishwasher 1 *human* *agrees with* white 2 dishwasher 2 *machine* *agrees with* white 1	NP^1 c	3
⑤	**0**	**S^0 a***	**the white 2 dishwasher 1 *NP1 human*** **the white 1 dishwasher 2 *NP2 machine***	**S^0 b**	**3**
⑥	**3**	**S^0 b**	**? (consult VP)**	**?**	**?**
⑦	3	VP^1 a	laughs *Requires animate subject*	VP^1 b	4
⑧	**3**	**S^0 b**	**laughs *Select NP1***	**S^0 c**	**4**

{S
[NP (ART the) (ADJ white 2) (N dishwasher 1)**]**
[VP (V laughs)]
}

SLecture 7.1-7.2 Supplement. Cataloguing and metadata. Bibliographic control

Another scheme: **O'Neill and Vizine-Goetz** 1989
Note: In the original, they start with *book* and end with *work*.

Work We define a work as a set of related texts with a common source. The term *work* is frequently used inconsistently and, as a result, the distinction between an edition, a printing, and work is often unclear. The term *literary unit* has also been used as a synonym for work. Carpenter found that the words *book* and *work* are used loosely in various definitions and that "sometimes they are even used interchangeably, with a corresponding confusion" (Carpenter, 1981, p. 118).

Using our definition, a work may be composed of substantially different texts. The texts, however, must have been derived either directly or indirectly from a common source. As the text undergoes successive revisions or reexpressions over time, the words and symbols forming later texts may be very different from the original but still represent the same work. In our discussion of text we identified *Moby Dick: La Ballena Blanca and Moby Dick: The White Whale* as separate texts, yet we consider them to be the same work. The translation is closely related to the original and was derived directly from it.

Text [FRBR expression] A text is a set of editions with similar content. The term *text* was introduced by Wilson (1968, p. 6) to describe the content of a book as independent from its physical form. A text is "a sequence of words and auxiliary symbols" which has "no weight and occupies no space" (Wilson, 1968, p. 7). For example, as Hagler and Simmons (1982, p. 74) point out, "the Bantam edition of *Bleak House,* or the 1923 edition, or the Limited edition, may all be identical, word for word, in their textual content, their differences being only in paper, typography, binding, price, and perhaps publisher's name." Thus, a single text comprises three editions. Any edition that has been revised or updated will form a new text. New texts formed by revisions are often identified by numbered edition statements or edition statements such as "New Edition" or "Revised Edition." A new text may also occur as the result of an adaptation or translation. Felix Sutton's abridgement and adaptation of *Ben Hur* for children is a new text. Similarly, *Moby Dick: La Ballena Blanca,* the Spanish translation of *Moby Dick: The White Whale,* is a new text.

Edition [FRBR manifestation] An edition is a set of printings that, at the time of publication, were bibliographically identical. An edition is usually associated with a text. Therefore, if the text changes, so does the edition. However, there are some changes which create a new edition without resulting in a new text. For example, a new edition will be created when a text is republished by a different publisher or with significant changes in type image, or both.

Printing A printing is a set of books by the same publisher which are either printed at one time or printed at different times using the original type image with no more than slight but well-defined variations. As a general rule, the variations permitted within a printing are limited to the correction of minor typographical errors. The books themselves may or may not contain printing information. Commercial publishers commonly display printing information on the verso of the title page. The printing information usually includes the printing number and may also include the printing date.

Book [FRBR item] A book, as defined here, is the bibliographic entity at the lowest level of the hierarchy and is the only one which corresponds to a physical object. All of the other bibliographic entities are abstract concepts. Various terms are used synonymously with *book,* and the term *book is* often used in ways incompatible with our definition. For instance, *item*, *bibliographic item, copy, volume,* and *document* as well as other similar terms have been used interchangeably with the term *book*.

It is the individual book that is used to derive the information necessary for cataloging since, for cataloging purposes at least, all of the books constituting a particular printing are assumed to be bibliographically identical. Therefore, any book can be used to determine the bibliographic properties of the printing.

SLecture 7.1b Advanced exercise: Thinking about rules for corporate entry

The following pages give a number of possible rules and examples for those students with a particular interest in cataloging of documents. (These rules will not be on any test in 571.)

Issue A The first question deals with **choice of main entry**.

A work emanating from a corporate body was obviously, in fact, produced by some person or a group of persons (possibly having a chairperson), and this information is sometimes available to the cataloger. Make a rule about when to make the main entry under person and when under corporate body. Make a rule when to make an added entry for corporate body for those works that have person or title as main entry.

Issue B The following questions deal with **form of entry**, whether main or added entry.

Note: B1, B2, B3 are sub-issues of B for which a rule is needed. B1.1 and B1.2 are alternate rules for sub-issue B1.

B1 Form of name for institutions

Consider the result of applying the following alternative rules for dealing with works entered under a corporate body (either main or added entry) in a large catalog or bibliography from the point of view of ease of searching in the catalog. Consult the examples on p. 241 and 250 which illustrate the problems.

Compare Rule B1.1 and Rule B1.2 with respect to how well they accomplish ease of search.

Rule B1.1. Enter publications emanating from an **institution** (i.e. school, church, radio station, art gallery, etc.) under the place where the institution is located, unless the first word after the initial article is a proper noun or proper adjective. In that case, enter the institution under its name with place added if necessary to distinguish it from other institutions of the same name. Enter the publications of societies (clubs, guilds, fraternities, professional groups, etc.) under the society's name.

	Name in document	Form of entry
B1.1-1	*Metropolitan Museum of Art*	New York, N.Y. Metropolitan Museum of Art
B1.1-2	*University of Maryland*	Maryland (State), University
B1.1-3	*Freer Gallery of Art*	Freer Gallery of Art
B1.1-4	*American Medical Association*	American Medical Association
B1.1-5	*Gardening Club of Haynesville*	Gardening Club of Haynesville

Rule B1.2. Enter a publication emanating from a corporate body under the name of the body.

