Dagobert Soergel

LIS 571
Organization and Control of Recorded Information

Spring 2011

Lecture Notes

January 19

Lecture 1.1

Introduction

Information professionals in the 21st century

Objectives	1 You should have an **appreciation for the wide variety of information tasks** that an education in information studies enables you to undertake and **the wide variety of information environments** you can work in. 2 You should have an **appreciation for the wide variety of information systems** that exist, including expert systems. 3 You should have an idea of Organization of Information concepts and skills that are needed in practice 4 You should have an overview of the course, what to expect and what is expected of you.

Outline

Introduction of students

Overview of the course

See sample CVs on UBlearns

What do information professionals do?

Answer questions, find things	• Explore the information need with the user: Understand the user's problem, understand what the user knows already, understand how the user thinks • Find answers in external and internal sources, such as ◦ Library catalogs, bookstore catalogs (mostly online now), ◦ Reference tools (bibliographies, biographical tools, almanacs, encyclopedias, etc.) , print or online ◦ Numeric databases, such as census databases, ◦ Maps ◦ The Web at large, intranets ◦ Archives (find records on a given subject even though they are not indexed by subject) ◦ A repository of instructional materials • Make a report that draws on several sources (extensive example: Congressional Research Service reports for Congress) • Organize the answer for quick perusal
Organize things so they can be found	• Catalog books using the MAchine-Readable Cataloging (MARC) format and the Anglo-American Cataloging Rules (AARC2-2002) • Catalog Web pages using Dublin Core • Catalog learning materials using educational metadata standards • Format documents using XML • Write abstracts for and index journal articles
Help people produce information	• Assist in editing and formatting documents • Help teachers in creating lesson plans (find instructional materials or learning objects, help format the lesson plan, help format materials for students, for example using graphic organizers) • Create Web pages (for the library or school media center on the organization)
Teach	Teach people • how to find information. Requires teaching them about information organization • how to assess and evaluate information • how to use and integrate information • how to present information
Develop and set up systems for all of the above	• Set up bibliographic and other databases, including library catalogs • Set up an intranet or a more ambitious enterprise portal that supports the work of all people in the organization • Set up document templates for easy creation of documents • Develop classification schemes, thesauri, taxonomies for special user groups (Each US agency must have a taxonomy to present its material to the public) • Help users with setting up their own personal information systems

On teaching and organization of information

Implications for curriculum and instruction

The findings from this study suggest that today's students need to learn – in a way that transcends their learning of specific content – a good deal about the structure of knowledge and about the importance of that structure. In order to learn in an information-rich environment, they need to learn

(1) that knowledge is indeed structured in meaningful ways;
(2) that various structures can be applied to various kinds of knowledge; and
(3) that a key part of learning is learning how to create personal structures that organize their own learning accurately and coherently.

They must learn that knowledge is an organized, systematically related set of ideas and that they need to work at building an understanding of that organization as well as learning the individual ideas. They must learn the nature and uses of various kinds of structures–for example, time lines, maps, and hyperlinks as well as traditional narrative structures–that they can use as tools for building their own knowledge. They must learn criteria and procedures for building appropriate and coherent structures that will allow them to integrate and communicate their thoughts. A curricular emphasis on teaching students how to structure information is, I believe, the most important implication for learning and teaching that stems from the presence of the information-rich environment in which we and our students live.

Learning theorists tell us that learning consists of constructing mental models or schemas, structures that are comprised of ideas and patterns or frameworks that organize and link those ideas. At some basic level, then, learning is the equivalent of organizing information. And no one in a school knows more about organizing information than the library media specialist best. Helping both teachers and students understand and learn to create a variety of ways to structure information is the key task for our profession in these best and worst of times.

From

Delia Neuman
Learning in an information-rich environment: Preliminary results
Treasure Mountain/Elms Research Retreat
Elms Resort and Spa
Excelsior Springs, MO
May 31, 2002

Types of information systems and information environments

Next page

Types of information systems and information environments

Information systems can be classified along many dimensions or facets. Any specific information system can be characterized by a combination, one concept from each dimension, for example

A system
- dealing with loosely structured information • using plain retrieval
- dealing with published or semi-published information
- serving a government agency
- information used for research and patient care
- dealing with the medical domain
- using paper technology for storage and accessing digital information

= **a traditional medical government library**

Sample dimensions (facets) for characterizing information systems

Types of information (such as bibliographic data, text and images, multimedia, numerical and other primary data, organization data and records);

> **Degree of structure of the information** (unstructured or loosely structured information as in text vs. tightly structured information as in numeric databases)

Processing to create answers: plain retrieval vs. drawing conclusions

Origin of information (such as generally published information - paper or online, government information, organizational information, information. about customers or patients);

> **internal vs. external information**

Users of the information, audience or organization served (groups - such as children, farmers, scholars, urban communities - or organizations - such as schools; universities or colleges; government agencies; businesses);

Uses of information (such as research, learning, problem solving, decision making, collaboration, day-to-day transactions);

Subject field (such as physics, medicine, or anthropology);

Technical means of providing access (such as paper vs digital).

The combinations are many, illustrating the flexibility and diversity of the information field

The table on the next page gives some examples for these dimensions. The information systems listed have characteristics for the other dimensions too but we do not list these.

Sample systems illustrating selected dimensions

Information & processing	**Expert systems** (medical diagnosis, computer configuration, detecting mineral deposits from satellite images, loan approval, etc.) **Software libraries** / databases for ease of access and reuse **Employment service databases** **Personnel information system** (usual personnel data plus skills and assignments to manage an organization's workforce) **Geographical information system (GIS)**
Users	**Information systems in organizations** Knowledge management: Make sure all applicable information is used to best advantage by organizing all types of internal and external sources of information – paper files and computer files no matter who keeps them, people and the knowledge they have in their heads – for access and usability. Information resources management Day-to-day transaction systems (order, inventory, etc.) Management information systems (MIS), Decision support systems Records management, archives (especially electronic records) A **personal information system** managing Web bookmarks, bibliographic references, downloaded Web pages, computer files, paper documents in personal collection, all kinds of notes, addresses, appointments
Use	**Instructional information systems** matching learner's needs with instructional materials In formal educational institutions In organizations for training (this is big business! Coordinate with personnel information system) For both: long-distance learning
Technology	**Paper libraries** of all kinds (public, academic and school, special) **Online information systems** Digital libraries Intranets An organizations Web site Any kind of **computer database** Bibliographic databases (e.g., Medline or OCLC's WorldCat) OCLC = Online Computer Library Center, the world's largest cooperative cataloging agency) Full-text databases (e.g., Westlaw or Lexis for law) Multimedia databases. Problem of retrieving still and moving images Substantive databases (directories, statistical data, material properties data)

Salaries of reporting professionals* by area of job assignment
Library Journal Oct. 2008, 2007 numbers. Full-time placements

ASSIGNMENT	No.	% of Total	Low Salary	High Salary	Average Salary	Median Salary
Acquisitions	18	1.3%	26K	70K	42K	39K
Administration	62	4.6%	18K	121K	44K	39K
Adult Services	44	3.3%	19K	48K	36K	36K
Archives	59	4.4%	14K	65K	40K	40K
Automation/Systems	21	1.6%	30K	93K	52K	48K
Cataloging & Classification	76	**5.6%**	18K	70K	40K	40K
Children's Services	75	5.5%	20K	55K	38K	38K
Circulation	51	3.8%	19K	55K	32K	33K
Collection Development	18	1.3%	30K	53K	41K	41K
Database Management	10	0.7%	24K	75K	41K	36K
Electronic or Digital Services	51	3.8%	24K	70K	45K	43K
Government Documents	8	0.6%	32K	50K	39K	38K
Indexing/Abstracting	6	0.4%	26K	26K	26K	26K
Info Technology	44	3.3%	32K	150K	53K	47K
Instruction	41	3.0%	17K	70K	42K	41K
Interlibrary Loans/ Doc. Del.	19	1.4%	21K	45K	34K	32K
Knowledge Management	7	0.5%	28K	51K	40K	41K
Other	110	8.1%	15K	115K	46K	43K
Reference/Info Services	293	**21.6%**	19K	70K	41K	40K
School Library Media Spec.	191	**14.1%**	25K	91K	44K	43K
Solo Librarian	51	3.8%	25K	57K	39K	39K
Usability/Usability Testing	15	1.1%	50K	90K	75K	78K
Web Design	1	0.1%	45K	45K	45K	45K
Youth Services	83	**6.1%**	20K	52K	36K	36K
TOTAL	**1354**	**100.00**	**14K**	**150K**	**42K**	**40K**

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	86K	95K
Deputy/Associative/Assistant Director	73K	80K
Dept Head/Branch Mgr/Coordinator/Senior Mgr	61K	61K
Manager/Supervisor of Support Staff	52K	54K
Librarian Who Does Not Supervise	48K	55K
Beginning Librarian	43K	45K

www.ala-apa.org/salaries/SalarySummary2008.pdf (Tables 1 and 2)

Some jobs in other environments (numbers from www.payscale.com, swz.salary.com, cbsalary.com, 2003 compilation by Roberta Shaffer)

Job title	From	To	Source
Chief Knowledge Officer	66K	130K	payscale
Chief Information Officer	90K	153K	payscale
Information Technology (IT) Manager	48K	92K	payscale
Chief Information Security Officer	127K	184K	salary
Information Architect	39K	103K	payscale
Ontologist	62K	84K	payscale
Senior content specialist	53K	74K	salary
Information Analyst	46K	126K	cbsalary
Consumer Information Director	55K	118K	cbsalary
Archivist	38K	52K	payscale
Strategic Information Planner	57K	75K	RS 2003
Business Intelligence Manager	55K	90K	RS 2003
Manager, Campus Technology and Academic Computing	62K	135K	RS 2003
Legal Information Specialist	50K	80K	RS 2003
Sarbanes-Oxley Compliance Manager, IT	89K	97K	payscale

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

Part 1. ***January 19 - January 26***

Foundations. Knowledge and knowledge representation

Lecture 1.2 ***January 19***

Information systems and information structure

Objectives	1 Gain an **appreciation for the variety of information systems** that exist, including expert systems. 2 **Understand the importance of information structure** / knowledge representation as the heart of an information system. 3 Have a **first idea of the entity-relationship approach** to knowledge representation.
Practical significance	Being knowledgeable about databases is a requirement for every executive assistant, let alone information professionals. Databases are the key to dealing efficiently with many types of information. Knowing about many types of information systems makes your skills more widely applicable and thus increases career opportunities. Expert systems are now widely used in many subject areas, for example, medicine, computer system configuration, and processing of loan applications; see the list at the end of Lecture 1.2 for some examples Designing or understanding the information structure of a system is key to building or using the system. The entity-relationship approach is the most natural and at the same time most general way for representing information.

Note on terminology: The Artificial Intelligence (AI) community speaks about *knowledge representation*, the database community speaks about *data modeling*.

Introduction

Purpose of an information system generally

Answer questions by either

finding an answer that exists ready-made in the database or

deducing an answer from multiple statements in the database.

Answering a question always involves going from something known to something unknown.

The lecture will show through examples **how information structure is used to find answers**.

We will look at three examples:

1 An expert system for medical prescriptions

2 A database that supports the operation of a university (Organizing Info., Chapter 3)

3 Medline, a bibliographic information system

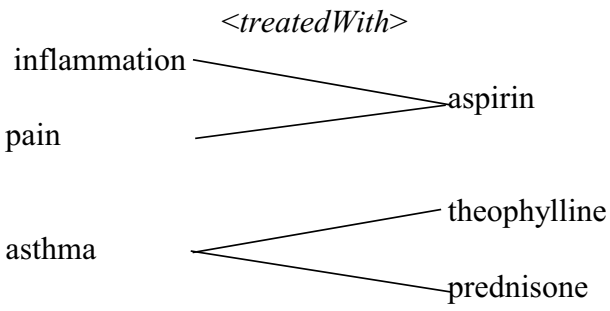
Example 1. **An information system for medical prescriptions**

Purpose	From the data in the patient record, including new diagnoses, find drugs the patient should take.
Questions	1 What drugs are used to treat asthma? Known: Disease asthma, unknown: Drug 2 What drugs should patient Fred take Known: Patient Fred, unknown: Drug

A Simplest system: only one kind of facts in the database, simple query

Question	What drugs are used to treat asthma? asthma <*treatedWith*> Drug X			
Facts	A1	inflammation	<*treatedWith*>	aspirin
	A2	pain	<*treatedWith*>	aspirin
	A3	asthma	< *treatedWith*>	theophylline
	A4	asthma	<*treatedWith*>	prednisone
Answer	theophylline, prednisone			

Graphical representation of facts



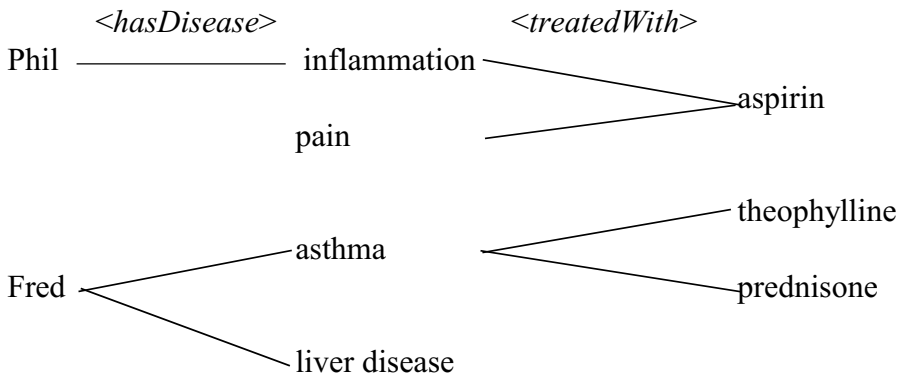
Note: It is customary to use the term "fact" in this context, even though "assertion" would be better. "Fact" implies an assertion is true, but not not all so-called "facts" stored in a database are true.

B More complex system: two kinds of facts that can be combined by a rule to answer more complex queries

Example: Add <*has disease*> facts, then answer queries about what drug(s) a person should take. Needs **rules on how to combine facts** (inference rules)

Question	What drug(s) should Fred take? Fred <*should take*> Drug X
Additional facts	B1 Fred <*hasDisease*> liver disease B2 Fred <*hasDisease*> asthma (New disease, basis for question) B3 Phil <*hasDisease*> inflammation
Rule	Person X <*shouldTake*> Drug Z IF Person X <*hasDisease*> Disease Y AND Disease Y <*treatedWith*> Drug Z
Answer	theophylline, prednisone

Graphical representation of facts



But what if

prednisone must not be taken by a person with liver disease, that is:

prednisone <*contraIndicatedWith* liver disease

C Still more complex system with still more facts and rules.

Question	What drug(s) should Fred take? (Now considering Fred's other diseases) Fred <*should take*> Drug X
Additional facts	C1 aspirin <*contraIndicatedWith*> peptic ulcers C2 theophylline <*contraIndicatedWith*> peptic ulcers C3 theophylline <*contraIndicatedWith*> arrhythmia C5 prednisone <*contraIndicatedWith*> liver disease
Rules	Person X <*shouldTake*> Drug Z IF Person X <*hasDisease*> Disease Y AND Disease Y <*treatedWith*> Drug Z AND Person X <*tolerates*> Drug Z Person X <*tolerates*> Drug Z IF Drug Z <*contraIndicatedWith*> Disease W AND NOT (Person X <*hasDisease*> Disease W) Person X <*tolerates*> Drug Z IF NOT (Drug Z <*contraIndicatedWith*> any Disease W)
Answer	theophylline

Note: If a drug is not ruled out based on the diseases seen from the patient record, a good system would **alert the physician to all contraindications** so that the patient can be checked out for these conditions.

Further refinements	• drug effectiveness • drug side effects and their severity • drug interactions and incompatibilities • drug cost A system containing all these data for a large number of drugs can prescribe as well as a human expert and would be called an **expert system.**

****

Example 2. **A database that supports the operation of a university**

Discussion of the example presented in Organizing Information, Chapter 3

Partial conceptual schema and some illustrative data for a university database

From Soergel, Organizing Information, Chapter 3

See next page

Two pages composed of material from DS85, Chapter3

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? Relates to ➜LIS 518 Reference Sources and Services (concepts important for how to search)

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 *<deals inclusively with>* 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.