	Name in document	Form of entry
B1.2-1	*Metropolitan Museum of Art*	Metropolitan Museum of Art, New York, N.Y.
B1.2-2	*University of Maryland*	University of Maryland
B1.2-3	*Freer Gallery of Art*	Freer Gallery of Art
B1.2-4	*American Medical Association*	American Medical Association
B1.2-5	*Gardening Club of Haynesville*	Gardening Club of Haynesville

B1a. What rationale can you perceive for each of the above two rules?

B1b. For each rule try to pin-point where the catalogers and, more importantly, the catalog users would have trouble making decisions. What terms in the rules are particularly difficult to define or interpret?

B2 Names of subsidiary corporate bodies

Consider the fact that corporate bodies are frequently subsidiaries or divisions of other corporate bodies, sometimes with names clearly indicating dependency (like "division") and sometimes with independent names, such as National Research Council, a branch of the National Academy of Sciences. Consider the following possible rules from the point of view of ease of search:

Rule B2.1. List all publications of a corporate body under the name of the parent body.

	Name in document	Form of entry
B2.1-1	*Catalog Code Revision Committee of the American Library Association*	American Library Association
B2.1-2	*National Research Council of the National Academy of Science*	National Academy of Sciences

Rule B2.2. List all publications by sub-divisions or subsidiary bodies **indirectly**. That is, as a sub-heading to the parent body.

	Name in document	Form of entry
B2.2-1	*Catalog Code Revision Committee of the American Library Association*	American Library Association. Catalog Code Revision Committee
B2.2-2	*National Research Council of the National Academy of Science*	National Academy of Sciences. National Research Council

Rule B2.3 List all publications of the divisions or subsidiaries of a corporate body under the subsidiary directly.

	Name in document	Form of entry
B2.3-1	*Catalog Code Revision Committee of the American Library Association*	Catalog Code Revision Committee. (American Library Association)
B2.3-2	*National Research Council of the National Academy of Science*	National Research Council

B3 Name changes of corporate bodies

Corporate bodies are prone to change their names or to use different forms of their name on different publications. Consider the following solutions from the point of view of ease of search:

Rule B3.1 Change all entries to the latest name with references from the older forms of the name.

Rule B3.2 Enter all publications under the original name of the body with references from the newer forms of the name.

Rule B3.3 Enter each publication under the name given on the title page with cross references to previous and later forms of the name.

What about the cost of each rule?

B4 Change in form of name due to a change in the rules

B3 is about name changes in the real world. But how the name of a corporate body is entered in a catalog record also depends on the cataloging rules, such as the rules discussed in this exercise. Rules analogous to Rules B3.1 - B3.3 can be made on how to deal with this problem.

Examples illustrating the problems of form for corporate names

KEY C: Name of the Corporate body
L: Location of the corporate body if it is an institution
P: Person associated with the work (for some help with question)
T: Title of the Work

1. C: Freer Gallery of Art
L: Washington, D.C.
T: Dictionary Catalog of the Library of the Freer Gallery of Art, Smithsonian Institution.

2. C: Center for Applied Linguistics
L: Washington, D.C.
T: Sociolinguistics (papers from a conference sponsored by the Center)

3. C: Freer Gallery of Art
L: Washington, D.C.
T: Eugene and Agnes E. Meyer Memorial Exhibition

4. C: University of Washington
L: Washington state (for a state institution, the location is the state under ALA rules)
P: Charles L. Grossman and others (authors)
T: Migration of College and University Students in the United States (Report of contract between the University of Washington and the U.S. Dept. of Education) The University of Washington is the main entry in the University of Maryland catalog.

5. C: Library of the University of Washington
L: Washington state
P: Freda Campbell, compiler.
T: Filing Rules for the Catalogs of the University of Washington Libraries

6. C: University of Washington
L: Washington state
T: Men and learning in modern society (Papers delivered at the inauguration of Charles E. Odegard as president of the University of Washington)

7. C: Public Library
L: Washington, D.C.
T: Index to "The Rambler" (a local newspaper feature)

8. C: American Library Association
T: Bulletin of the ALA

9. C: American Library Association and others
P: C. Sumner Spalding, general editor
T: Anglo-American Cataloging Rules (North American Text)

10. C: American Library Association
P: none, or assume issued by president
T: Annual Conference Summary Report

Entries according to AACR2 rules XXX In progress, some items still need to be checked

Rule 21.1B2 deals with **whether to make an entry for the corporate body** (whether to establish a relationship)
Rule 24 deals with the **form of entry** (the form of the entity identifier for the corporate body)

	Entry	**AACR2 Rule**
1	Freer Gallery of Art	24.1
2	Center for Applied Linguistics Assuming this is an independent body. If it is part of a university, it would be different. Would need to research this	24.1
3	This one I'm not sure. I found a rule that said an exhibition should be treated as a corporate body if it reoccurs under the same name. So, if this is true for this exhibition, the entry would be: Eugene and Agnes E. Meyer Memorial Exhibition. If not, the entry would be: Freer Gallery of Art. In order for an exhibition to be the main entry, it must first meet the criteria to be considered a corporate body as stated in AACR2 21.1B1: "[For] art exhibitions, treat as corporate bodies only those that recur under the same name (e.g., Biennale di Venezia, Documenta)." If the exhibition is establishable as a corporate body, it may be used as the main entry heading under categories a) and d) of rule 21.1B2 of AACR2. from http://www.stanford.edu/~kteel/guidelines_mainentry.html	21.1B1
4	Grossman, Charles I am assuming this work to not be administrative in nature or the collective thought of the body	
5	University of Washington. Library. I am considering the library to be a subordinate body.	24.6b, 24.13A
6	University of Washington	24.6b
7	Washington, D.C. Public Library should this entry have a "government" designation? I'm not sure how that should be indicated Also, the preferred name for the locality may be District of Columbia(as used in the name of the library on their Web site)	24.18
8	American Library Association	24.1
9	American Library Association. I am considering AACR to be the collective thought of the body	24.1
10	American Library Association	24.1

SLecture 7.2b Dublin Core elements. Definitions

Element Name:	**Title**
Label:	Title
Definition:	A name given to the resource.
Comment:	Typically, a Title will be a name by which the resource is formally known.
Element Name:	**Creator**
Label:	Creator
Definition:	An entity primarily responsible for making the resource.
Comment:	Examples of a Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity.
Element Name:	**Subject**
Label:	Subject
Definition:	The topic of the resource.
Comment:	Typically, Subject will be expressed as keywords, key phrases or classification codes that describe a topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme.
Element Name:	**Description**
Label:	Description
Definition:	An account of the content of the resource.
Comment:	Examples of Description include, but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.
Element Name:	**Publisher**
Label:	Publisher
Definition:	An entity responsible for making the resource available
Comment:	Examples of Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity.
Element Name:	**Contributor**
Label:	Contributor
Definition:	An entity responsible for making contributions to the content of the resource.
Comment:	Examples of Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity.
Element Name:	**Date**
Label:	Date
Definition:	A date of an event in the lifecycle of the resource.
Comment:	Typically, Date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date value is defined in a profile of ISO 8601 [W3CDTF] and includes (among others) dates of the form YYYY-MM-DD.
Element Name:	**Type**
Label:	Resource Type
Definition:	The nature or genre of the content of the resource.
Comment:	Type includes terms describing general categories, functions, genres, or aggregation levels for content. Recommended best practice is to select a value from a controlled vocabulary (for example, the DCMI Type Vocabulary [DCT1]). To describe the physical or digital manifestation of the resource, use the FORMAT element.

Element Name:	**Format**
Label:	Format
Definition:	The physical or digital manifestation of the resource.
Comment:	Typically, Format may include the media-type or dimensions of the resource. Format may be used to identify the software, hardware, or other equipment needed to display or operate the resource. Examples of dimensions include size and duration. Recommended best practice is to select a value from a controlled vocabulary (for example, the list of Internet Media Types [MIME] defining computer media formats).
Element Name:	**Identifier**
Label:	Resource Identifier
Definition:	An unambiguous reference to the resource within a given context.
Comment:	Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system. Formal identification systems include but are not limited to the Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL)), the Digital Object Identifier (DOI) and the International Standard Book Number (ISBN).
Element Name:	**Source**
Label:	Source
Definition:	A Reference to a resource from which the present resource is derived.
Comment:	The present resource may be derived from the Source resource in whole or in part. Recommended best practice is to identify the referenced resource by means of a string or number conforming to a formal identification system.
Element Name:	**Language**
Label:	Language
Definition:	A language of the intellectual content of the resource.
Comment:	Recommended best practice is to use RFC 3066 [RFC3066] which, in conjunction with ISO639 [ISO639]), defines two- and three-letter primary language tags with optional subtags. Examples include "en" or "eng" for English, "akk" for Akkadian", and "en-GB" for English used in the United Kingdom.
Element Name:	**Relation**
Label:	Relation
Definition:	A reference to a related resource.
Comment:	Recommended best practice is to identify the referenced resource by means of a string or number conforming to a formal identification system.
Element Name:	**Coverage**
Label:	Coverage
Definition:	The extent or scope of the content of the resource.
Comment:	Typically, Coverage will include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity). Recommended best practice is to select a value from a controlled vocabulary (for example, the Thesaurus of Geographic Names [TGN]) and to use, where appropriate, named places or time periods in preference to numeric identifiers such as sets of coordinates or date ranges.
Element Name:	**Rights**
Label:	Rights Management
Definition:	Information about rights held in and over the resource.
Comment:	Typically, Rights will contain a rights management statement for the resource, or reference a service providing such information. Rights information often encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights. If the Rights element is absent, no assumptions may be made about any rights held in or over the resource.

SLecture 8.2b 8 WordNet pages

UB LIS 571 Supplement

Assignments

Assignment 1 Hypermedia explorations: Perseus and Freebase

Perseus 3 has more features than Perseus 4. An older version of the guided exploration, for Perseus 3, can be found at
www.dsoergel.com/UBLIS571DS-01.2-3Assignment1PerseusAssignmentOldVersion3.pdf

Supplement Assignment 6a

From conceptual data schema to database definition and querying

Objectives: 1 Understand the translation of an entity-relationship conceptual schema into table definitions for a relational database.

2 Understand how a relational database can be queried, especially, how data from several tables can be combined in an answer.

Tasks: 1 Define a table in Microsoft Access and enter some data.

2 Run some SQL queries in a MS Access database.

Deliverables: Printouts of the Table T5COFInstructor (showing the data you entered), of the definitions of Queries 4a - c/d, and of the query results..

This assignment is primarily a tutorial, Tasks 1 and 2 are at the end.

1 Introduction. The process of designing a relational database

Before you can start using a relational database management system (DBMS), such as Microsoft Access, you need to understand the process of designing a database.

Database design starts with developing an entity-relationship conceptual schema. This schema specifies the types of data to be covered in the data base, but it does not say how the data should be stored. In a relational database data are stored in **tables**. A table has **rows and columns**. Any given cell can have only one value in it. We thus need to transform the E-R conceptual schema into table definitions. Then we can enter data and query the database.

Sample table: T1CourseOffering

COF	Course	Semester	Room	TimeSlot	Limit	Enrolled
COF01	FDST101	1979SP	HBL1125	13	35	23
COF02	FDST257	1979SP	F1101	05	50	45
COF03	FDST663	1979SP	F1150	08	15	15
COF04	FDST101	1979SM	F0113	15	40	31

Tables are very simple yet very flexible data structures that are easy to manipulate. An **object-oriented database** can have more complex data structures but is harder to manipulate.