Some general concepts

Information systems by extent of processing	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**, or **expert systems** The term **decision support system** is also used, particularly in connection with systems that use simulation and modeling to support business decisions.
Types of information processing	• inferential reasoning • mathematical computations • statistical analysis • simulation and modeling • neural networks and genetic algorithms
Plain information retrieval or database system	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 <*dealtWithIn*> Hearing tests Could also use links from words in the text or from person who is author.

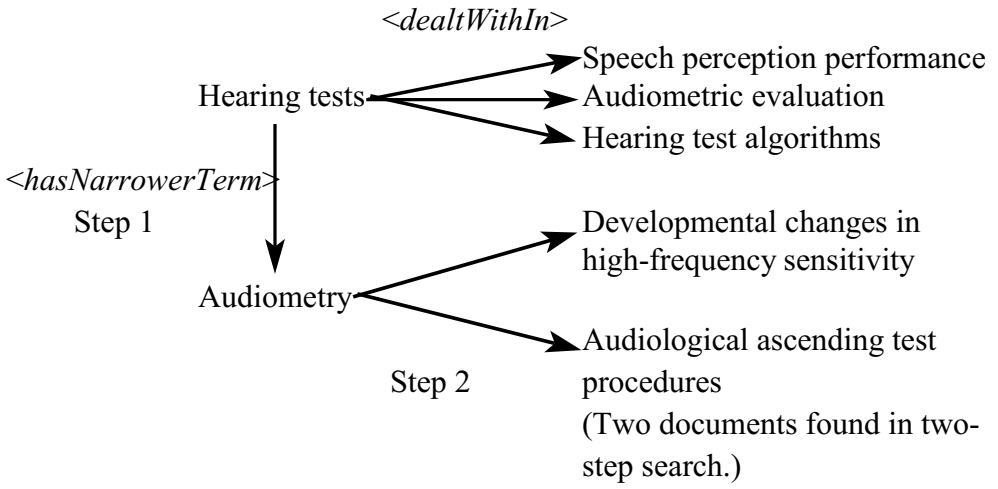
Expert system

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 librarian would.

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.
- 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.
- Learns.

Methods

- Information structure

  Knowledge representation - conceptual level

  Data structure and plain retrieval - access level

- Information processing

  inferential reasoning

  mathematical computations, statistical analysis, simulation and modeling, neural networks, and genetic algorithms

- Input/output methods based on information structure and processing:

  Understanding language for data acquisition and understanding user requests

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:

1. 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

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

3. Identifying plants, fish, insects etc.

4. Selecting foods for particular occasions.

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

6. 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?)

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

8. 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

9. Configuring a computer or other machinery

10. 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

January 19, 2011 Name (optional)

Free-write 1

Lecture 1.1 Introduction and Lecture 1.2 Information structure

- **Reflect** – what you learned, what was most important, what was most interesting, what was extraneous;
- **Ask questions** – ask for more explanation, how is a concept connected to other concepts, why is a concept important, how can it be applied, why is a reading important;
- Offer **critique and suggestions**;
- Say anything else you want to.

Over

Lectures 2.1 and 2.2 ***January 26***

The nature of knowledge and knowledge representation

Objectives	1 Understand the characteristics and facets of different types of knowledge and be able to apply this understanding to an analysis of information needs, the organization of information, and the evaluation of information found. 2 Understand findings from cognitive psychology on the way people form and deal with concepts and to apply these findings to a better understanding of information needs, to the design of classifications, and to information presentation. 3 Solidify the understanding of the approaches to knowledge representation as a basis for evaluating knowledge representation schemes.
Practical significance	Knowing about types of knowledge is important for • understanding information needs (as in interviewing a library patron before doing a search – reference interview); • analyzing and assessing information found; • determining how to organize and process information /knowledge in accordance with its type; • matching documents to the needs of the patron according to the type of information they contain. Knowing about types of concepts is important for understanding how people think and, therefore, how they ask questions, how they determine relevance, and how they process information. The answers, in turn, determine how information should be retrieved (retrieval should approximate human relevance judgment) and what information should be presented to a user in what form. Knowing about system-internal knowledge representation and conceptual data schemas is important for organizing a body of knowledge for retrieval and beyond that, for inference, that is, for a system that can draw conclusions from the knowledge stored (and thus create new knowledge), rather than simply retrieving what is there.

Outline

1 Types of knowledge: characteristics/facets/dimensions

- 1.1 Types of knowledge by content
- 1.2 Types of knowledge by scope of applicability
  - 1.2.1 Knowledge about regularities (laws, rules) vs. knowledge about individual detail from which the regularities can be derived
  - 1.2.2 Scope of applicability to natural or social phenomena. Scope of validity of a statement in space and time.
- 1.3 Types of knowledge by degree of "vagueness" of knowledge
- 1.4 Types of knowledge, other aspects

2 The nature of concepts / categories

- 2.1 Types of concepts. Individual concepts and class concepts
  - 2.1.1 Individual concepts – individual entities. Persistence over time
  - 2.1.2 Class concepts / categories. Simplified account
  - 2.1.3 Mass concepts (oil, flour, sugar) vs. count concepts (sugar cubes, books).
  - 2.1.4 Abstract concepts (freedom, justice). Can define the concrete class of all countries in which freedom prevails.
- 2.2 Objectivist vs. organism-centered view of categories
- 2.3 Explicit definition of categories vs. prototypes and fuzzy membership. Radial categories
- 2.4 Basic level categories (Eleanor Rosch)

3 Knowledge representation

- 3.1 Definition of knowledge representation (in the mind, on paper, for computers)
- 3.2 Approaches to knowledge representation
- 3.3 Some mechanisms in knowledge representation
- 3.4 Some criteria for describing and evaluating knowledge representations

1 Types of knowledge: characteristics/facets/dimensions

➜LIS 518 Reference Sources and Services: Selection of reference tools

1.1 Types of knowledge by content

Definitional knowledge (dictionary) vs. **assertive knowledge** (encyclopedia, world almanac).
Essential vs accidental attributes
(These are relative distinctions, see discussion of concepts in Section 2.1 below.)

Knowledge about **static relationships** vs. knowledge about **events and actions**

Knowledge by subject area or by relationship type used in a statement

1.2 Types of knowledge by scope of applicability

The more widely applicable an item of knowledge, the more important to obtain it, validate it, and store it in an easily accessible form. There are several aspects or facets of scope.

1.2.1 Knowledge about regularities (general laws, rules) vs. knowledge about individual detail from which the regularities can be derived

Knowledge about regularities that can be applied to many cases or throughout a system (such as a medical expert system), knowledge about individual detail can be applied only to the individual case.

Examples:

Regularity:	asthma <*treatedWith*> theophylline	[Applies to all asthma patients]
Individual detail:	Fred <*hasDisease*> asthma	[Applies just to one patient.]
Regularity:	FDST 257 <*hasPrerequisite*> FDST 101 FDST 257 <*isOfferedAs*> COF02	[Applies to all students in any section] [Applies to all students in this section]
Individual detail:	COF02 <*hasStudent*> R. Smith	[Applies to this student]
Regularity:	Hearing tests <*hasNarrowerTerm*> Audiometry	[Applies to all searches for these concepts.]
Individual detail:	Document 4 <*dealsWith*> Audiometry	[Applies to retrieval of just one document.]
Regularity:	Kepler's laws of planetary motion	[Applies to all planets at all times]
Individual detail:	The observational data about planet positions	[Each observation applies to the position of one planet at one time]
Regularity:	Burglary is punishable with 3 - 10 years of prison.	[Applies to all burglaries]
Individual detail:.	Weaver broke a large Window and entered the house. He took . . .	[Applies to this particular burglary]

To generalize from two of these examples.

Domain	Type of fact	Examples
Medical	**Regularities, general facts**	Facts about symptoms and diseases and treatments are broad; they apply to many cases.
	Individual detail facts	Facts about an individual patient are narrow; they apply only to one case.
Subject access to documents	**Regularities, general facts**	Facts about concept relationships are broad; they apply to all searches for the concepts involved and affect the retrieval of all documents indexed by one of the concepts.
	Individual detail facts	Facts linking a document to a concept (indexing facts) are narrow; they affect only the retrieval of this one document.

Ways to reason from past experience	
Regularities or laws are known	**Reasoning from general laws (deductive):** Draw conclusions on specific cases to which the laws can be matched.
Regularities or laws are <u>not</u> known	**Case-based reasoning (inductive)**: Find similar past cases and assume the new case will have similar outcomes. Examples: Weather forecast Decide a legal case where the law is inconclusive

Two important specific kinds of knowledge about regularities	
Type of knowledge	**Examples**
Knowledge about restrictions on data values	*A male individual of a mammal species cannot be pregnant.* *A two-year-old human cannot weigh more than 30 pounds.*
Default knowledge	Default knowledge: *A car has **four** wheels.* Specific knowledge about an individual case / knowledge about exceptions: *The Runabout has **three** wheels.*

1.2.2 Scope of applicability to natural or social phenomena. Scope of validity of a statement in space and time

Regularities can differ in the scope in which they apply:

Examples

Narrow scope	Broad scope
A law describing the free fall of objects towards the earth applies only on the earth (strictly speaking, only on a given point on the earth). Kepler's laws apply only to the movement of objects moving in an orbit around another object (originally they were conceived as applying only to the movement of the planets around the sun).	The general law of gravity applies to many phenomena throughout the universe; many more specific laws, like the two mentioned, can be derived form it.
A property value for a specific material (such as the electrical conductivity of copper) applies only to phenomena involving that material.	The gravitational constant holds through the entire universe (or so physicists think) and is involved in many phenomena.
A social rule, custom, or etiquette rule that applies only in one country	A social rule, custom, or etiquette rule that applies world-wide
Many rules of grammar apply to only one language family (such as all Indo-European languages or to only a single language).	Some linguists believe that some principles apply to all languages (language universals).

Related distinction	
Domain-specific knowledge For example: Effects of a drug; how to teach math to fourth-graders	**Common sense knowledge** For example: Use cost-benefit analysis, general principles of management

Example:

In a project of the World Bank to improve schools, the country found the knowledge they gained on general management procedures, especially procurement, just as useful as domain-specific (in World Bank speak: sector-specific) knowledge in education.

How are scope and usefulness of knowledge related?

1.3 Types of knowledge by degree of "vagueness" of knowledge

"Vagueness" is a vague umbrella term for the more well-defined distinctions listed below.

Concepts of "vagueness" of knowledge can be applied, for example, to knowledge of document relevance to a user's request (see below).

1.3.1 Precise vs. imprecise knowledge

This has to do with the error range with which variable values are given. Precise knowledge may be known or unknown. What degree of precision does the user need?

Examples

Knowing that a certain Shakespeare quote is from Macbeth is less precise than knowing it is from Macbeth, Act 2, Scene 3.

Numeric values derived from empirical data, such as physical measurements or poll results, are subject to error; they have an error range. In reporting such numbers, give only significant digits and preferably indicate the error range to avoid conveying unwarranted precision.

1.3.2 Certain vs. uncertain knowledge

Linked with risk. Combined with precision: Confidence intervals

Yes/no statements (such as facts or rules in an experts system) vs. probabilistic statements.

We can assume that a document either is relevant to a user's question or it isn't, with no shades of gray in between (see next point for a different stance), yet still say that document X has a probability of 0.7 of being relevant. Our knowledge about relevance is uncertain.

1.3.3 Graded assertions

For example, a document can be highly relevant or somewhat relevant. This can be expressed by a numerical score between 0 and 1. In other words, the set of relevant documents had no sharp boundary but rather is a fuzzy set. We cannot say that a document is a member of the set or that it is not a member of the set; rather, membership in a fuzzy set is a matter of degree.

1.3.4 Unambiguous vs. ambiguous statements (including intentional ambiguity)

Oder-Neisse line as border between Poland and Germany after WW2

1.3.5 Facts ("true", "objective") or statements asserted as facts vs. opinion

"Hard" statements vs. judgment statements. News page vs. editorial page

1.3.6 Knowledge about the accuracy, certainty, or trustworthiness of facts or rules

1.4 Types of knowledge, other aspects

1.4.1 **Beliefs.** Need to indicate whose belief.

1.4.2 **Modality of knowledge items** (descriptive, prescriptive, statement of possibility)

Descriptive statement:	The car is going 50 miles per hour (what is)
Prescriptive statement:	The speed limit is 45 miles per hour (what should be) (a prescription for drivers as to speed of their cars)
Possibility statement:	The car can go 100 miles per hour (what could be)

More examples

Descriptive knowledge:	Knowledge about the effects of calorie intake, specific nutrients (such as vitamin E), and exercise.
Prescriptive knowledge:	Guidelines on nutrition and exercise.

People writing to an advice columnist report the facts of the case as they see them – descriptive knowledge. The advice columnist tells them what to do – prescriptive knowledge.

The law is prescriptive knowledge

Politicians and planners deal with the art of the possible; they need knowledge of what is possible. For example, some people claim to know that it is not possible to change Social Security because public opinion is against it and the votes to pass legislation are not there. On the other hand, visionary politicians may defy conventional knowledge of what is possible and make things possible. Time horizon of statements about possibility.

1.4.3 **Knowledge about what kinds of knowledge are important**: Conceptual data schema (introduced in Lecture 1.2)

2 The nature of concepts / categories / classes

Importance: The nature of concepts is fundamental to information processing in people and in machines (see readings, particularly Skemp). Another way of looking at types of knowledge.

2.1 Types of concepts. Individual concepts and class concepts

2.1.1 **Individual concepts – individual entities**. Persistence over time

2.1.2 **Class concepts / categories**. Simplified account

See Sections 2.2 - 2.4 for a discussion of the complexities of the structure of categories.

Concepts have	
Intension, **intensional definition,** "meaning"	**Definitional knowledge** as opposed to world knowledge (empirical knowledge) A concept or class defined in terms of attributes or characteristics all entities must possess in order to be members of the class. These are called **essential attributes or characteristics**. A characteristic of an individual entity can expressed in several ways: (1) the entity possesses an attribute (2) the entity is capable of entering a given relationship (occupy a comparable place in a network of relationships) A **query formulation is a definition**. It defines what it means for a document (or a person, or a computer program) to be relevant for the user. It encapsulates the user's intention.
Extension	The set of individual entities belonging to the category For example, the set of all relevant documents In some cases it is possible to define a category by exhaustively listing all its members. This is called an **extensional definition**.

Example definitions (isa = is a type of):

Class	Definition
red balls	All objects that meet two conditions: Object <*hasType*> ball **and** Object <*hasColor*> red
pews	All objects that meet two conditions: Object <*isa*> bench **and** [Object <*locatedIn*> Building, Building <*isa*> church]
government documents	All documents that meet the conditions: Document <*publishedBy*> Organization, Organization <*hasType*> government agency.
water-soluble substances	All chemical substances that meet the condition Substance <*solubleIn*> water.
?	All English words that meet the condition Word <*canServeAsObjectFor*> refine.

Essential attributes	Attributes that are used in the definition of a class
Accidental attributes	Any other attribute that one or more members of a class may possess

It often happens that all members of a class share an accidental attribute, that is an attributes that is not defining but nevertheless present in each member of the class. Such a general law can be determined by observation.

Example: Assume it is true that all government documents are authoritative.
Thus, if we have ascertained that a given document meets the definition for government document, we can conclude that the document is authoritative (knowledge of a regularity).

It is this ability to predict the behavior of an entity once it has been identified as belonging to a concept/category that makes for the usefulness of concepts; concepts are essential for economy of mental operations.

In law:

Fit facts of the case to a legal concept, for example determine that the facts of a case meet the definition of burglary.
Then apply the legal rule applicable to that concept.

Erroneous generalization: Stereotypes.

Relationship of definition to empirical data. Examples: One can define an animal species through a list of attributes such that no animal existing in nature fits the definition. Yet such fictitious animals are written about and depicted and become objects of searches. *Imaginary animals* is a very popular search topic in the International Children's Digital Library (ICDL).

2.1.3 **Mass concepts** (oil, flour, sugar) vs. **count concepts** (sugar cubes, books)

2.1.4 **Abstract concepts** (freedom, justice). Can define the concrete class of all countries in which freedom prevails.