2 Developing the conceptual schema and defining tables

Conceptual schema for a university database

Entity types	Relationship types			
Course offering (ID: string, COF...) Course (ID: string, AAAA999) Semester (ID: str 9999[SP, SM, F] Room (ID: string) Time slot (ID: number [1 .. 20]) Count number	Course	<*offered as*>	Course offering	(1:N)
	Course offering	<*takes place in* >	Semester	(N:1)
	Course offering	<*meets in*>	Room	(N:1)
	Course offering	<*scheduled in*>	Time slot	(N:1)
	Course offering	<*has limit*>	Count number	(N:1)
	Course offering	<*has enrolled*>	Count number	(N:1)
Text Term	Course	<*has title*>	Text	(1:1)
	Course	<*deals with*>	Subject	(N:1)
Subject (ID: string S...)	Subject	<*designated by*>	Term	(1:1)
Person (ID: string)	(Course offering, Person)	<*has grade*>	Grade	n/a
Grade (ID: number [0,1,2,3,4]	Course offering	<*has instructor*>	Person	(N:N)

Explanation of symbols (by way of examples)

(1:1) A Subject is <*designated by*> exactly one (1)Term.
A Term <*designates*> exactly one (1) subject (in this DB).

(N:1) A Course offering <*meets in*> exactly one (1) Room.
A Room may <*serve as meeting place*> for many (N) Course offerings.

(1:N) A Course may be <*offered as*> many (N) Course offerings.
A Course offering <*is offering of*> exactly one (1) Course.

(N:N) A Course offering may <*has instructor*> many (N) Persons.
A Person may <*be instructor for*> many (N) Courses.

Note: These symbols are meaningful only for binary relationships. "Many" in this context means "more than one".

From entity-relationship conceptual schema to tables in a relational database

A table has rows and columns. Any given cell can have only one value in it. We could make a table for every relationship type, resulting in 11 tables. However, it is more efficient to keep the number of tables small by combining several relationship types into one table where possible. Here are the rules; as we apply these rules, you will see the rationale behind them.

Table definition rules

1 A multi-way relationship (3 or more) needs its own table, one row per statement.

This leaves binary relationships as candidates for combining into one table.

2 A N:N relationship needs its own table, one row per statement.

3 Formulate all remaining binary relationships so they are (N:1) or (1:1). All relationships with the same entity type on the left hand side can be combined into one table.

Applying these rules to the relationship types in the example:

By Rule 1, we select the three-way relationship *<has grade>* and make a table for it. This is easy: To each *<has grade>* statement corresponds a row in the table. Columns 1, 2, 3 correspond to the argument positions 1, 2, 3 in the relationship.

Table T4COFStudentGrade

(COF01, Tarr, L.) *<has grade>* 4
(COF01, Lund, M.) *<has grade>* 2

COF	Student	Grade
COF01	Tarr, L.	4
COF01	Lund, M.	2

All remaining relationships are binary. So, by Rule 2, we select *<has instructor>* since it is N:N (a course offering can have more than one instructor, an instructor can teach more than one course offering). Again, to each *<has instructor>* statement corresponds one row in the table; the table has just two columns. (See next page)

The remaining relationships are (1:1), (N:1), or (1:N); we can apply Rule 3 to them.

Course *<offered as>* Course offering (1:N) can be turned around to

Course offering *<is offering of>* Course (N:1)

The relationships can then be grouped into three blocks (starting at the top):

Block 1 in which all relationships start with Course offering
Block 2 in which all relationships start with Course
Block 3 in which all relationships start with Subject

Each block can be represented by a single table (see next page).

Conceptual schema for a university database (repeated)

Entity types	Relationship types			
Course offering (ID: string, COF...) Course (ID: string, AAAA999) Semester (ID: str 9999[SP, SM, F] Room (ID: string) Time slot (ID: number [1 .. 20]) Count number	(First relationship turned around) **Course offering**	***<is offering of>***	**Course**	**(N :1)**
	Course offering	*<takes place in >*	Semester	(N :1)
	Course offering	*<meets in>*	Room	(N :1)
	Course offering	*<scheduled in>*	Time slot	(N :1)
	Course offering	*<has limit>*	Count number	(N :1)
	Course offering	*<has enrolled>*	Count number	(N :1)
Text Term	Course	*<has title>*	Text	(1: 1)
	Course	*<deals with>*	Subject	(N :1)
Subject (ID: string S...)	Subject	*<designated by>*	Term	(1: 1)
Person (ID: string)	(Course offering, Person)	*<has grade>*	Grade	n/a
Grade (ID: number [0,1,2,3,4]	Course offering	*<has instructor>*	Person	(N :N)

Explanation of symbols (by way of examples)

(1:1) A Subject is *<designated by>* exactly one (1)Term.
A Term *<designates>* exactly one (1) subject.

(N:1) A Course offering *<meets in>* exactly one (1) Room.
A Room may *<serve as meeting place>* for many (N) Course offerings.

(1:N) A Course may be *<offered as>* many (N) Course offerings.
A Course offering *<is offering of>* exactly one (1) Course.

(N:N) A Course offering may *<has instructor>* several (N) Persons.
A Person may *<be instructor for>* several (N) Courses.

Note: These symbols are meaningful only for binary relationships. "Many" in this context means "more than one".

Table T5COFInstructor (for relationship *<has instructor>*, which is (N:N)

COF01 *<has instructor>* Kahn, L.
COF02 *<has instructor>* Kahn, L.
COF03 *<has instructor>* Simms, B.
COF03 *<has instructor>* Zog, H.

COF	Instructor
COF01	Kahn, L.
COF02	Kahn, L.
COF03	Simms, B.
COF03	Zog, H.

Note that COF03 needs two lines in Table T5.

The next three tables each represent a block of (N:1) relationships. The first block consists of six relationships, all starting with Course offering. We could express a group of statements formed with these relationships through six 2-column tables, but we can also express them as **one** table with 7 columns. Each row corresponds to a Course offering value. The column 1 takes the Course offering ID; columns 2 - 7 each take the right-hand value of one of the statements:

Table T1CourseOffering

COF01 *<is offering of>* FDST101
COF01 *<takes place in >* 1979Sp
COF01 *<meets in>* HBL1125
COF01 *<scheduled in>* 13
COF01 *<has limit>* 35
COF01 *<has enrolled>* 23

COF	Course	Semester	Room	TimeSlot	Limit	Enrolled
COF01	FDST101	1979SP	HBL1125	13	35	23

Thus, COF01 *<belongs to>* FDST101, COF01 *<takes place in>* 1979SP, etc. Put differently, each column is defined by a relationship type.

Why not add the relationship

Course offering *<has instructor>* Person (N:N)

as an 8th column to this table? (Hint: How would you handle COF03)

The two relationship types in the next block can also be represented as one three-column table:

Table T2Course

FDST101	*<has title>*	Introduction to food processing
FDST101	*<deals with>*	S12

Course	**Title**	**Notation**
FDST101	Introduction to food processing	S12

Each row corresponds to a course

Each of the other relationships needs its own table as follows (see explanation with Table T4):

Table T3Subject

S12 *<designated by>* Food processing

Subject	**Term**
S12	Food processing

Note that Tables T4 and T5 look just like Tables T1- T3. All tables are handled by the DBMS in the same way. But the interpretation of the tables is quite different:

Each row in a table can be seen as an instance of a frame. (The columns define the slots.) Tables T4 and T5 correspond to *minimal frames* (as defined in Lecture 4): Each frame represents one statement, each slot one argument of the relationship with which the statement is formed. Table 1 corresponds to an *extended frame*: Each row represents a block of binary statements, all starting with the same entity (the *focal entity*), each slot/column represents a binary relationship to the focal entity. In many contexts, the terminology is *file*, *record*, *field*. In object-oriented databases, the terms are *object* and *variable*. The following table shows the correspondences:

Table	File		
Row	Record	Frame	Object
Column	Field	Slot	Variable

Complexity increases from left to right: A field in a record can be *repeating* (have multiple occurrences), frame slots and object variables can have whole frames or complex objects as values (as in nesting boxes within boxes). Furthermore, a frame slot can have *procedural attachments*, and an object can have associated procedures, called *methods*, that process the data defined by the object's variables.

Complete tables with data

Table T1CourseOffering

COF	Course	Semester	Room	TimeSlot	Limit	Enrolled
COF01	FDST101	1979SP	HBL1125	13	35	23
COF02	FDST257	1979SP	F1101	05	50	45
COF03	FDST663	1979SP	F1150	08	15	15
COF04	FDST101	1979SM	F0113	15	40	31
COF05	FDST101	1979F	HBL1125	03	35	34
COF06	FDST257	1979F	F0112	15	15	12
COF07	CMSC620	1979F	HBL4115	20	25	18
COF08	CMSC424	1979F	HBL0109	05	60	45
COF09	CMSC420	1979F	HBL0103	14	15	13
COF10	FDST663	1980SP	F1150	04	15	12
COF11	CMSC424	1980SP	HBL0109	04	60	47
COF12	FDST101	1980SP	HBL1125	09	35	33
COF13	CMSC824	1980SP	HBL0109	07	20	15

Table T2Course

Course	Title	Subject
FDST101	Introduction to food processing	S12
FDST663	Seminar in meat canning	S17
FDST257	Vegetable pickling	S13
CMSC424	Database design	S19
CMSC620	Problem solving methods in artificial intelligence	S20
CMSC420	Data structure	S18
CMSC824	Relational database design	S19

Table T3Subject

Subject	Term
S12	Food processing
S13	Vegetable pickling
S17	Meat canning
S18	Data structure
S19	Database management
S20	Artificial intelligence

Table T4COFStudentGrade

COF	Student	Grade
COF01	Tarr, L.	4
COF01	Lund, M.	2
COF01	Kolb, T.	3
COF01	Doe, V.	0
COF02	Doe, J.	4
COF02	Smith, R.	4
COF03	Clay, S.	3
COF03	North, A.	3
COF03	Zipf, E.	1
COF04	Manet, J.	0
COF04	Kim, A.	4
COF04	Phillip, N.	3
COF05	Sprotto, L.	2
COF05	Jones, R.	4
COF06	Doe, V.	4
COF06	Jones, R.	4
COF06	Zipf, E.	3
COF07	Wang, L.	4
COF07	Meyer, P.	3
COF07	Gonzalez, A.	3
COF08	Gonzalez, A.	4
COF08	Hsiao, T.	2
COF08	Dellum, T.	3
COF08	Bush, M.	4
COF09	McCall, H.	4
COF09	Andreotti, S.	3
COF09	Yeltsin, B.	1
COF09	Sun, Y.	3
COF10	Tarr, L.	3
COF10	Doe, J.	4
COF10	Kolb, T.	3
COF11	McCall, H.	4
COF11	Yeltsin, B.	4
COF11	Chu, W.	3
COF12	Simon, R.	4
COF12	Gold, D.	3
COF12	Darrell, F.	1
COF12	Kovak, J.	3
COF12	David, J.	3
COF13	Gonzalez, A.	3
COF13	Hsiao, T.	3
COF13	Andreotti, S.	4
COF13	Sun, Y.	4

Table T5COFInstructor

COF	Instructor
COF01	Kahn, L.
COF02	Kahn, L.
COF03	Simms, B.
COF03	Zog, H.
COF04	Clay, S.
COF05	Kahn, L.
COF06	Simms, B.
COF07	Charniak, E.
COF07	Winston, P.

COF	Instructor
COF08	Date, C.
COF09	Minker, J.
COF10	Simms, B.
COF11	Minker, J.
COF12	Clay, S.
COF13	Codd, E.
COF13	Date, C.

3 Using Microsoft Access

3.1 Tutorial

All tables except T5COFInstructor are already defined in a database called University (on the distributed diskette) and "populated" (filled) with data. So you will first try running some queries, both predefined and created by you. You can open the University database from drive a: or copy it to your hard disk for faster operation.

Start MS Access.

Be sure the radio button *Open an existing file* is turned on.

If *University.mdb* is not already on the menu, click on *More files*, navigate to the directory where the database is stored. Double-click on *University.mdb*. A small database window opens; observe the navigation bar at the left.

Access opens with a list of tables. Double-click on *T1CourseOffering*. Examine the table, then close it. Right-click on the table; in the pop-up menu, click on *Design View* and examine the design window, then close it. Open the other tables and examine them.