2.2 Objectivist vs. organism-centered view of categories

Next page

2.2 Objectivist vs. organism-centered view of categories

(Quotes from Lakoff, *Women, fire, and dangerous things*. U. of Chicago Pr.; 1987)

Important: **Information is not just transmitted but needs to be actively processed and assimilated by the receiver or learner** (see the last paragraph of this section).

Objectivist view of categories (as characterized by George Lakoff)

- Symbols that correspond to the external world are *internal representations of external reality*.
- Abstract symbols may stand in correspondence to things in the world independent of the particular properties of any organism.
- Since the human mind makes use of internal representations of external reality, the mind is *a mirror of nature*, and correct reason mirrors the logic of the external world. (p. XIII)

Organism-centered view (DS term) of categories (George Lakoff)

- Do meaningful thought and reason concern merely the manipulation of abstract symbols and their correspondence to an objective reality, independent of any embodiment (except, perhaps, for limitations imposed by the organism)?
- Or do meaningful thought and reason essentially concern the nature of the organism doing the thinking – including the nature of its body, its interactions in its environment, its social character, and so on? (p. XV - XVI)
  Embodied cognition versus symbolic representation

A balanced view (D. Soergel)

- Interacting with the physical, social, and intellectual world around us as well as with our own selves, we form complex mental models which allow us to better understand the world around us and better understand ourselves and thus help us to take actions in the world towards achieving our objectives. This formation of mental models has a social dimension; it is often done in interaction with or building on the models of others – as in group learning.
- These mental models, which include concepts / categories, reflect a structuring of experience in ways useful to the person. A person's experience is shaped by perceptions of the world (within the limitations of the person's faculties for perception and thought) and by the modes of interaction with objects in the world. Thus, **a mental model is not simply a mirror image of the outside world but rather an actively shaped image, adapted to the person's needs**, often distorted, often enriched (or contaminated, depending on one's point of view) with elements that have no counterpart in the "real" world (but might well be realized as the person shapes the world).

The sense-making approach to information service
The view of mental models, concepts, and categories presented above is important for an understanding of how people use information and what information should be provided to people. In this view, a person must assimilate information into her mental model; a person **must make personal sense of the information**. Different people may get different things out of the same document. In that sense, one might say that information does not exist objectively, but only as it gives rise to a change in a person's mind. Or that a book does not convey information as much as it is a stimulus for the reader to create and elaborate her own information in her own mind.

The sense-making approach in related disciplines	
In literary theory	This is the position taken by **reader response theory**. The expression "I did not get much out of it" is in tune with this "active reader" position. The art of giving the reader a "relevant" book, then, is to find a book that allows this reader with his background and mental models to "get something out of" the book, to construct his own knowledge, updating his mental models in a way that will help him to find better solutions to the problems he faces.
In education	The **constructivist theory of learning** holds that we learn best by constructing or reconstructing knowledge for ourselves. **Discovery learning** is a closely related approach. It holds that a students learn best when they explore a subject and discover facts and relationships for themselves. In science this means that students discover scientific laws through their own experiments. The **job of the teacher or information specialist** then is to create an environment, including access to information, that enables students or users to do their own discovery and knowledge construction with guidance provided only to the extent necessary ("scaffolding").

2.3 Explicit definition of categories vs. prototypes and fuzzy membership Radial categories

Prototype. Example *Chair*:

Chair, living room chair, kitchen chair, lawn chair, easy chair, rocker, armchair, chaise longue, bar stool, stool?, footstool??, ottoman??

Necessary attributes vs. sharing a sufficient number of attributes.
Knowledge of concepts stored in memory as explicit definitions or prototypes?
In reality a mixture of both?

Importance of examples in thesaurus scope notes

Radial category. Example**:** *Mother*
(a category that has a central case but then many cases deviating more or less in different directions)

There are many "models" of what a *mother* is (Lakoff 1987, p. 83).

> "The central case, where all the models converge, includes a mother who is and always has been female, and who gave birth to the child, supplied her half of the child's genes, nurtured the child, is married to the father, is one generation older than the child, and is the child's legal guardian."

The following cases share some, but not all, of these features. The first four emphasize a biological perspective, the others a social perspective.

- Biological mother (also called natural mother, but that term was abandoned)
- Birth mother (term for biological mother in the context of adoption)
- Surrogate mother
- Genetic mother
- Rearing mother
- Stepmother
- Adoptive mother
- Foster mother
- Unwed mother

What would you search for if the user says he wants documents about *mother*?

Application to retrieval:

Consider the category consisting of the documents (or statements, such as statements of fact or hypotheses, etc.) relevant to a query. We can define such a category in two ways:

- through a query formulation that explicitly specifies the features that make a document relevant (expressing the intent of the user, intensional definition). (This query formulation could be applied in a Boolean search (to be retrieved, a document needs to meet all conditions) or in a ranked retrieval search (documents are retrieved even if they do not meet all conditions exactly, and are ranked by how closely they meet the conditions);
- through a sample document that serves as a prototype of relevant documents ("more like this") or several documents that serve as examples.

The category "relevant documents" can be a radial category when there are different ways in which a document can be relevant to the query. Needs several query formulations.

2.4 Basic level categories (Eleanor Rosch)

This theory considers the **application of categories or concepts to action**.
From this perspective, what categories are most useful and worth the effort to learn?

Example

If somebody tells you that there is a piece of *furniture* in a room you have been assigned, that does not tell you much. You still do not know what you can do in the room. But, if somebody tells you there is a *chair* in the room, that tells you a lot more; you know you can sit down. If somebody tells you there is a *easy chair* in the room, you know a little more, but not much more; you still know only that you can sit down (perhaps a bit more comfortably).

There is a big information gain from *furniture* to *chair*, but a small gain from *chair* to *easy chair*. So it is worthwhile to learn about the category of *chair*, but the added benefit of knowing all the specific types of chair would be low and the learning effort would be very high.
chair is at the optimal level in the hierarchy, it is a **basic level category**.

More examples

Superordinate	**Basic level**	**Subordinates**	
Furniture	Chair	Kitchen chair	Living room chair
		Lawn chair	Armchair
		Easy chair	Bar stool
		Stool?	Footstool??
	Table	Dining room table	Kitchen table
		Card table	Folding table
		Pool table	Operating table

Empirical results of studies in cognitive psychology

- Subjects were given words that name categories of objects, such as *furniture*, *chair*, *lawn chair*, and were asked to list **attributes** of that category.
  For **superordinate categories**, such as *furniture*, subjects listed **few attributes**.
  For **intermediate categories**, such as *chair*, subjects listed **many attributes**. **Basic level**
  For **subordinate categories**, such as *lawn chair* or *easy chair*, subjects listed **a few additional attributes** beyond those for *chair*.
  "Basic level categories are the most inclusive level of classification at which **objects have a significant number of attributes in common**." (p. 214)
- Basic level categories are the most inclusive level of classification at which **objects share substantive functionality**. Example:
  There are **few, if any movements** or other things you do **in common** to all types of *furniture (table, chair, cabinet)* (**superordinate category**). (p. 217)
  But people make the **same kind of movements** (sitting down) for all *chairs* (**basic level**)
  These **movements are hardly different** for a *easy chair* (**subordinate category**)
- Basic level categories are **learned first**.

Level	Number of attributes	Number of instances	Number of categories at that level	Usefulness for action
Superordinate	Few attributes	Huge number of instances	Few categories	Low
Basic level	Many attributes (high information gain)	Large number of instances	Medium number of categories	High
Subordinate	Only a few more attributes (low information gain)	Low number of instances	Very large number of categories	Only slightly higher

Note: Basic level may depend on group - culture and subculture.

Optional

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)

3 Knowledge representation (KR) ➜ LIS 506 IT, LIS 569 Data Management

3.1 Definition of knowledge representation (in the mind, on paper, for computers)

Knowledge representation is the expression of knowledge through a system of symbols or signs, such as words, Dewey numbers, or icons. A knowledge representation scheme must provide

- symbols that refer to objects in the world or objects in the mind (put differently, symbols that refer to entity values, roughly, nouns);
- symbols that refer to relationship types (roughly, verbs);
- a syntax that allows for the expression of statements consisting of entity identifiers linked through relationship symbols.

Natural language is a very expressive knowledge representation system, but it is hard for a computer system to figure out what a natural language text means and then act on this knowledge. Need simpler KR systems for useful computer systems.

Approaches
Entity-relationship representation (very common in the database field)
Semantic network representation
Frame representation (artificial intelligence & object-oriented programming)

Note: In 571 we talk about knowledge representation in the abstract. Implementation in databases is treated briefly in 506 Information Technology and extensively in 569 Data management. On the 571 Web site there is an assignment that takes you through creating and querying a simple Microsoft Access implementation of the University Database.

Knowledge representation examples (for computer systems)

A small example with data on birds

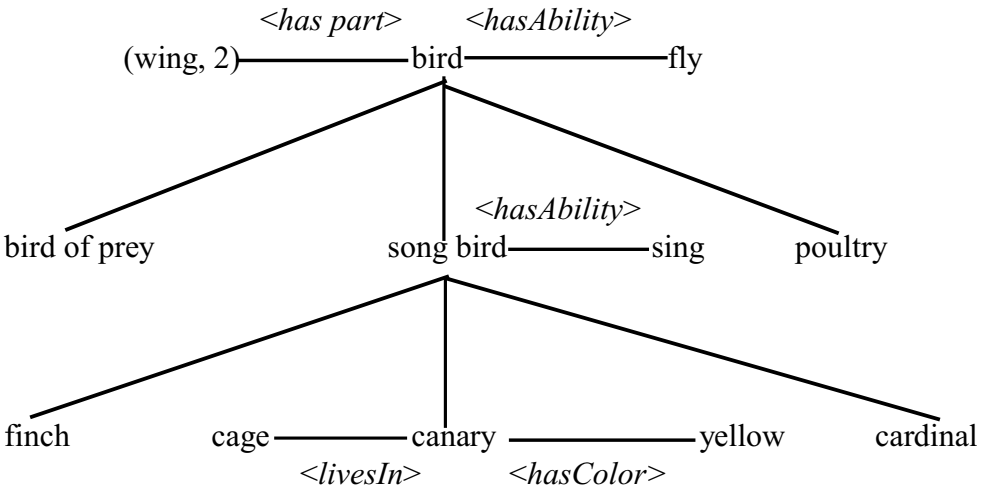
A large example with data on food products

Simple example: Representation of data on birds

Entity-relationship representation of data on birds

canary	<*isa*>	song bird	canary	<*hasColor*>	yellow
finch	<*isa*>	song bird	song bird	<*hasAbility*>	sing
cardinal	<*isa*>	song bird	bird	<*hasAbility*>	fly
songbird	<*isa*>	bird	bird	<*hasPart*>	(wing, 2)
bird of prey	<*isa*>	bird	canary	<*livesIn*>	cage
poultry	<*isa*>	bird	bird	<*isa*>	animal

Semantic network representation of data on birds



Hierarchical inheritance	A node lower in the hierarchy **inherits** the characteristics of nodes above it. For example, song bird <*hasAbility*> fly and songbird <*has part*> (wing, 2) are both true; song bird **inherits** these characteristics from bird. Note: The existence of a hierarchy alone does not make hierarchical inheritance. Only when the hierarchy is used to pass down characteristics from higher nodes to lower nodes is there hierarchical inheritance
Spreading activation	Activation (or attention) may spread from a node to a neighboring node: A person thinking about yellow (yellow is activated) may be reminded of a canary (canary is activated), and then of song bird and then of a bird singing (sing is activated) and then, again starting from song bird, of any bird and thus of a bird flying)

Frame representation of data on birds with hierarchical inheritance

Frame for: **bird**	
isa:	animal
includesSpecific:	song bird; bird of prey, poultry /* does not inherit down */
hasColor:	
hasPart:	(wing, 2)
hasAbility:	fly
livesIn:	

Frame for: **song bird**		[inherited]
isa:	bird; [animal]	
includesSpecific:	canary; finch; cardinal	
hasColor:		
hasPart:	[(wing, 2)]	Anything that is true for a bird is true for a songbird through inheritance
hasAbility:	sing; [fly]	
livesIn:		

Frame for: **canary**		[[inherited from two levels above]]
isa:	song bird; [bird]; [[animal]]	Anything that is true for a bird or a song bird is true for a canary through inheritance
includesSpecific:		
hasColor:	yellow	
hasPart:	[[(wing, 2)]]	
hasAbility:	[sing]; [[fly]]	Anything that is true for a bird or a song bird is true for a canary through inheritance
livesIn:	cage	

Frame for: **penguin** /* added to illustrate inheritance override */	
isa:	bird; [animal]
includesSpecific:	
hasColor:	white; black
hasPart:	[(wing, 2)]
hasAbility:	swim; NOT fly; [fly] (example of **overriding** an inherited piece of data)
livesIn:	Antarctica

See http://percevia.duncraft.com/ for a bird database using more bird attributes

This page is intentionally blank

More elaborate example: Representation of data on food products

FP stands for Food Product

Purposes of the food information system	**Government agency**: Determine the safety of a food product **Consumer:** Find food products to be avoided with a given allergy **Cook:** Prepare a food product **Food manufacturer:** Produce the ingredient label
Sample questions	Find all products to be avoided by people allergic to eggs. I have cauliflower, onions, and tomatoes I need to use up. Find a good recipe.

Entity-relationship (E-R) representation

Conceptual data schema (entity types and relationship types covered; *isa* is short for *is a*)

Entity types	**Relationship types**		

Name
Food product (FP)
Organism
  Person
Organism part
Substance
Form
Process
Purpose
Container
Good, commodity
  Portion
Legal entity
  Person
Money number
	Food product	<*hasName*>	Name

	Food product	<*isa*>	Food product
	Food product	<*comesFromSource*>	Organism
	Food product	<*comesFromPart*>	Organism part
	Food product	<*isExtractedSubstance*>	Substance
	Food product	<*isMadeFrom*>	Food product[1]
	Food product	<*hasIngredient*> <*withPurpose*>	Food product Purpose
	Food product	<*contains*>	Substance (omitted in the examples)
	Food product	<*processedBy*> <*withIntensity*> <*withPurpose*>	Process Intensity Purpose
	Food product	<*hasForm*>	Form
	Food product	<*packedIn*>	(container, volume)
	Organism	<*eat*>	Portion or Substance or FP
	<*buy/sell*> (Legal entity [seller], Legal entity [buyer], Good, Money no.)		