Note: The main advantage of a database management system (DBMS) is that it can display data in any combination and format the user requires; it can present many views on the data. (This is a general principle of using computers for providing information. The driving force behind XML is structuring information so that it can be displayed in many ways, reused, "repurposed"). In the tutorial, you will display data using existing queries. We will start with a query for data from a single table and move on to queries that combine data from several tables, which is where the real power lies. One further point: In Access, queries can be shown in *Design view* or in *SQL* (Structured Query Language); you will look at queries both ways, but mostly in SQL.

In the navigation bar, click on *Queries*. Then double-click on *Query1a*; this will *run* the query and extract and format data from one or more tables as specified in the query. You will see a different display of data from table T1: Only some columns of data are displayed, the columns are in a different order, and the rows are sorted by course number.

Now, right-click on *Query1a*; in the pop-up menu, click on *Design View*. In the design view you can see how the query is specified in a format that approaches WYSIWYG (**W**hat **Y**ou **S**ee **I**s **W**hat **Y**ou **G**et). To see the SQL presentation, in the top menu bar, click on *View*, in the drop-down menu click on *SQL View*. You should see this (not as nicely formatted):

```
SELECT     T1CourseOffering.Course, T1CourseOffering.COF,
           T1CourseOffering.Semester, T1CourseOffering.Room
FROM       T1CourseOffering
ORDER BY   T1CourseOffering.Course;
```

This should be self-explanatory. In this example only, SQL keywords are bold.

SELECT is followed by the columns (fields) we want to show,

FROM introduces the table(s) from which data are do be displayed
ORDER BY specifies the *sort key(s)* for sorting the rows displayed

Check out Query 1b the same way. In SQL view you see (the line added to Query 1a is bolded):

```
SELECT    T1CourseOffering.Course, T1CourseOffering.COF,
          T1CourseOffering.Semester, T1CourseOffering.Room
FROM      T1CourseOffering
WHERE     T1CourseOffering.Course) >= "FDST"
ORDER BY  T1CourseOffering.Course;
```

Query 1b adds a selection condition so that only selected rows are displayed. The SQL keyword is **WHERE**.

It would be nice to see the course titles in the display. But Table T1CourseOfferings has only the course number. We could use the course number to access Table T2Course and get the course title from there. A relational database supports just this kind of combination of data from several tables. Check out Query 2a. (To see the full display, maximize the window.)

Query 2a (in SQL view, minus the extraneous [] MS Access puts in)

Note: From now on, the queries will be just in SQL because it is easier to deal with combination of data from several tables. All SQL queries are given here so you need not look the up in MS Access.

The part added to Query 1a is bolded. It is the matching condition that selects the correct row from Table T2 so we get the correct table.

```
SELECT    T1CourseOffering.Course, T2Course.Title,
          T1CourseOffering.COF, T1CourseOffering.Semester,
          T1CourseOffering.Room
FROM      T1CourseOffering, T2Course
WHERE     T2Course.Course = T1CourseOffering.Course
ORDER BY  T1CourseOffering.Course;
```

Note 1: This query corresponds to a two-step search in a graphical representation of the data: From Course offering to Course (based on Table T1), from Course to Title (based on Table T2)

Note 2: We could have added another column to Table T1CourseOffering. How often would the title for FDST01 appear in the revised table? What would this do to storage space and, more importantly, input effort and error possibilities?

Query 2b

```
SELECT      T1CourseOffering.Course, T2Course.Title,
            T1CourseOffering.COF, T1CourseOffering.Semester,
            T1CourseOffering.Room
FROM        T1CourseOffering, T2Course
WHERE       T2Course.Course=T1CourseOffering.CourseAND
            T1CourseOffering.Course) >= "FDST"
ORDER BY    T1CourseOffering.Course;
```

We will now build, step by step, a query to produce transcripts. That means, we will need to show grades by student, so we will start with a query on Table T4COFStudentGrade. Check out the data display for each query. You can see the SQL form here.

Query 3a

```
SELECT      T4COFStudentGrade.Student, T4COFStudentGrade.COF,
            T4COFStudentGrade.Grade
FROM        T4COFStudentGrade
ORDER BY    T4COFStudentGrade.Student;
```

A transcript should show the course number and the semester. We can get these pieces of information from Table T1CourseOffering, matching on COF:

Query 3b (additions to Query 3a bolded)

```
SELECT      T4COFStudentGrade.Student, T4COFStudentGrade.COF,
            T1CourseOffering.Course, T1CourseOffering.Semester,
            T4COFStudentGrade.Grade
FROM        T4COFStudentGrade, T1CourseOffering
WHERE       T1CourseOffering.COF=T4COFStudentGrade.COF
ORDER BY    T4COFStudentGrade.Student;
```

Finally, the transcript should contain the course title, which we can get from Table T2Course, matching on Course (a three-step search/navigation). And there are three smaller things to fix:

(1) The course offering number (COF) is not needed in the transcript, so we take it out from the list of fields following SELECT. (But is still plays a vital role in linking the tables.)

(2) The courses on a transcript should appear by semester, and within semester in course number order, so we add a second and third sort key after ORDER BY.

(3) The semester column should appear after the student, before the course number, so we rearrange the order of the fields after SELECT.

Here is the final query:

Query 3c (additions to Query 3b are bolded)

```
SELECT    T4COFStudentGrade.Student, T1CourseOffering.Semester,
          T1CourseOffering.Course, T2Course.Title,
          T4COFStudentGrade.Grade
FROM      T4COFStudentGrade, T1CourseOffering, T2Course
WHERE     T1CourseOffering.COF=T4COFStudentGrade.COF AND
          T2Course.Course=T1CourseOffering.Course
ORDER BY  T4COFStudentGrade.Student, T1CourseOffering.Semester,
          T1CourseOffering.Course;
```

3.2 Your tasks

Task 1. Define and populate a table

Define Table T5COFInstructor (see Complete tables with data at the end of Section 2)

In the Navigation bar to the left, double click on *Tables*

Double click on *Create table in Design view*

In the window that opens, enter a line for each of the two fields/columns.

When you are done, click the x in the upper right hand corner, answer *Yes* to save, in the box that opens enter the table name *T5COFInstructor,* answer *No* to primary key.

Double click on the new table and enter data in the window that comes up.

Hint: To speed up data entry, copy COF, then paste it every time you need it, using the shortcut key Ctrl-V.

Task 2. Define some queries

Query 4a: An alphabetical list of instructors with the course offerings they teach.

Query 4b: Add a column for course number to the display.

Query 4c: Add course titles, omit the course offering from the display.

Extra challenge

Query 4d: Instructors and the subjects they teach (as seen from the subjects of their courses).

Note: To print deliverables

Printing query definition: In design view you can see the SQL definition. Just copy and paste into a document.

Table content and query results can be printed directly.

Note: If you ever need to print table definitions, here is how to do it:

Tools > Analyze > Documenter.

Select the type of object (table, query, etc), check the specific objects you want to document, and click OK.

You will see a possibly lengthy display which can be printed.

gold **Assignment 7** ***Assigned: Feb. 25***
Lecture 6.1b ***Due: Mar. 4***

Applying linguistic techniques to retrieval problems

Objectives	Understand, through exploration, the possible improvements in free-text retrieval that can be achieved through **linguistic techniques** from Lecture 6.1b such as (for complete list see Lecture 6.1b) (P2.3.3,1) 1. Using all terms that designate a query concept (all **synonyms** of the query term). (P2.3.8,2#) 2. **Word Sense Disambiguation (WSD)** by syntactic analysis to determine part of speech (POS) and/or noun phrases (NP) and by semantic interpretation (from the multiple meanings of a homonym or polyseme, pick out the one that applies in the context. PXXX 3. **Resolution of anaphoric references** (what do *it*, *she*, *they*, *the machine*, ... refer to). PXXX
Materials: Explanation of the query.	The proximity operator **WS** requires that the two words occur **within the same sentence**. Thus the query formulation *forest* **WS** *fire* retrieves all passages in which the two **words** occur in the same sentence. This is the operator used in the baseline query formulation in the assignment. Most IR systems will take this query quite literally and look for the words (and that is how you need to analyze retrieval performance in Task 1. But the user is interested in the close mention of two concepts. That is where linguistic techniques come in.
Tasks	Explore possible improvements in free-text retrieval through **linguistic techniques**, using the examples in Table 1, which give some short passages of text and a query to be applied to this "collection". (1) analyze retrieval performance of a query using the **WS** operator and (2) (main task) suggest linguistic techniques that could be added to the retrieval system to improve retrieval. See the next page for more detailed instructions. You should still adhere to the requirement that the two concepts must be mentioned somehow in the same sentence.
Deliverables	The filled-out Tables 1 -3 with some analysis of Table 3.
Time	2 hours

over

UBLIS 571 Soergel Spring 2015 Assignment 7 Name
gold

Task 1 **Prelude**:

In Table 1 (facing page) for all passages that are relevant to the user's need as expressed in the query, put Y in the *Relevant* column; for all other passages put N.

Then for all passages that are retrieved by the query formulation, put Y in the *Retrieved* columns; for all other passages, put N

Fill in the 3x3 grid in Table 2 and compute performance measures: recall, discrimination, precision.

Task 2 **Main point:**
What **linguistic techniques** could be used to improve free-text retrieval performance? (Adding index terms to the passages is not an option.)
In Table 2, analyze each passage in turn; check for each the applicable linguistic technique(s).
In Table 3 summarize retrieval effects. For each technique, list all affected passages and indicate the effect: If the passage is now correctly retr

Query statement (description of information need / topic): **Forest fires**

Query formulation: forest **WS (W**ithin same **S**entence) fire* (fire* finds fire or fires)

Take out Table 1 (on next page) for passages to be retrieved and do Task (1), then fill in Table 2.

Then do Task (2).

In Table 1, check for each passage the linguistic technique(s) that would improve retrieval.
Then summarize the effects for

Table 1. Passages to be retrieved

Query formulation: forest **WS** fire* (fire* finds fire or fires)

Linguistic technique to use

• For each passage, in col.1 put Y if it is relevant, N if it is not • Then for each passage, in col.2 put Y if it is retrieved, N if it is not • For passages that are *relevant and retrieved* or *not relevant and not retrieved* no action is needed, no linguistic technique needs to be used. • For passages that are *relevant but not retrieved* or *not relevant but retrieved*, check linguistic techniques that would solve the problem (but preferably not introduce new problems). Check all techniques that apply. **Passage**		Relevant	RetrIeved	None needed	Synonyms	
P1	Forest fires in Indonesia cause serious air pollution in South East Asia.					
P2	The fire in Yellowstone Park destroyed 25% of the forest.					
P3	The fire station is located behind the city forest.					
P4	With fire in her eyes she chased him through the forest.					
P5	The soldiers opened fire into the forest.					
P6	The fire went out of control. It reached the forest and destroyed many acres.					
P7	The animal got scared by the fire burning in the field. It ran into the forest.					
P8	He asked whether he should fire the forest workers.					
P9	Many square miles of forest in the West are burning.					
P10	The dry wooded area went up in flames.					

Table 2. Recall, discrimination, precision

	Relevant	Not relevant	All
Retrieved			
Not retrieved			
All			

Recall: _____ **Discrimination:** _____ **Precision:** _____

$$\frac{\textit{relevant correctly retrieved}}{\textit{all relevant}} \qquad \frac{\textit{irrelevant correctly rejected}}{\textit{all irrelevant}} \qquad \frac{\textit{relevant correctly retrieved}}{\textit{all retrieved}}$$

Table 3. Linguistic techniques effect on individual passages

In the following table, enter only passages whose retrieval status changed by applying the technique. The row for synonym expansion is already filled in

	Passage relevant		**Passage not relevant**	
	Good change	Bad change	Good change	Bad change
	Was not retrieved Now retrieved	Was retrieved Now not retrieved	Was retrieved Now not retrieved	Was not retrieved Now retrieved
Synonym expansion	P9, P10	none	none	none
Noun phrase				
WSD				
Part of speech				
Anaphora resolution				

Table 4. Linguistic techniques effect summary.