[1] Not used in the lecture examples, but in the reading on a food description language

Entity values can be seen from the E-R statements (FP0, FP1, etc, Plant, . . .)
< > around relationship names omitted for brevity

E-R statements [] inherited from one level above [[]] inherited from two levels above

FP0	***hasName***	**Food product**
FP1	***hasName***	**Vegetable product**
FP1	*isa*	FP0 Food product
FP1	*comesFromSource*	Plant
FP2	***hasName***	**Meat product**
FP2	*isa*	FP0 Food product
FP2	*comesFromSource*	Animal
FP2	*comesFromPart*	Carcass
FP3	***hasName***	**Egg product**
FP3	*isa*	FP0 Food product
FP3	*comesFromSource*	Animal
FP3	*comesFromPart*	Egg
FP4	***hasName***	**Prepared food**
FP4	*isa*	FP0 Food product
FP4	*processedBy*	*
FP5	***hasName***	**Soup**
FP5	*isa*	FP0 Prepared food
FP5	*processedBy*	*
FP5	*hasForm*	Liquid OR Semiliquid
FP11	***hasName***	**Diced carrots**
FP11	*isa*	FP1 Vegetable product
FP11	*comesFromSource*	Carrot plant
FP11	*comesFromPart*	Root
FP11	*hasForm*	Diced
FP12	***hasName***	**Cut green beans**
FP12	*isa*	FP1 Vegetable product
FP12	*comesFromSource*	Bean plant
FP12	*comesFromPart*	Immature fruit
FP12	*hasForm*	Cut
FP13	***hasName***	**Chicken broth**
FP13	*isa*	FP2 Meat product
FP13	*comesFromSource*	Chicken
FP13	*comesFromPart*	Meat
FP13	*comesFromPart*	Bones
FP13	*isExtractedSubstance*	{Fat, Protein, Flavor}
FP13	*processedBy*	Cooking
FP13	*hasForm*	Liquid
FP14	***hasName***	**Cubed cooked chicken**
FP14	*isa*	FP2 Meat product
FP14	*comesFromSource*	Chicken
FP14	*comesFromPart*	Skeletal meat
FP14	*processedBy*	Cooking
FP14	*hasForm*	Cubed
FP15	***hasName***	**Eggs**
FP15	*isa*	FP3 Egg product
FP15	*comesFromSource*	Chicken
[FP15	*comesFromPart*	Egg]
FP16	***hasName***	**Durum wheat flower**
FP16	*isa*	FP1 Vegetable product
FP16	*comesFromSource*	Durum Wheat
FP16	*comesFromPart*	Seed, kernel
FP16	*hasForm*	Ground
FP17	***hasName***	**Noodles**
FP17	*isa*	FP4 Prepared food
FP17	*hasIngredient*	FP16 Durum wheat flower
FP17	*hasIngredient*	FP15 Eggs
FP17	*processedBy*	Mixing
FP17	*processedBy*	Extruding
FP17	*processedBy*	Drying
FP18	***has name***	**Flavoring** (detail omitted)
FP19	***hasName***	**BHT** (detail omitted)
FP20	***hasName***	**Chicken noodle soup**
FP20	*isa*	FP5 Soup
FP20	*hasIngredient*	FP13 Chicken broth
FP20	*hasIngredient*	FP14 Cubed cooked chicken
FP20	*hasIngredient*	FP11 Diced carrots
FP20	*hasIngredient*	FP12 Cut green beans
FP20	*hasIngredient*	FP17 Noodles
FP20	*hasIngredient*	FP18 Flavoring
FP20	*hasIngredient*	FP19 BHT
	w/ purpose	Preservation
FP20	*processedBy*	Boiling
	w/ intensity	Fully cooked
	w/ *purpose*	Make edible, Preservation
FP20	*hasForm*	Liquid with solid chunks

FP21	**has name**	**Diced parsley** (statements not shown)
FP22	***hasName***	**Campbell's Chicken Noodle Soup**
FP22	*isa*	FP20 Chicken noodle soup
[FP22	*hasIngredient*	FP13 Chicken broth]
[FP22	*hasIngredient*	FP14 Cubed cooked chicken meat]
[FP22	*hasIngredient*	FP11 Diced carrots]
[FP22	*hasIngredient*	FP12 Cut green beans]
FP22	*hasIngredient*	FP21 Diced parsley
[FP22	*hasIngredient*	FP17 Noodles]
[FP22	*hasIngredient*	FP 18 Flavoring]
[FP22	*hasIngredient*	FP 19 BHT
	withPurpose	Preservation]
[FP22	*processedBy*	Boiling
	w/ intensity	Fully cooked
	w/ *purpose*	Make edible, Preservation]
FP22	*packedIn*	Steel can

[ ] inherited from one level above, [[ ]] inherited from two levels above

Portion-1	***isa portion of***	**FP22 Campbell's chicken noodle soup**
[[Portion-1	*hasIngredient*	FP13 Chicken broth]]
[[Portion-1	*hasIngredient*	FP14 Cubed cooked chicken meat]]
[[Portion-1	*hasIngredient*	FP11 Diced carrots]]
[[Portion-1	*hasIngredient*	FP12 Cut green beans]]
[Portion-1	*hasIngredient*	FP21 Diced parsley]
[[Portion-1	*hasIngredient*	FP17 Noodles]]
[[Portion-1	*hasIngredient*	FP18 Flavoring]]
[[Portion-1	*hasIngredient*	FP19 BHT *purpose* Preservation]]
[[Portion-1	*processedBy*	Sterilized by heat
	WithPurpose	{Make edible, Preservation}]]
[Portion-1	*packedIn*	(Steel can, 10 fl oz)]

FP23	***hasName***	**Frozen cut green beans**
FP23	*isa*	FP12 Cut green beans
[FP23	*comesFromSource*	Bean plant]
[FP23	*comesFromPart*	Immature fruit]
[FP23	*hasForm*	Cut]
FP23	*processedBy*	Freezing
FP23	*packedIn*	Carton

buy/sell **(Safeway, Fred, Portion-1, $1)**

Fred	*eats*	Portion-1
Chicken	*eats*	Hormone
Chicken	*eats*	FP24 Fish meal
FP24	*contains*	Mercury

Semantic network here

Some sample frames (not all data represented in frames)

A minimal frame: **Instance of a frame for buy/sell** (a relationship with four arguments)
A minimal frame gives information for one relationship.

SourceOfGoodOrService/ReceiverOfMoney:	Safeway
ReceiverOfGoodOrService/SourceOfMoney:	Fred
GoodOrService:	Portion-1 (a particular can of Campbell's chicken noodle soup)
MoneyAmount:	$1

All slots are essential; each value depends on all the others. The same information cannot be expressed in two separate frames.

Relationships can have two, three, or more pieces of information (called arguments) needed to make a complete statement. In the entity-relationship version, we wrote the same information as:

buy/sell (Safeway, Fred, Portion-1, $1)

The frame is just a different way of writing this, with specification of the role each piece of information plays. None of the pieces of information can be separated out and given separately. That is why the frame is called minimal.

Linguists specify for each verb or group of verbs the slots that must be filled in order to make a complete statement with the verb; they call this specification a **case frame**. So the above is a case frame for the verb buy and for the verb sell. (Both verbs describe the same transaction, just from different perspectives.)

An extended frame: **Instance of the food product frame for FP20**

An extended frame combines information from several relationships. Many of the pieces of information in an extended frame could be stored separately.
Frame slots are defined through relationship types.
Corresponding slot codes and names from the paper on food description language in []
In that paper, additional slots giving still more information about a food are defined.
In the sample frame, many slots are empty.

An extended frame: **Instance of the food product frame for FP20**		**Corresponding facet from the food classification in Reading**
Slot	**Value**	
FP20 *has name:*	**Chicken noodle soup**	
isa:	FP4 Soup	*A Product type*
Slots dealing with food origin		
comesFromSource:		*B1 Food source*
comesFromPart:		*B2 Part*
isExtractedSubstance:		
isMadeFrom:		
hasIngredient:	FP13 Chicken broth	*B3 Ingredient*
hasIngredient:	FP14 Cubed cooked chicken	*B3 Ingredient*
hasIngredient:	FP11 Diced carrots	*B3 Ingredient*
hasIngredient:	FP12 Cut green beans	*B3 Ingredient*
hasIngredient:	FP17 Noodles	*B3 Ingredient*
hasIngredient	FP18 Flavoring	*B3 Ingredient*
hasIngredient *withPurpose Preservation*	FP19 Preservative BHT	*D4 Method of preservation*
End food origin		
processedBy:	Boiling	*D2 Cooking method*
processedBy withIntensity:	Fully cooked	*D1 Degree of preparation*
processedBy withPurpose Preservation:	Sterilizing by heat (Boiling)	*D4 Method of preservation*
hasForm:	Liquid with solid particles	*C Phys. state, shape, form*
packedIn:		*E2 Container, wrapping*

Another instance of the food product frame, FP22
(inherits most of its information from FP20; inherited slots are not repeated)

An extended frame: **Instance of the food product frame for FP22**	
Slot	**Value**
FP22 *has name:*	**Campbell's Chicken Noodle Soup**
isa:	FP20 Chicken noodle soup
hasIngredient:	FP21 Diced parsley
packedIn:	Steel can

Think of this type of inheritance in the context of recipes.

3.2 Approaches to knowledge representation

Summary of concepts covered in examples

Entity-relationship approach

Semantic networks

Frames

Role of frames

Grid for data acquisition

Template for data output (for example, city data frame in Wikipedia)

Activation of all frame elements when one element is activated
(Seeing *parsley* may activate in a person's mind the whole frame for *Campbell's Chicken Noodle Soup*)

Types of frames

Minimal frames (DS term)

A minimal frame represents an n-ary relationship – each slot corresponds to one argument of the relation. No slot could be omitted without making the frame incomplete, that is, making at least one other slot value indeterminate.

Extended frame (DS term)

An extended frame includes additional slots that represent further relationships, usually binary relationships from the focal entity to other entities.

3.3 Some mechanisms in knowledge representation

Spreading activation

Hierarchical inheritance

Restrictions on values

Default values (for example, telephone area code in the database of a local charity)

Procedural attachments (procedures to be called when data are entered in the slot)

3.4 Some criteria for describing and evaluating knowledge representations (advanced)

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)

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

The material at the end of Lecture 1.2 is related.

January 26, 2011 Name (optional)

Free-write 2
Lectures 2.1 + 2.2. Nature of knowledge. Knowledge representation

- **Reflect** – what you learned, what was most important, what was most interesting, what was extraneous;
- **Ask questions** – ask for more explanation, how is a concept connected to other concepts, why is a concept important, how can it be applied, why is a reading important;
- Offer **critique and suggestions**;
- Say anything else you want to.

Over

Part 2. *February 2 - February 14*

Information retrieval: General principles and methods

Lecture 3.1 *February 2*

The structure of information systems (Organizing Information, Section 5.1)

Objectives	1 Know and understand the functional components of information systems and be able to use this framework • to analyze information systems; and • to integrate the subject matter from this and other courses. 2 Understand the wide variety of information systems .
Practical significance	• To **design, operate, or use an information system** or a specific function in it, you must understand the information system components, their inputs, output, and functioning. • To **take advantage of all available career opportunities**, you must understand the multitude of information systems and information environments in which the knowledge and skills acquired in CLIS can be applied. • The information system diagram provides a framework for organizing information from many courses.

In-class exercises (in preparation for Assignment 4)

Referring to Text Figures 5.1c, p. 47 and 5.6, p. 58 (both integrated on the following page)

1 Analyze a **Web search system** (like Google or Yahoo) as an information system.

2 Analyze a **special library** as an information system.
Question: How does interlibrary loan fit into the information system framework?

3 Additional question: Determine the percentage of resources allocated to each of the following functions (Refer to Section 2.6 of the text).

(1a) Assist the user in identifying relevant documents (intellectual access). (A user needing information is given a list of references to documents relevant for her problem or topic. She must then consult these documents and extract the information needed.)

(1b) Make available known documents (physical access). (A user requests specific documents, often documents found through 1a, and is given copies to keep or borrow.)

(2) Provide tailor-made packages of substantive data. (A user needs information on a certain topic and is given a report that contains just the information she needs, no more nor less. This report may be prepared by information center staff by extracting information from documents or it may be the result of a search of a substantive database.) (See Section 2.5 of the text for an explanation of "substantive data".)

What kind of data do you need to answer this question?

4 Analyze the **production and use of a textbook** in the information system framework.

Student questions on Organizing Information, Chapter 5

Note on diagram conventions

Denotes a file, data, inputs and outputs. Could be a group of people and their problems as an input

Denotes a process that transforms inputs into outputs

Sequence of processes and files, flow of data

Control of processes or file organization

Combined information systems diagram here

Lecture 3.2 ***February 2***

Objectives and performance measures for information systems

(Organizing. Information, Ch. 8)

Objectives	1 Understand the purpose and objectives of IR systems so that you can examine the functioning of the individual IR system components in light of these objectives. 2 Understand both the importance and the difficulty of defining suitable measures of information system performance and of applying such measures to actual systems.
Practical significance	A clear understanding of objectives and evaluation criteria for both individual searches and for an information system as a whole is important for the following: 1 Conducting individual searches. • Determining user requirements. • Selecting an information system (database and search system) that can be expected to meet these requirements. • Evaluating search results and determining when to stop searching. • Determining the amount of resources that should be allocated to a search. 2 Selecting information systems to be acquired, including reference tools and online databases (see Lecture 1.1 for a list of types). (Acquisition includes leasing or contracting for use, as well as training in the use of the system.) 3 Designing information systems or communicating requirements to systems analysts. See Text Chapter 8, Introduction, and Section 8.5 for elaboration.

Discussion questions: see next page

Discussion questions

1 Consider the definition of relevance and of performance measures in general in the context of an information system with data on the structure of a nuclear power plant to be used in case of malfunctions. The system gives detailed information about all components down to the last pipe and valve, their functions and interrelationships.

2 Consider performance measures for the following information system. The purpose of the information system is to assist in solving crimes. The system stores reports of past crimes — both solved and unsolved — and indexes them by various features of the modus operandi. To use the system, the detective formulates a query based on the features of the unsolved crime. The system provides reports of similar crimes; these might assist in solving the crime in question.

3 Consider a Web search service (such as AltaVista or Lycos) that produces ranked retrieval output. Picture two users. User 1 needs a quick answer to a question, and user 2 needs a comprehensive list of materials (for example, a listing of all classifications schemes and thesauri available on the Web). What performance measures would be appropriate for each type of user?

Organizing Information, Chapter 8 review, especially

Deriving performance measures, Figures 8.2 and 8.3

The concept of relevance

Relevance criteria of teachers selecting oral history materials

Relevant to teaching content and method

- **Relationship to theme**
- **As part of broader curriculum**
  - Relates to other schoolwork
  - Variety for the classroom
  - Vocabulary
- **Characteristics of oral history interviews**
  - Flow of interview
  - Expressive power
  - Language & verbal expression
  - Diction
  - Nonverbal communication
- **Characteristics of the story**
  - Positive message for students
  - Role of interviewee in Holocaust events
- **Relationship of story to student**
  - Students connect with passage
  - Students identify with interviewee
  - Radical difference from students' reality

Represents different populations

- Race
- Age of interviewee during Holocaust events

Appropriateness

- Developmental appropriateness
- Acceptability to stakeholders

Technical production quality

Length-to-contribution ratio

Topical relevance for scholars

Types of topical relevance	
Topic: Food in Auschwitz	
Relevance type	Example
"Classical relevance" (TREC definition)	
• Provides direct evidence	Describes types of food and portions served
• Provides indirect/circumstantial evidence	Describes undernourished people
Additional relevance types	
• Provides context	(1) Reports on intensity of manual labor (2) Availability of food in the area around the camp
• Useful as a basis for comparison	Food situation in a different camp
Pointer relevance	
• Provides pointer to a source of information (The information pointed to can be relevant in any of the ways listed above)	Mention of a study on the subject

TREC = **T**ext **RE**trieval **C**onference
A yearly competition of information retrieval systems performing specified retrieval tasks on a given test collection held at NIST (National Institute of Standards and Technology)

CLEF = **C**ross-**L**anguage **E**valuation **F**orum
The corresponding activity in Europe. Uses the MALACH speech retrieval test collection for one task.

February 2, 2011 Name (optional)

Free-write 3

Lecture 3.1. The structure of information systems

Lecture 3.2. Objectives of information systems

- **Reflect** – what you learned, what was most important, what was most interesting, what was extraneous;
- **Ask questions** – ask for more explanation, how is a concept connected to other concepts, why is a concept important, how can it be applied, why is a reading important;
- Offer **critique and suggestions**;
- Say anything else you want to.

Over

Lecture 4.1 ***February 9***

An integrated information structure model

Objectives	1 Understand a general model of information retrieval; 2 Be able to analyze specific systems and information retrieval operations in terms of this general model; 3 Integrate knowledge across types of information systems and developing an overall vision of retrieval possibilities.
Practical significance	This knowledge will enable you • to use existing systems in new and imaginative ways, in particular, to use several different systems in synergistic ways; • to design new systems with increased power. Note: Pay attention to the first bullet; it will make you a better searcher. You may never be able to use a unified integrated system of the kind described. But you can use existing systems in combination to achieve improved search results enabled by the way of thinking presented here. In other words, you can build your own "virtual" integrated information structure whenever a search requires it.

This lecture will present the material from the reading

Design of an integrated information structure interface

as if it was a presentation at a conference. Please read the Prologue and p. 1 - 13 beforehand; these sections give examples. The lecture does not assume that you have read the remainder of this reading; rather, the reading is a back-up reference.