	Effect on recall	**Effect on discrimination**
Synonym expansion	Always increase	No effect in sample, but could decrease (if an added synonym has other meanings)
Noun phrase		
WSD		
Part of speech		
Anaphora resolution		

Assignment 11 Request-oriented Indexing

Task/step 2. Build a precombined descriptor

For purposes of this assignment, we will consider only 3 of the 13 facets for arrangement; so we will combine only the elemental descriptors from these three facets:

B Division by mode of transportation

E Transportation system elements

J Passenger transport vs. freight transport

To represent the precombined descriptor, we combine the notations of the assigned elemental concepts into a notation string (see the example; for more examples see the Model Catalog, indexing with the London Education Classification). Descriptors from the other facets are not used for arrangement but they provide a more complete document representation and can be used for retrieval in a computerized system.

This raises the question: **In what order should the elemental notations be combined**? (**combination order**; the technical term used in the theory of faceted classification is *citation order*.) The combination order determines the arrangement of documents and document records (on the shelves, in a Web subject directory, in a display of search results, etc.).

The combination order must be determined in accordance with user needs. Assume the users are **engineers**. Civil engineers deal with building roads, harbors, airports, etc. while mechanical engineers deal with building vehicles. So the primary basis for arrangement should be the distinction between *traffic facilities* and *vehicles* made in Facet E, transportation system elements. Next in importance from the point of the engineer is *mode of transportation – ground*, *water*, *air*. (From an engineering standpoint, a passenger airplane has more in common with a cargo airplane than with passenger automobile.) So the combination order should be **E - B - J**.

Question: From the point of view of a **user of transportation services**, what should the combination order be? Or, as another example, consider **education with** three facets: *Grade level*, *Subject*, and *Type of student* (gifted, handicapped, etc.). Pick a user group for educational materials and determine what the combination order should be. Keep in mind that material on the facet listed first is kept together in the arrangement, material on the facet listed second is scattered, but still in "clumps", and material on the last facet is completely scattered in tiny pieces.

Arrangement of sample precombined descriptors in transportation

E1	**Traffic facilities**
E1B1	. Traffic facilities > Ground transport
E1B1J3	. . Traffic facilities > Ground transport > Passenger transport
E1B1J4	. . Traffic facilities > Ground transport > Freight transport
E1C1	. Traffic facilities > Water transport
E1C1J3	. . Traffic facilities > Water transport > Passenger transport
E1C1J4	. . Traffic facilities > Water transport > Freight transport
E1D1	. Traffic facilities > Air transport
E1D1J3	. . Traffic facilities > Air transport > Passenger transport
E1D1J4	. . Traffic facilities > Air transport > Freight transport
E2	**. Traffic ways**
E2B1	. . Traffic ways > Ground transport
E2B1J3	. . . Traffic ways > Ground transport > Passenger transport
E2B1J4	. . . Traffic ways > Ground transport > Freight transport
E2C1	. . Traffic ways > Water transport
E2C1J3	. . . Traffic ways > Water transport > Passenger transport
E2C1J4	. . . Traffic ways > Water transport > Freight transport
E2D1	. . Traffic ways > Air transport
E2D1J3	. . . Traffic ways > Air transport > Passenger transport
E2D1J4	. . . Traffic ways > Air transport > Freight transport
E3	**. Traffic stations**
E3B1	. . Traffic stations > Ground transport
E3B1J3	. . . Traffic stations > Ground transport > Passenger transport
E3B1J4	. . . Traffic stations > Ground transport > Freight transport
E3C1	. . Traffic stations > Water transport
E3C1J3	. . . Traffic stations > Water transport > Passenger transport
E3C1J4	. . . Traffic stations > Water transport > Freight transport
E3C7	. . . Traffic stations > Ocean transport
‣E3C7J4	**. . . . Traffic stations > Ocean transport > Freight transport**
E3D1	Traffic stations > Air transport
E3D1J3	Traffic stations > Air transport > Passenger transport
E3D1J4	Traffic stations > Air transport > Freight transport
E6	**Vehicles**
E6B1	. Vehicles > Ground transport
E6B1J3	. . Vehicles > Ground transport > Passenger transport
E6B1J4	. . Vehicles > Ground transport > Freight transport
E6C1	. Vehicles > Water transport
E6C1J3	. . Vehicles > Water transport > Passenger transport
E6C1J4	. . Vehicles > Water transport > Freight transport
E6D1	. Vehicles > Air transport
E6D1J3	. . Vehicles > Air transport > Passenger transport
E6D1J4	. . Vehicles > Air transport > Freight transport

Assignment 13.1 Dewey Decimal Classification

Case L: **Special education, level of education, and subject** (Advanced, optional)

<table>
<tr><th colspan="2">Dewey rules</th></tr>
<tr><td>At 370 Education</td><td>Remember from the instruction with 370 Education:

Class special education in a specific subject in 371.9 Special education.</td></tr>
<tr><td>At 371.9 Special education</td><td>To each subdivision identified by * add (append) the numbers following 371.904 in 371.904 3 - 371.904 7.
At 371.904 4 Programs in specific subjects we are further instructed to add (append) the number following 372 in 372.3 - 372.8

Example 1: the book
High school physics for blind students

would be classed under
371.911| 4|35 Special education > Blind students > in specific subjects/ Science and technology

Class number built following instructions at 371.9, which inherit down to 371.911 as indicated by the *:

4 is from 371.904 4 Programs in specific subjects

35 is from 372.35 [Elementary education in] Science and Technology

Example 2: the book
Accommodations for blind students in high school

would be classed under

371.911| 73

73 is from 371.904 73 Special education at secondary level</td></tr>
<tr><td>Note</td><td>Class 371.9 provides a rich example for this kind of analysis. DDC 20 allowed specifying both educational level and subject within subdivisions of special ed.</td></tr>
</table>

L **Analyze Case L; write your answers in the proper slots.** (Advanced, optional)

(1) Combination order

(2.1) Exhaustivity of indexing — which facets are represented in the class?

(2.2) Specificity of indexing — how specific is the concept from each facet?

:

(3) Effect on retrieval (recall and discrimination)

Invent some query topics for illustration

Write a very brief analysis