Design of an integrated information structure interface

Part 1. Basic structure and search commands

Lecture 4.2 ***February 9***

Conceptual data schemas and input, storage, and output/presentation formats

(Organizing Information, Sections 9.1, 9.2, 9.4, and 9.5)

Objectives	1 **Be able to analyze or design the conceptual data schema** of an information system • analyze the conceptual data schema underlying an information system; • judge the adequacy of this schema with respect to the queries to be answered; • use the knowledge of the schema to exploit fully the possibilities of obtaining answers from the information system; • design a conceptual data schema for an information system based on user requirements. 2 **Be able to analyze and design the input formats and output formats** used to interact with an information system: • input formats that make data entry complete, error-free, and easy • output formats (for reports, such as recurring bibliographies, or the display of search results) that contain all the information needed (and no more) in an easy-to-read form.
Practical significance	1 **For designing information systems:** The success of any information system depends vitally on the completeness of the information included. The conceptual data schema determines what information can be included in the system and what information is elicited from the people that enter data into the system. Input and output formats determine how easy it is to interact with the system. 2 **For using information systems (including reference tools):** To get the most out of an information system in terms of being able to do different types of searches, you need to know its conceptual data schema. To select the appropriate information system, you need to be familiar with the conceptual data schemas of many information systems. To do the kind of power search that draws on multiple information systems simultaneously requires even more knowledge of conceptual data schemas.

Schema Arrangement of parts in some order, showing interrelationships.

This topic is closely related to document structure design, to be discussed in Lectures 5.2-6.2.

In-class exercise

Developing the conceptual data schema for the information system of a large computer users' group

Next page

In-class exercise

Developing the conceptual data schema for the information system of a large computer users' group (such as the Washington Apple Pie)

A computer users' group has the purpose of helping members to better use their computers.

Some functions of a computer users group

- a library for members to use
- a newsletter with articles and product reviews
- special interest groups (hold meetings, have a chair)
- a group purchase program
- a list of experts on specific subjects that have agreed to be on call to answer member questions

Sample questions / outputs from the database

Entity types	Relationship types

Uses of the different types of information in an information system

A type of information (a fact type, see Lecture 1.2), as defined by a relationship type, may be used for one or more of the following functions.

- **Retrieval, drawing inferences, statistical analysis**

  Example: From drug prescription expert system

  Disease <*treated with*> Drug

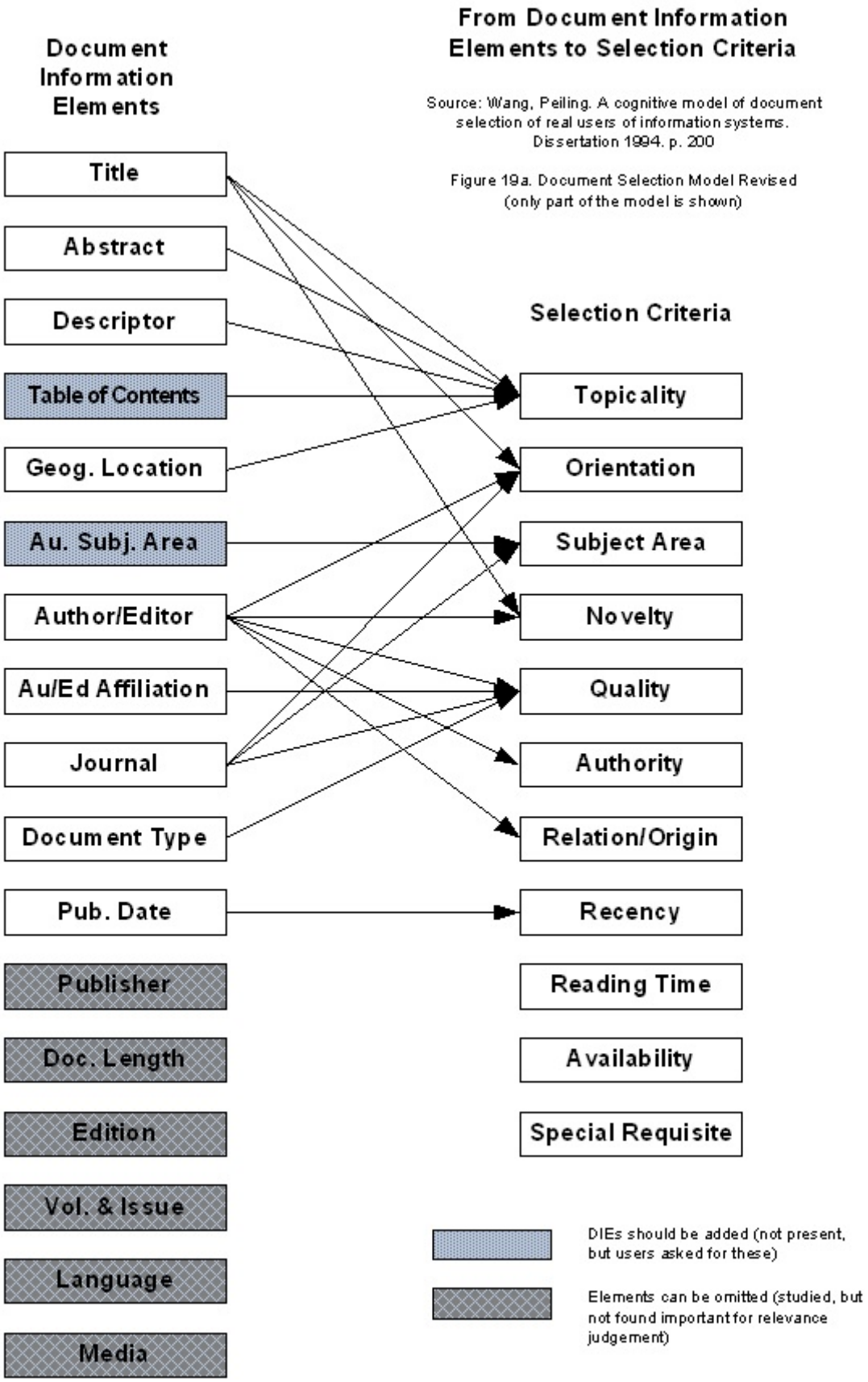
  This piece of data is used for
  - plain retrieval of medical knowledge;
  - inference in conjunction with patient data.

- **Arranging retrieval output**
  - Example: Arranging a long list of document records retrieved from an OPAC (Online Public Access Catalog) by call number
  - Example: Arranging output from a Web search by URL (Uniform Resource Locator), which would bring pieces of one Web document that consists of several pages together in the output list

- **Providing information to the user**, either the substantive data sought or information about a document that enables the user to judge the relevance of the document.

The conceptual data schema designer must weigh the cost for including a type of information against the benefit in terms of these three functions.

As an example, look at Wang's list of "Document Information Elements" users considered in document selection or wished to have available (next page). These results should be used as a guideline in systems design when deciding what information to include in the system and what information to present to users in the output format. Users did consider information elements that are linked to the document indirectly, such as the subject area of the document author. As will be further elaborated below, the information system must assemble all this information about a document, possibly obtaining information from other databases, such as a database about persons.


From Document Information Elements to Selection Criteria
Source: Wang, Peiling. A cognitive model of document selection of real users of information systems. Dissertation 1994. p. 200
Figure 19a. Document Selection Model Revised (only part of the model is shown)
Document Information Elements
Title
Abstract
Descriptor
Table of Contents
Geog. Location
Au. Subj. Area
Author/Editor
Au/Ed Affiliation
Journal
Document Type
Pub. Date
Publisher
Doc. Length
Edition
Vol. & Issue
Language
Media
Selection Criteria
Topicality
Orientation
Subject Area
Novelty
Quality
Authority
Relation/Origin
Recency
Reading Time
Availability
Special Requisite
DIEs should be added (not present, but users asked for these)
Elements can be omitted (studied, but not found important for relevance judgement)

From conceptual data schema to records (Organizing Information, Section 9.2)

record format	• A record is an assembly of information about a given entity, such as an event, a person, or a document for input, storage, communication, or display. • The **record format** determines how the different pieces of information are arranged in the record. • A record is a simple frame. Slot in a frame = data field in a record. • The evaluation criteria for schemes of knowledge representation (Lecture 2.2, Section 3.4) apply to records; see also Organizing Information, Section 9.3. • Many records are extended frames: they incorporate many binary statements that link the focal entity of the record to another entity. Each statement could stand on its own. The presentation of these data in a record is more concise and may be more intuitive and more easily grasped than a series of statements. See the MARC record format (facing page) and the examples in text Chapter 9.
input record input format	The best way for eliciting input from the system operator (for example, a cataloger) is often an **input record** , a form with blanks (slots) to fill in.
storage record, storage format	Some systems store data internally in tables (relational database), where information about a given entity may be distributed over several tables. Other systems store data internally in records or frames, assembling all information about one entity, e.g., a book, in one place (but making it more difficult to assemble information about entities of other types, e.g., persons).
communication record	A record in a common **communication format** to transfer data from one system to another. Each system may use its own internal format. Examples: MARC, Z39.50
output/ display record, display format	To present information about an entity in a format easily understood by the user, the information system must assemble the desired types of information into an output record. The information elements in the output record may be linked to the entity directly (for example, the direct link between a document and the person who authored it) or indirectly (for example, the indirect link between a document and the organization with which the author is affiliated).

MARC format	Opposite is a sample record format, the MARC (MAchine-Readable Catalog) format. (MARC was developed by the Library of Congress starting in 1962 for the interchange of bibliographic data and has become a widely used standard). For each data field, the corresponding statement template (relationship type plus entity types related) is given. A more complete list of MARC fields is found with the description of the model catalog in the general readings.

From relationship types to data fields in a MARC record for documents. (simplified)

Relationship type	Corresponding MARC field
Document <*authoredBy*> **Person** (who is main entry)	100 Main Entry-Personal Name
Document <*emanatedFrom*> **Organization**	110 Main Entry-Corporate Name
Document <*emanatedFrom*> **Meeting**	111 Main Entry-Meeting Name
Document <*hasTitle*> **Text** (if Title is main entry)	130 Main Entry-Uniform Title
Document <*hasTitle*> **Text**	245 Title Statement
Document <*publishedBy*> **Organization** this Organization <*locatedIn* **Place** (chain) Document <*publishedIn*> **Date**, ... (distinguished by using subfields)	260 Publication, Distribution, etc. (Imprint)
Document <*partOf*> **Document** (which is a Series)	490 Series Statement
Document <*dealsWith*> **Person**	600 Subject Added Entry-Personal Name
Document <*dealsWith*> **Organization**	610 Subject Added Entry-Corporate Name
Document <*dealsWith*> **Meeting**	611 Subject Added Entry-Meeting Name
Document <*dealsWith*> **Document**	630 Subject Added Entry-Uniform Title
Document *dealsWith*> **Topic**	650 Subject Added Entry-Topical Term
Document <*dealsWith*> **Place**	651 Subject Added Entry-Geographic Name
Document <*authoredBy*> **Person**	700 Added Entry--Personal Name
Document <*emanateFrom*> **Organization**	710 Added Entry-Corporate Name
Document <*emanateFrom*> **Meeting**	711 Added Entry-Meeting Name
Document <*hasTitle*> **Text**	730 Added Entry-Uniform Title
Document <*heldBy*> **Organization**	850 Holding Institution

The MARC record is an extended frame; it incorporates many binary statements that link a document to another entity. Each statement could stand on its own.

February 9, 2011 Name (optional)

Free-write 4

Lecture 4.1. An integrated information structure model

Lecture 4.2. Data schemas and formats

- **Reflect** – what you learned, what was most important, what was most interesting, what was extraneous;
- Ask questions – ask for more explanation, how is a concept connected to other concepts, why is a concept important, how can it be applied, why is a reading important;
- Offer **critique and suggestions**;
- Say anything else you want to.

Over

Lecture 5.1 ***February 16***

Access to information: data structure and search modes

(Organizing Information, Chapters 10 and 11) (85 min.)

➜LIS 518 Reference Sources and Services

Objectives	1 Understand the purpose of a data structure as answering questions with retrieval as a special case. 2 Understand the basic principle of searching: use all available evidence to predict the degree of relevance of some entity Ranked retrieval and plain Boolean retrieval as special cases. 3 Be able to formulate simple Boolean queries. 4 Be able to analyze the storage structures (tables, record formats) of an information system and design simple storage structures. 5 Be able to analyze data access structures (indexes) in an information system and use the understanding gained for efficient searching. 6 Be able to design simple data structures for access. 7 Be able to apply the principle of hierarchical inheritance to achieving more compact storage.
Practical significance	• When searching for X, use a reference tool where X can be found in the index. • Store data with minimal redundance by using hierarchical inheritance. (OCLC and other bibliographic databases are tremendously redundant since they do not use a data structure that exploits hierarchical inheritance.)

Outline

1 Retrieval as prediction.

2 Review of Boolean retrieval

3 Ranking documents by expected relevance

4 Search modes and data structures

4.1 Review of Organizing Information, Chapter 11

4.2 Further elaboration of data structures

4.3 Using hierarchical inheritance for efficient storage

1 Retrieval as prediction.

Query formulation: find good predictive clues

Retrieving a document or other entity is predicting that it is relevant to the problem to be solved at least to a certain degree. The challenge in formulating a query is to find the clues that predict that a document or other entity is relevant:

For documents:	What clues can predict that the document will be helpful in solving the problem at hand?
For persons to fill a job:	What clues can predict that a person will do well in the job?

Finding the right clues requires knowledge and may involve some guesswork:

- When searching for documents using free-text retrieval, the searcher must determine what words and expressions the author of a relevant document may have used in the title, the abstract, and the full text; this requires knowledge of how language is being used both in general and by specific schools and even individual authors.
- When using descriptors assigned by an indexer, the searcher must determine what descriptors an indexer would have assigned to relevant documents; this requires knowledge of the index language, the indexing instructions, and the actual indexing practices. (Request-oriented indexing, to be discussed later, seeks to increase the probability that descriptors corresponding to user needs are included in the indexing language and assigned correctly in indexing.)

Of course, free-text terms and assigned subject descriptors are only two kinds of evidence. Many other clues can be considered, such as publication date, reputation of the author or the author's institution, reputation of the journal or publisher, etc.

2 Review of Boolean retrieval (Organizing Information, Chapter 10)

3 Ranking documents by expected relevance (as in Google)

Boolean retrieval: YES or NO – division of all documents in the system into **two classes**

A document either scores 1 and is retrieved or it scores 0 and is rejected:

class 1: retrieved - expected to be relevant
class 2: rejected - expected not to be relevant

Using three queries to get YES, MAYBE, NO – division of all documents into **three classes**

Problems with formulation of queries, especially if interactive retrieval not possible. Consider the following query formulation consisting of four descriptors

A Simulation AND *B* Traffic flow AND *C* Passengers AND *D* Rail transport

Perhaps documents that contain any three of the four descriptors might be of some interest; could run the broader query (ABC OR ABD OR ACD OR BCD) in parallel:

Class 1: retrieved in narrow search ABCD –	expected to be clearly relevant
Class 2: retrieved in broad search –	expected to be somewhat relevant
Class 3: not retrieved at all –	expected to be not relevant

More refined ranking by expected relevance – continuous scale

Compute for each document a **quantitative measure of expected relevance** to the given search request. Instead of having 3 classes of documents, we then get a **list of documents ranked according to expected relevance**. (In many systems the ranking is poor and does not approximate the user's intuition.) Measure is computed as the nearness or **similarity** between query formulation and document representation, based on the number of descriptors they have in common. Different formulas are possible; very simple: percentage of query descriptors found in document record. For each formula:

(a) crude form (uses no knowledge of concept term relationships) and

(b) knowledge-based form (uses knowledge of concept and term relationships, for example to match a query term with a synonym or narrower term in the document).

Problems of OR descriptor combinations, as in the following query formulation:

Traffic congestion AND Terminals AND Air Traffic AND (New York
OR
Boston
OR
Washington)

It does not matter if only New York or both New York and Boston occur in the document description. The query has four **conceptual components**; these, rather than the terms, should be the basis for comparing the query with documents.

Query formulation Q and document representations D1 - D4

Q	**B2** Rail transport	AND	**E3** Traffic stations	AND	**J3** Passenger transport	AND **U15** US		
D1	**B2** Rail transport	**D1** Air transport	**E3** Traffic stations		**J3** Passenger transport	**U15** US	**U20** Europe	**Q24** Traffic simulation
D2	**B2.7** Local rail transport		**E1** Traffic facilities		**J3** Passenger transport	**U15** US		
D3	**B1** Ground transport		**E6** Vehicles		**J3** Passenger transport	**U15** US		
D4	**B2** Rail transport		**E3** traffic stations		**J3** Passenger transport	**U15** US		

Formulas for computing expected relevance

Base formulas **1** and **2**, descriptor matching rules **a** and **b**, gives four formulas: **1a**, **1b**, **2a**, **2b**.

Base formula 1: R = # of descriptors in common/ # of query descriptors

Base formula 2: R = # of descriptors in common / (# of query descriptors + # of doc. descriptors)

Matching rule a: (crude) **Exact descriptor match**: A query descriptor produces a match only if the document representation contains exactly the same descriptor

Matching rule b: (knowledge-based) **Hierarchically expanded match**: A query descriptor produces matches as shown in the following examples:

Query descriptor	**Document descriptor**		**Match value**
B2 Rail transport	**B2** Rail transport	Same	1
	B2.7 Local rail transport	Narrower	1
	B1 Ground transport	Broader	0.5

Note: The numbers in the column "Match value" are set arbitrarily for this exercise. One might count a narrower descriptor as 0.75 of a match, for example.

In-class exercise: **Ranking of retrieved documents**

Purpose	1. To give you a better "feel" of how formulas for the computation of expected relevance work and what a rank list of documents looks like. 2. To have you compare the effectiveness of four formulas (two base formulas, each applied with two matching rules).
Task	Given are a query formulation and four document representations (descriptors assigned to the documents) and four formulas for the computation of expected relevance. The formulas are deliberately very simple; many more complex formulas are being used. 1. Using your own judgement, rank the documents 1-4 by their relevance to the query. 2. Compute for each document the coefficient of expected relevance according to four different formulas and list the documents in rank order by decreasing expected relevance. 3. Compare the four rankings with your intuitive ranking. State which is better. Briefly state why one formula works better than the other.

Results

Expected relevance score for the query Q

Document	**Formula**			
	1a	**1b**	**2a**	**2b**

Ranking

	Intuitive	**Formula**			
		1a	**1b**	**2a**	**2b**
1					
2					
3					
4					

AltaVista uses a different ranking method considering three elements; in order of importance

- **The rarity of query words**: A document gets more weight for a rare word than for a word occurring in many documents. Extremely common words are ignored for ranking.
- **The absolute and relative location of query words**: A document gets more weight for a query word in the HTML title, in an HTML META tag, or in the top portion of the document than for a query word someplace down in the body of the document. Also, a document gets more weight if two query words occur close together.
- **The frequency of query words in the document**: Does the word occur once or more than once (2, 3, 4, etc. are all treated equal).

4 Search modes and data structures

To execute a search, a retrieval system must operate on stored data. The problem is to devise a data structure and a search process that makes retrieval fast. We will discuss data structures for Boolean retrieval and data structure in semantic networks.

4.1 Review the data structures described in Chapter 11

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.

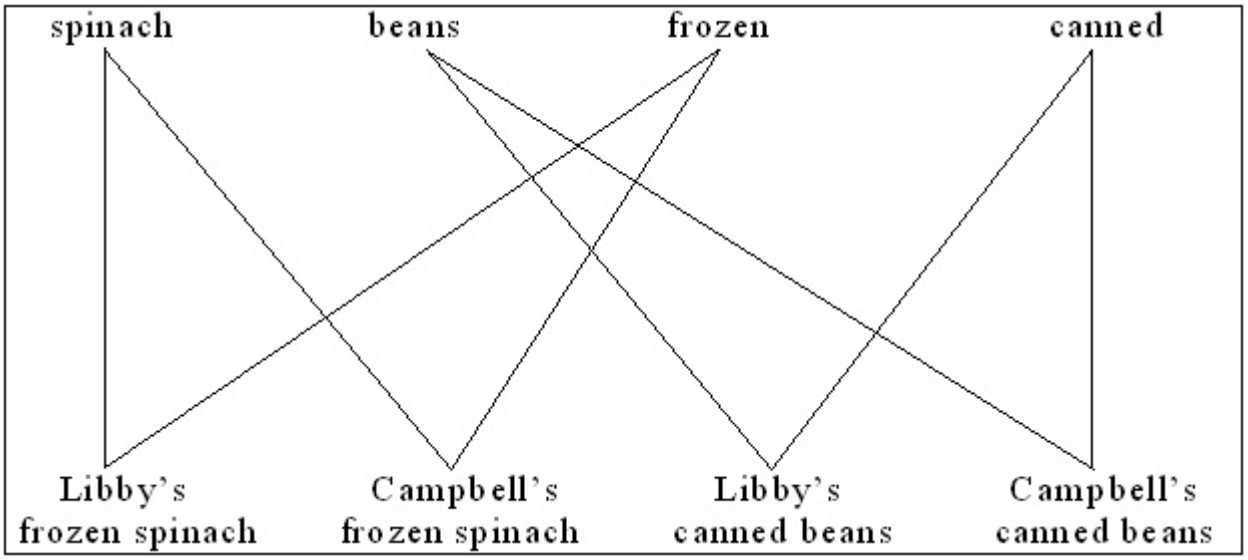
Object-oriented databases are based on frames with hierarchical inheritance (see Lecture 2.2). They are closer to the record model than to the relational model.

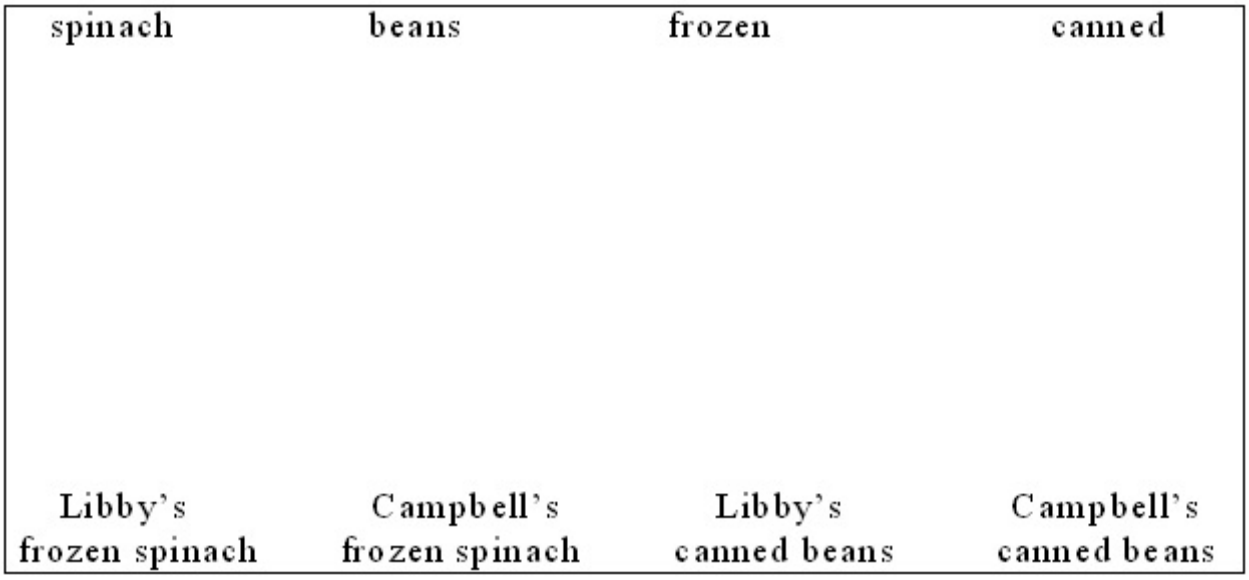
Searching printed indexes vs. searching by computer.

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*.)

4.3 In-class exercise: Using hierarchical inheritance for efficient storage

4.3.1 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.





Original database	Restructured database
Food product 1. Libby's frozen spinach *Food*: spinach *Preservation*: frozen *Manufacturer*: Libby	
Food product 2. Campbell's frozen spinach *Food*: spinach *Preservation*: frozen *Manufacturer*: Campbell	
Food product 3. Libby's canned beans *Food*: beans *Preservation*: canned *Manufacturer*: Libby	
Food product 4. Campbell's canned beans *Food*: beans *Preservation*: canned *Manufacturer*: Campbell	

Note: The restructured database has more records, but they are much shorter

A good example for hierarchical inheritance is a cookbook which may have a basic recipe for potato soup and then many variations that say,

> "Use the recipe for potato soup but add X ingredient."

4.3.2 Look at the six bibliographic records from OCLC on the next page.
How could the bibliographic data be restructured for more efficient storage?

Reorganize these bibliographic records, using hierarchical inheritance for efficient storage

Document 1

100 1 Mager, Robert Frank, $d 1923-
245 10 Developing attitude toward learning / $c Robert F. Mager.
260 Belmont, Calif. : $b Fearon/Pitman Publishers, $c c1968.
300 vii, 104 p. ; $c 22 cm.
650 0 Interaction analysis in education.
650 0 Learning, Psychology of.
650 0 Group work in education.
650 0 Classroom management.

Document 2

100 1 Candelora, D[eborah] M.
245 10 Hands-on technology program $h [computer file]
246 HOT program
260 [Ramsey, NJ]: $b [Galaxy Networks], $c 1996-
500 Title from the home page HTML title
500 Material copyrighted by D. M. Candelora
500 Accessed 1998 Feb. 2
650 0 Science - Study and teaching (Elementary) - Aids and devices
650 0 Science - Experiments
650 0 Computers - Study and teaching (Elementary) - Aids and devices
650 0 Mathematics - Study and teaching (Elementary) - Aids and devices
650 0 Learning by discovery
650 0 Active learning
856 4 $u www.galaxy.net/~k12/ $n Ramsey, NJ

Document 3

100 1 Mager, Robert Frank, $d 1923-
245 10 Developing attitude toward learning : $b or SMATs 'n' SMUTS / $c Robert F. Mager.
250 2nd ed.
260 London : $b Kogan Page, $c 1991, c1990.
300 116 p. ; $c 23 cm.
650 0 Interaction analysis in education.
650 0 Learning, Psychology of.
650 0 Group work in education.
650 0 Classroom management.
650 0 Students $a Motivation

Document 4

100 1 Conant, James Bryant, $d 1893-1978
245 10 The comprehensive high school; $b a second report to interested citizens $c by James B. Conant.
260 New York, $b McGraw-Hill $c [1967]
300 vi, 95 p. $c 21 cm.
650 0 Education, Secondary
650 0 Education $z U.S. $y 1945-
650 0 Comprehensive High Schools $z U.S. $y 1945

Document 5

100 1 Mager, Robert Frank, $d 1923-
245 10 Developing attitude toward learning, $b or, SMATS "n" SMUTS / $c Robert F. Mager
250 2nd ed.
260 Belmont, Calif. : $b David S. Lake, $c c1984.
300 x, 116 p. : $b ill. ; $c 24 cm.
490 1 The Mager library
500 Rev. ed. of: Developing attitude toward learning. 1968.
650 0 Interaction analysis in education.
650 0 Learning, Psychology of.
650 0 Group work in education.
650 0 Classroom management.

Document 6

100 1 Mager, Robert Frank, $d 1923-
240 10 Developing attitude toward learning. $l Spanish
245 10 Desarrollo de actitudes hacia la ensenanza / $c Robert F. Mager.
260 Barcelona : $b Martacinez Roca, $c c1985.
300 158 p. : $b ill. ; $c 19 cm.
650 0 Interaction analysis in education.
650 0 Learning, Psychology of.
650 0 Group work in education.
650 0 Classroom management.

Part 3 *February 16 - March 2*

The nature, design, and management of documents and records

➜LIS 506 Information Technology

Lectures 5.2a - 6.2b ***February 16 - February 23***

Document function, structure, analysis, and design (No text chapter)

Scope	This part of the course requires a clarification of the scope, particularly what is meant by "document," and how this topic is approached by many disciplines from many angles.
Broad definition of "document"	Text has been defined as: "Any passage, spoken or written, of whatever length, that [forms] a unified whole." (Halliday) **Document: any presentation of information in any form:** • written or spoken text, still or moving images, or music and sound (a multimedia document combines all of these) • in any medium – print, computer screen, TV, radio, etc.
Disciplines/ fields dealing with information presentation / document design	Text linguistics, discourse analysis Rhetoric, English composition Document design, including Web design Information architecture User interface design Instructional design Advertising design Graphics design, including, for example, guidelines for transparencies Formatting documents for interpretation by computer programs

Objectives and practical significance for Lectures 5.2a - 6.2b are given on the following pages.

Discussion question

Design question for multimedia documents:
What combination is best for given communicative task, for example teaching a concept, persuading people to do something or quit doing something, or instructing in the use of a device? Generalization of text, which refers only to language.

Objectives, Lectures 5.2a-6.2b (inherit down to each lecture)	1 (With lectures 1.1-2.2) Understand the principles for **developing a good conceptual structure for a body of knowledge** and representing that structure for human understanding and for machine processing. 2 **Understand the importance of document structure and presentation** for the efficient transmission of information. 3 Be able to **analyze the structure and design of existing documents** as one aspect in assessing the quality and usability of the document. 4 **Appreciate the importance of good document design** and what is needed to achieve good design as a basis for further study.

Practical significance	**Lectures 5.2a - 6.2b** especially in the context of the Web. Inherit down to each lecture
General	• Well-designed information presentation facilitates assimilation and understanding of information and helps people cope with the ever-increasing amounts of information needed to function in a modern society.
Document production	• Assisting in the **manual production of documents**, a very important problem particularly in large organizations such as the World Bank. Also important in helping students studying English composition. Note: On a day-to-day level, **most information specialists** (including librarians) must **produce documents** all the time as seen from the list below. • **Automatic** or computer-assisted **generation of text** and documents • Devising **guidelines for document design** (as in a text on English composition)
Document retrieval	• **Structure for storing information and selecting specific document parts for retrieval and display**. Many organizations now organize their documents into large text databases available and searchable on an intranet. • Devising systems that help users to find just the right documents or portions of documents for a given purpose • Devising computer systems that can be the users agent for assessing relevance, assimilate information from a document, abstract or index a document
Document analysis and assimilation	• **Understanding how people process documents** (assess relevance, assimilate information from a document, abstract or index a document) • Serving as the user's agent in judging the relevance and appropriateness of a document to the user's situation (background and purpose) • **Guidance in the analysis of documents.** Reading and evaluating documents, for example, scientific articles or news stories, is much easier if one understands their structure. Document processing by human indexers or machine indexing systems is based on document structure.

Documents produced by information specialists

- Presentation of search results (bibliography or substantive data)
- New acquisitions list
- Guide to the library, instructional materials
- Guide to information on . . .
- Promotional materials
- Library newsletter
- Meeting notes
- Classification schemes and thesauri

Library Web site

Increasingly, libraries set up Web sites for use by their patrons; these Web sites include (but are not limited to) the kind of information listed above. (Hint: other libraries' Web servers are very useful information sources, for example www.lib.uchicago.edu/LibInfo/)

Outline for Lectures 5.2a - 6.2b

Lecture 5.2a. Knowledge (re)presentation in text and images. Text linguistics (35 min.)

Micro

Lecture 5.2b. Text analysis overview and examples (30 min.)

Lecture 6.1a. Natural language processing. Syntactic and semantic parsing (50 min.)

Macro

Lecture 6.1b. Document macrostructure and inter-document relationships (40 min.)

Document macrostructure. Document templates

Hypermedia

Inter-document (inter-textual) relationships

Lecture 6.2a Document design. Information design (45 m)

Lecture 6.2b. Formatting documents for interpretation by computer programs. Document markup languages (15 min, more fully covered in ➜ 506)

Lecture 5.2a (35 minutes) ***February 16***

Knowledge (re)presentation in text and images. Text linguistics

Objectives (In addition to the objectives inherited from 5.2-6.2)	1 Be aware of the different types of text and the communication purposes they serve. 2 Understand text coherence and cohesion and their role in text understanding by people and by computer programs. 3 In particular, understand the problems of anaphoric reference.
Practical significance	• Matching text type with user needs is important in answering questions. This has implications for cataloging. ➜ LIS 518 Reference Sources and Services: Judging appropriateness • Understanding text coherence and text cohesion is important for evaluating texts and for good writing. • Knowing about anaphoric references points out limitations in using adjacency commands in free-text searching and the need for linguistic processing to overcome these limitations. ➜ LIS 518 Reference Sources and Services: Query formulation

Context	Lecture 2.2 focused on knowledge representation in computer systems and in the brain, so did Lecture 4.2 on data schemas and formats. This sequence of lectures focuses on external knowledge representation with the purpose of helping people to assimilate knowledge. But there is much overlap and interrelationship: • Semantic networks are used as external knowledge representation in the form of concept maps. • Record formats can be useful for external representation. • Internal representation often serves the main purpose of letting a computer program create multiple external knowledge representations (Lecture 6.2b on XML).

over

Outline

1 Functions of documents
2 Document or text types
3 Text structure: cohesion and coherence

Discussion questions

1 In the context of a hypertext system, what is a text?
   (1) Each individual segment or
   (2) the total sequence of text segments (and perhaps images, etc.) the user/reader constructs in the interaction with the system?

2 In the second example from Crombie, what rhetorical devices are used to express coherence relationships? How could knowledge of such devices assist in text analysis?

3 How could the principles of text structure be applied to the structure of images?

1 Document analysis. Document functions

Perspectives in document analysis	

	• Internal document structure
	• External context or communicative situation of the document:
	  • the creator (writer/speaker/designer) and the peruser (reader/listener/viewer) and their relationship
	  • the function the document fulfills

Document functions	
Informing, educating	Enable the reader/viewer to construct, reconstruct, or otherwise update his or her own mental image, to make sense out of the message presented • Reporting results of research or scholarly endeavors. Describing objects or events • Educating: helping the reader understand a new field or topic • Providing small pieces of information quickly as needed • Reporting the discourse at a meeting and its results (minutes) • Laying out a plan of action • Giving instructions/prescriptions/orders (transmitting norms) (also below) • Informing documents may be designed for reading in context or to enable the reader/viewer to quickly locate a needed piece of information.
Instructing, persuading	Creating or changing beliefs, attitudes, or behaviors. Persuasion. • Giving instructions/prescriptions/orders (transmitting norms) (repeated here) • Persuading somebody to do something (vote for a person or issue, buy something) • Assisting in the treatment of mental or behavioral disorders by enabling the reader/viewer/listener to construct or reconstruct emotional/attitudinal structures, including self-image (bibliotherapy)
Entertaining	Providing entertainment or enjoyment A document can serve multiple functions, especially it can entertain while it educates (and educate the better for it, “edutainment”)

2 Document or text types (summarized from Beaugrande)

The type of a text is determined by its internal structure and the communicative situation, especially the function of text. Major text types are listed and defined briefly below. The viewer/reader/listener can process a document more efficiently if s/he knows the its type
➜LIS 518 Reference Sources and Services: Matching text type with user's purpose.

Major text types
(There are many, many text types; this list is just the tip of the iceberg.)

Type	Examples
Descriptive:	Review article; newspaper article; dictionary definition
Argumentative:	Logical proof; Legal argument
Didactic:	Textbook
Narrative:	Fairy tale; Letter
Conversational:	Reference interview
Literary:	Prose (e.g., Novel); Poetry (e.g., Limerick)
Scientific:	Research study report

The classification of text types parallels roughly the classification of functions, but there is not a perfect one-to-one correspondence; for example, a poem may educate or persuade.

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*.
Argumentative	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 established 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.)
Conversational	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.

3 Text structure: cohesion and coherence

Cohesion and coherence as the key devices in determining the internal structure of texts

Cohesion (Grammatical)	Elements of a text are properly linked grammatically: Properly structured sentences Inter-sentence relationships
Anaphoric reference	Use of a pronoun or general noun to refer to an object, action, or thought previously identified in the text. Example: President Bill Clinton gave a speech at Concord High School. **He** emphasized the need for crime prevention and for the restoration of family values. **This** made the Republicans angry. **They** accused **him** of stealing **their** issues. Meanwhile, Patrick Buchanan addressed a rally in Manchester. **He** hammered away at the theme that the jobs of American workers must be protected from low-wage foreign competition. **This theme** has propelled him to the front in the polls. Importance in the context of information systems: - Detecting the relationships in a text. - Proximity searching.
Coherence (Lexical-semantic)	Does the document make sense? Does an argument proceed in a logical fashion? • If a section requires background knowledge the reader cannot be expected to possess, does the document provide this background knowledge before the reader gets to that section? • Are there proper transitions to prepare the reader's mind set for new information? • Do illustrations fit with the text? • In a conversation: Is a question properly answered? Does the contribution of one participant build on previous contributions? Importance in the context of information systems: Design hypermedia systems that support the user in constructing coherent documents Coherence related to document/text macrostructure

Incohesive text

President Bill Clinton gave a speech at Concord High School. **They** emphasized the need for crime prevention and for the restoration of family values. **This** made the Republicans angry. **She** accused **him** of stealing **her** issues. Meanwhile, Patrick Buchanan addressed a rally in Manchester. **She** hammered away at the theme that the jobs of American workers must be protected from low-wage foreign competition. **This scandal** has propelled him to the front in the polls.

Cohesive but incoherent text

President Bill Clinton gave a speech at Concord High School. **He** talked about playing the saxophone and mused about Plato. **This** made the Republicans angry. **They** climbed the Mount Everest. Meanwhile, Patrick Buchanan addressed a rally in Manchester. **He** ran down the street and smashed two cars. **This courageous action** has propelled him to the front in the polls.

Two related principles from composition:

Frame-style paragraph	Sentences in such a paragraph all have the same grammatical subject or main focus. The paragraph presents a frame focused on one entity; each sentence is a frame slot giving information on that entity, allowing the user to maintain focus rather than jumping back and forth.
	Example: **Cattle** (called cows in vernacular usage) are domesticated ungulates, a member of the subfamily Bovinae of the family Bovidae. **They** are raised as livestock for meat (called beef and veal), dairy products (milk), leather and as draft animals (pulling carts, plows and the like). In some countries, such as India, **they** are subject to religious ceremonies and respect. **Cattle** are estimated to number 1.3 billion in the world.

Spreading activation paragraph	Alternatively, the sentences in a paragraph should be strung together so that the entity mentioned in the previous sentence is taken up at the beginning of the next sentence, like a path through a semantic network.
	Example: **Cattle** are raised for beef and **milk**. Their **milk** is an important source of **calcium**. **Calcium** is important for growing strong bones and healthy teeth in children and preventing osteoporosis. Calcium is also important for many functions in the body, for example, **muscle contraction**, **which** is especially important for athletes.

Relationship types and their expression in text

The relations given in the Crombie reading and illustrated by two annotated text examples on the next two pages elucidate further the concept of text cohesion and coherence. To understand a document, a person or a computer program must ascertain the relationships between individual elements. For a machine language understanding system we would need rules that link the type of relationships with grammatical features and the relationships between words in the lexicon. The same information is needed for the converse process of text generation. Some of these relationships have also been proposed and used for *relational indexing*, which indicates not only the concepts treated in a document but also the relationships between these concepts.

Page from Crombie

Page from Crombie

Lecture 5.2b ***February 16***

Text analysis overview and examples (30 minutes)

Objectives (In addition to the objectives inherited from 5.2-6.2)	1 Be aware of important text analysis methods. 2 Have an appreciation for and understanding of the potential of automated text analysis for processing vast quantities of information through automated translation, automated indexing, and extraction of data from text
Practical significance	Increasing amounts of text need sophisticated linguistic tools for intelligent indexing and data extraction, for example, Convera RetrievalWare, www.convera.com/solutions/retrievalware/default.aspx , Inxight, www.inxight.com/pdfs/linguistics_adding_value.pdf Specifically, linguistic techniques can help with the following functions: • Preparing a description of the document Descriptive cataloging (e.g. from optically scanned title page) Subject indexing Abstracting, text summarization (e.g., Tools > AutoSummarize in Word, Copernic Summarizer, www.copernic.com/en/products/summarizer/) • Determining the reading level of a document (more generally: the audiences for which the document is appropriate) • Determining the attitudes, beliefs, or emotions underlying the document (content analysis in sociology and political science or in psychoanalytical methods) • Determining authorship or other characteristics of the origin of the document • Preparing a hypertext version of a document, possibly for incorporation into a larger hypertext • Extracting data from a document. Representing the relationships expressed in a document in a more explicit and more easily manipulated way • Machine translation, for example on-the-fly translation of Web documents

Scope and limitations of lectures 5.2b-6.1a	
Scope	This pair of lectures introduces tools and methods for performing linguistic and statistical analysis of text. This includes • The use of textual analysis in the building of information retrieval systems and knowledge-based systems. • Frame-based data extraction from text. • Rudiments of parsing sentences
Limita-tions	These lectures concentrate on a subarea of document analysis, namely natural language processing applied to machine-readable text (text available as a stream of individually encoded characters). Text available as sound or graphics can be automatically converted (optical character recognition, speech recognition). Analysis of images (for example, object recognition)and analysis of sounds are other subareas of document analysis. Literary and artistic analysis also falls in the broader context of document analysis.

Approaches to text analysis

Most of these techniques are used by human readers and machine systems alike for the purposes outlined under *Practical significance*.
Human readers may analyze a text for indexing, abstracting, extracting a specific fact or proposition, or for assimilating all the facts or propositions expressed in the text.

- **Statistical**
  - Word / phrase / concept frequency
  - Frequency of words that connote an attitudinal/emotional dimension (**content analysis** in psychology/sociology/political science).
  - Differential frequency.
  - Looking for the unexpected (such as weighting rare words highly in ranking retrieval results), as in AltaVista's ranking method, Lecture 5.1

  The statistical approach is used mostly by computer systems, but perhaps also implicitly by human readers

- **Based on text macrostructure - positional approach**

  For example:
  - Introduction and conclusions useful source for abstract.
  - Section headings and figure captions useful source for index terms.
  - First and last paragraphs of sections, first and last sentences of paragraphs

- **Cue words, cue phrases, and cue sentences**

  For example, "method", "important result", "new"

- **Syntactic and semantic analysis**
  - Parsing of sentences or partial parsing to detect noun phrases
  - Parsing with semantic interpretation
  - Inter-sentence parsing, resolution of anaphoric references

- **Slot filling in frames using parsing or cues**

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

A further technique, not shown in the examples, is
searching for a word or phrase and its synonyms

Some of the techniques mentioned here (**in bold**) are applied in Assignment 7

Examples start on next page

1 Importance of resolving *anaphoric references* in free-text searching with proximity operators

Proximity operators used here (syntax varies from system to system)

WS two words occurring in the same sentence

WP two words occurring in the same paragraph

Texts are from the Columbia University College of Physicians and Surgeons *Complete medical home guide*

Query statement / information need:	What to do about sticky eyelids
Query formulation to search free-text:	eyelid! WS stick!

BLEPHARITIS

Blepharitis is an infection of the edges of the **eyelids**. *They* become red, **sticky**, and crusty, and sometimes the victim has to unstick them to see anything in the morning.

The WS query formulation misses this entry because eyelid and sticky do not occur in the same sentence.

Implications for free-text searching

Query formulation: calcium WS excret! WS within same sentence

Query formulation: osteoporosis WP vertebr! WP within same paragraph

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.

2 Extracting data through slot-filling in frames: examples

Understanding and summarizing stories by machine

Based on distinguishing types of stories, such as *corporate merger*, *disaster*, *state visit*.
Each type of story has a list of items to be included in a summary; these are arranged in **a frame** specific for that type of story.

The summarizing process then proceeds in two steps:

1 Detect basic type of story, for example *story about disaster*, and pull up the proper frame

2 For filling each slot, fill the slots following the instructions given

Disaster frame – general pattern

Slot	**Instructions: What to look for to find slot fillers**
Type of disaster	indicator word such as *earthquake*, *aftershock*, *hurricane*
Where	place name (from a large dictionary of place names)
When	date line plus words such as *today, yesterday, Sunday*, *recent*
Number of dead	*killed* or *dead* or *fatality*, and a number close by
A*mount of damage*	(*$* or *dollar* or and number before or after) or *much* or *heavy*, esp. when close to *damage* or *worth* or *destroyed*

Disaster frame – Event 345

Slot	**Slot filler** (for story on facing page)
Type of disaster	earthquake aftershocks
Where	central Italy
When	October 6, 1997
Number of dead	10
Amount of damage	$1 billion

Disaster frame – Event 406

Slot	**Slot filler** (for story on facing page)
Type of disaster	hurricane
Where	Mexico's Pacific Coast, Acapulco
When	October 9, 1997
Number of dead	120
Amount of damage	untold millions of dollars

Aftershocks Jar Central Italy; Repair Cost Put at $1 Billion

Associated Press

ROME, **Oct 6**—The ground in **central Italy** rumbled again **today**, and officials said repairing buildings **damaged** by a series of earthquakes could cost more than **$1 billion.**

The aftershocks in the Umbria and Marches regions have prompted more people to seek temporary shelter, 11 days after a pair of quakes **killed 10** people. The National Geophysics Institute said today's tremors hit about every 30 minutes before dawn, the strongest with a magnitude of 3. No new destruction was reported.

The Sept. 26 quakes damaged the beloved Basilica of St. Francis in Assisi, along with thousands of other buildings.

The less severely damaged buildings will be repaired so that as many people as possible can return to their homes before winter, civil defense chief Franco Barberi said at a news conference. He said it will cost $875 million to $1.15 billion to repair damaged buildings.

The government will move about 3,000 units of prefab housing into the region in the next few weeks. Tents and camping vehicles already in place can shelter as many as 50,000 people.

Hurricane Devastates Mexico's Pacific Coast

Floods Kill at Least 120, Most in Acapulco

By Chris Kraul
and Mary Beth Sheridan

Los Angeles Times

ACAPULCO, Mexico, **Oct 9**-
Bearing 115 mph winds and torrential rain, Hurricane Pauline roared out of the Pacific through this coastal resort region before dawn **today**, leaving at least **120** people **dead**, thousands homeless and **untold millions of dollars** in **damage**.

Most of the dead were counted in and around Acapulco, a sunny port city usually filled with carefree Mexican and foreign tourists. The powerful storm left Acapulco, a city of about 1 million people, "unrecognizable," according to one report—a tangle of uprooted trees, downed power lines, overturned cars and bodies.

Morning light revealed corpses and garbage and the wreckage of countless wood-frame homes floating in oily, four-foot-deep floodwaters that coursed through the streets and washed over La Costera Miguel Aleman, a fabled promenade skirting Acapulco's ocean-front. City officials said there had been some isolated instances of looting, and army units were called out to patrol the streets. A deluge of rain—20 inches in less than 24 hours—sent floodwater, mud, gravel and boulders rushing down drought-parched hills surrounding Acapulco through several slum neighborhoods, smashing poorly constructed shanties and more substantial houses to flinders and washing away anything not firmly anchored. At least seven mudslides reportedly caused heavy property damage around the city, and local officials fear thick layers of mud coating many neighborhoods may conceal dozens of bodies as well.

While the official death stood at 120 late last night local authorities said it would certainly climb—and perhaps double—as search parties comb through the debris left by the storm. The U.S. Embassy in Mexico City said that no Americans were reported among the dead or missing.

The Red Cross issued a plea for

See HURRICANE, A29, Col. 4

From the Washington Post, October 7, 1997 and October 10, 1997, respectively

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 Extracting data from text, esp. importance of resolving anaphoric references

Next page

3 Extracting data from text, esp. importance of resolving anaphoric references

Consider the following text (only the example marked with | at the left margin is treated in class; explore the other **bold** / *italic* pairs on your own):

VASCULITIS

VASCULITIS, as the name implies, is an inflammation of the blood vessels — both the arteries and the veins. Diseases in this category are relatively rare and comprise some of the most baffling and poorly understood disorders in medicine. Very often, the diagnosis remains unsuspected for long periods because of the variable way in which these disorders behave.

Inflammation of a blood vessel, particularly a small artery, can cause a narrowing of its lumen (internal diameter). If the vessel becomes completely closed, the tissue normally nourished by the diseased artery will die or be severely damaged.

Some forms of vasculitis are believed to result from an allergy or hypersensitivity, such as an adverse reaction to certain drugs. Sulfa drugs were very common causes of vasculitis, particularly in the early days of their use when the preparations were more crude and the dosages given were higher than today.

Patients with vasculitis, particularly when it involves widespread areas in the body, many be extremely ill with a generally poor prognosis. One particular type of vasculitis, which affects older people, involves inflammation of the cranial or temporal arteries, the vessels that serve a portion of the facial, jaw, and tongue muscles, the scalp, and most important, the retina. **Cranial arteritis** is the most common cause of sudden blindness in the elderly. Usually only one eye is involved but sometimes it occurs in both. *This condition* is successfully treated with corticosteroids, provided that treatment is started before there is significant loss of vision. *It* is often associated with a syndrome of severe muscle pain and stiffness called **polymyalgia rheumatica**. *This illness* is also largely confined to the elderly. It is almost always associated with a very high sedimentation rate, which measures the amount of inflammation, and it usually responds dramatically to cortisone-type drugs in low doses. Polymyalgia may occur without cranial arteritis, but because of the association, arteritis should be suspected in patients with polymyalgia.

Another form of vasculitis is called **Wegener's granulomatosis**. *This* is an extremely rare disorder which attacks the respiratory tract, the nasal sinuses, and the kidney in a progressively destructive process. Wegener's granulomatosis was once invariably fatal but now most patients can be treated successfully with cytotoxic or immunosuppressive drugs.

Patients with generalized or systemic vasculitis will often have paralysis of a foot or a wrist as a result of loss of blood supply to the peripheral nerve serving that limb. The blood vessels of the lung may also be affected, resulting in asthmalike symptoms. The development of asthma relatively late in life is very unusual, and may signify vasculitis.

There is another type of vasculitis known as **Takayasu's disease**, which occurs almost exclusively in young women. The inflammation is largely restricted to the branches of the great artery which leaves the heart (the aorta). *It* has also been called "pulseless" disease, for the diseased arteries may be so narrowed that a pulse cannot even be detected at the wrist. Patients with *this disease* will very frequently have symptoms of dizziness, light-headedness, weakness, and difficulty in using the arms, due to muscle pain from even slight physical effort. This is a direct result of lack of oxygen to the muscles, as the narrowed arteries are unable to deliver the increased amount of blood required during muscular effort. Corticosteroid therapy may be effective against Takayasu's disease, but the disease may go into remission without treatment.

These diseases are a few examples of the very broad spectrum of disorders included in the category of vasculitis. They are often difficult to diagnose, for their onset and evolution may be vague and ill-defined. The more classic types are easier to identify, but because of their relative rarity they are often not suspected until late in the course of the illness. Biopsy of an involved organ such as the kidney, muscle, or liver may be required in order to establish that a vasculitic process is indeed present

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, . . .

4 Examples: Importance of recognizing noun phrases for retrieval and translation

Noun phrases and semantic interpretation (word sense disambiguation)
Note: To search for phrases, in most systems use " ".

Example 1

information retrieval, retrieval of information, retrieval of legal information
but: information on the retrieval of sunken treasures

Example 2. Noun phrases expressing a unit of meaning

hepatitis A
vitamin A
twelve-step program
route of administration (also known as administration route, medication route, route of drug entry, method of drug application)
gene pool
breath test
motivational interviewing
blue law
social control
boundary layer flow (aerodynamics)
data link layer (data communication)

peer pressure, pressure by peers social pressure vapor pressure benefits program safety program conference program computer program	Meaning of polysemous words determined by context in a phrase

Example. Importance of parsing complete sentences for noun phrase identification

1 The green vegetables supply calcium. **NP** The green vegetables **V** supply
2 The green vegetables supply calcium to the body.
3 The green vegetables supply digestible calcium.
4 The green vegetables supply determines sufficiency of calcium. **NP** The green vegetables supply

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.

5 Importance of semantic interpretation for retrieval and automated translation

Example:	**Importance of semantic interpretation for disambiguating homonyms in searching** (sense disambiguation, meaning disambiguation)
Query statement / information need:	Passages referring to white (race/ethnic group)
Query formulation to search free-text:	white

Passages retrieved:

White students were found to hold prejudices against their black and Hispanic peers.

White cars are preferred by middle-aged buyers.

The **white** dishwasher laughs

The **white** dishwasher is broken.

The black congresswoman won election in a majority **white** district.

Douglas **White** won the race.

A **white** knight came to the rescue of CSX Corporation in its take-over fight.

The family unit is the basis for American society. **White** units make up 53% of all family units in the state.

GE makes microwave ovens. Half the units sold are **white**.

The **white** drinking fountain

a. In a story set in the historic segregationist South
b. In a travel guide to Italy

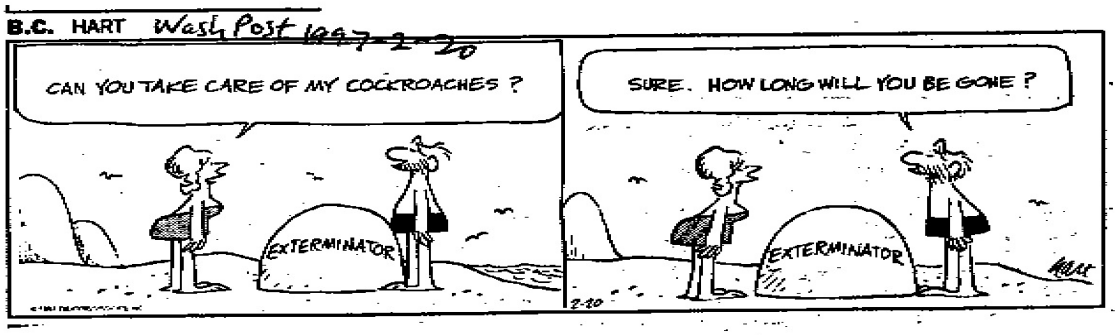
A sophisticated free-text retrieval system would analyze the text to determine the meaning of **white** in each passage and tag the passage accordingly. It would ask the user what meaning of **white** she was after and find only properly tagged passages. Mistakes in the analysis may cause retrieval of erroneous passages and rejection of relevant passages.

Example: Importance of semantic interpretation for automated translation	
The white *dishwasher* laughs.	**The white *dishwasher* is broken.**
German: Der weisse *Tellerwäscher* lacht.	German: Die weisse *Spülmaschine* ist kaputt.
French: Le *plongeur* blanc rit.	French: Le *lave-vaisselle* blanc est détraqué.

Semantic interpretation often requires the parsing of complete sentences.

More examples: Semantic interpretation rules

Jokes are often based on deliberately misconstruing the meaning of a word in a given context. Thus, they can focus the attention of the language analyst on words with multiple meanings and on the semantic interpretation rules that distinguish between these meanings. With this in mind, analyze the comic strip episodes below.



DOONESBURY
By Garry Trudeau
Washington Post 2000-4-2
ANOTHER GREAT SPEECH, DAD.
I DUNNO. THIS CAMPAIGN NEEDS A JOLT.
WHAT KIND OF JOLT?
I WAS THINKING OF ANNOUNCING MY VICE PRESIDENT.
BUT WE HAVEN'T BEEN MADE UP A LIST YET.
I'VE ALREADY GOT IT FIGURED OUT! I WANT KATHIE LEE GIFFORD.
EXCUSE ME?
WHY NOT? SHE'S TOUGH, SHES COMPETITIVE AND SHES AVAILABLE!
PLUS, SHES MAJOR HEADLINE BAIT. A PROVEN NUMBER TWO, AND TOUGH ENOUGH TO RUN A SWEATSHOP!
YOU KNOW, THAT'S NOT SUCH A CRAZY IDEA, DAD.
DOES SHE HAVE ANY CONVICTIONS?
WHO CARES? I HAVE FIVE, AND IT HASN'T SLOWED ME DOWN!

February 16, 2011 Name (optional)

Free-write 5

Lecture 5.1. Data structure and search modes

Lecture 5.2a. Knowledge (re)presentation in text and images. Text linguistics

Lecture 5.2b. Text analysis overview and examples

- **Reflect** – what you learned, what was most important, what was most interesting, what was extraneous;
- **Ask questions** – ask for more explanation, how is a concept connected to other concepts, why is a concept important, how can it be applied, why is a reading important;
- Offer **critique and suggestions**;
- Say anything else you want to

.

Over

Lecture 6.1a (45 min) ***February 23***

Natural language processing. Syntactic and semantic parsing

Objective inherited from Lect. 5.2-6.2 and	Have a general idea how syntactic parsing and semantic interpretation work. Note: Parsing is the linguists' term for sentence diagramming
Practical significance	Natural Language Processing (NLP) is booming both in commercial applications and in academic research. It is used for the following purposes (see also list in Lecture 5.2b and example on next page): • Automatic indexing /classification, including document categorization / automatic cataloging / automatic metadata generation • Automatic abstracting, automatic summarization, including creating unified summaries from multiple documents (e.g., multiple news stories on the same topic or event) • Automatic extraction of formatted data from text (information extraction, fact extraction, relationship extraction) (See example on next page) • Question-answering: Within a large document, find the specific sentence or paragraph that answers a question • Automatic translation (e.g., www.google.com/language_tools) • Grammar checking in a word processor (or computer-assisted essay-grading) • Creating textual answers from the data returned by a database query • Automatic speech recognition
Note	Working through some detailed examples is necessary to create a good sense of what is going on in natural language processing. However, there is no need for memorizing the details. You will not be required to produce the step-by-step sequence of a parse. *Parsing* is just a different word for *diagramming sentences*. *Parsing* is generally used with the connotation that the diagramming is done by a computer program.

Introductory remarks on Natural Language Processing (NLP)

Practical significance

Example of information extraction (entity-relationship statement extraction, relationship extraction)

Done by hand to illustrate what we want a machine to do

Patients with vasculitis, particularly when it involves widespread areas in the body, may be extremely ill with a generally poor prognosis. One particular type of vasculitis, which affects older people, involves inflammation of the cranial or temporal arteries, the vessels that serve a portion of the facial, jaw, and tongue muscles, the scalp, and most important, the retina. **Cranial arteritis** is the most common cause of sudden blindness in the elderly. Usually only one eye is involved but sometimes it occurs in both. *This condition* is successfully treated with corticosteroids, provided that treatment is started before there is significant loss of vision. *It* is often associated with a syndrome of severe muscle pain and stiffness called **polymyalgia rheumatica**. *This illness* is also largely confined to the elderly. It is almost always associated with a very high sedimentation rate, which measures the amount of inflammation, and it usually responds dramatically to cortisone-type drugs in low doses. Polymyalgia may occur without cranial arteritis, but because of the association, arteritis should be suspected in patients with polymyalgia.	vasculitis <*mayCause*> extreme illness vasculitis <*hasPrognosis*> poor cranial arteritis <*definedAs*> vasculitis that involves inflammation of the cranial or temporal arteries, the vessels that serve a portion of the facial, jaw, and tongue muscles, the scalp, and most important, the retina cranial arteritis <*occursIn*> elderly cranial arteritis <*mayCause*> (sudden blindness, elderly) (sudden blindness, elderly) <*causedBy*> (cranial arteritis, high percentage of cases) (blindness <*causedBy*> cranial arteritis) <*preventedWith*> (corticosteroids, given early) cranial arteritis <*associatedWith*> polymyalgia rheumatica polymyalgia rheumatica <*definedAs*> syndrome of severe muscle pain and stiffness polymyalgia rheumatica <*occursIn*> elderly polymyalgia rheumatica <*causedBy*> (cranial arteritis, medium percentage of cases) sedimentation rate <*measures*> degree of inflammation polymyalgia rheumatica <*associatedWith*> (sedimentation rate, high) cranial arteritis <*mayCause*> polymyalgia rheumatica polymyalgia rheumatica <*treatedWith*> (cortisone-type drugs, dosage: low, response: very good)

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

P. 145 - 154 The parsing game (take these pages out of your binder)

P. 156 - 171 More detail about the syntactical analysis (look at together with the parsing game)

571 Soergel

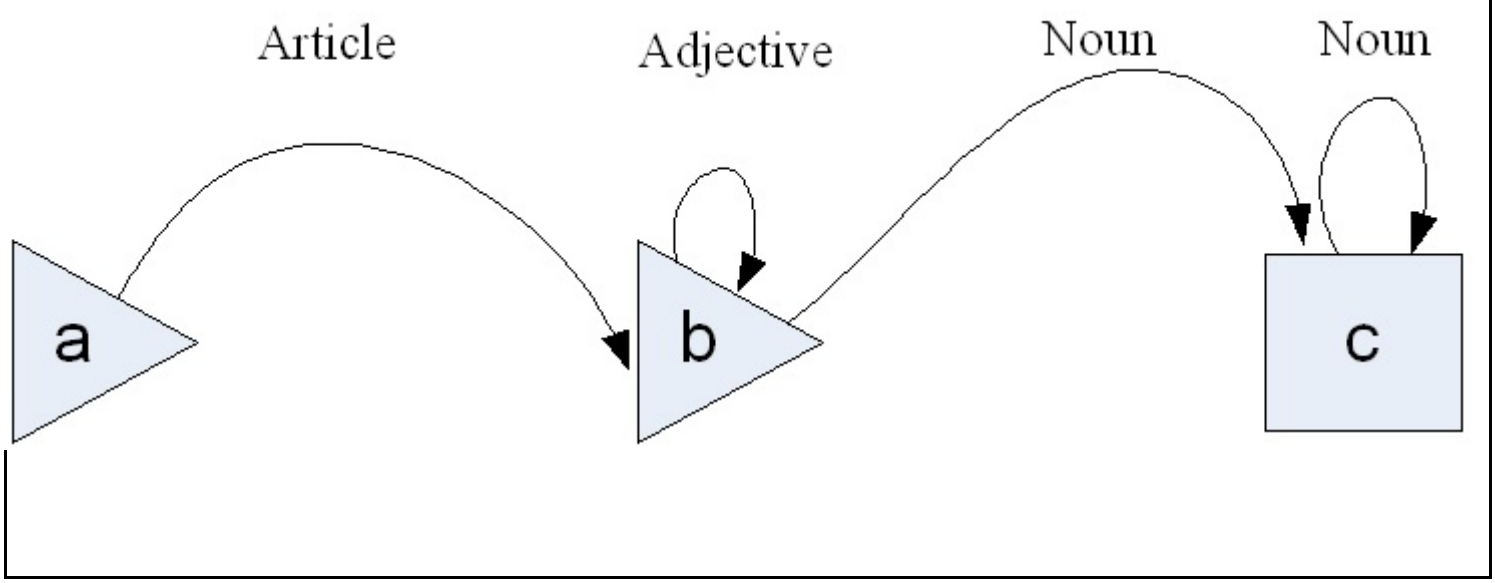
The parsing game

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 $_3$

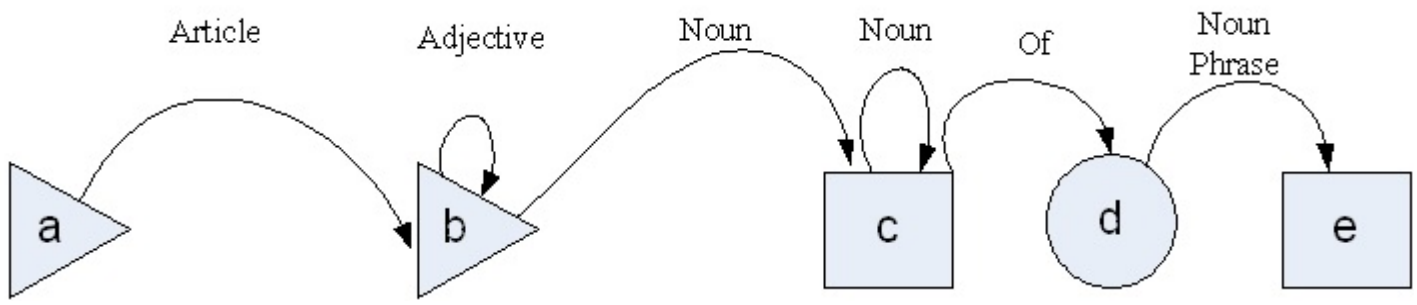
7 $_0$ bone $_1$ mass $_2$

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)



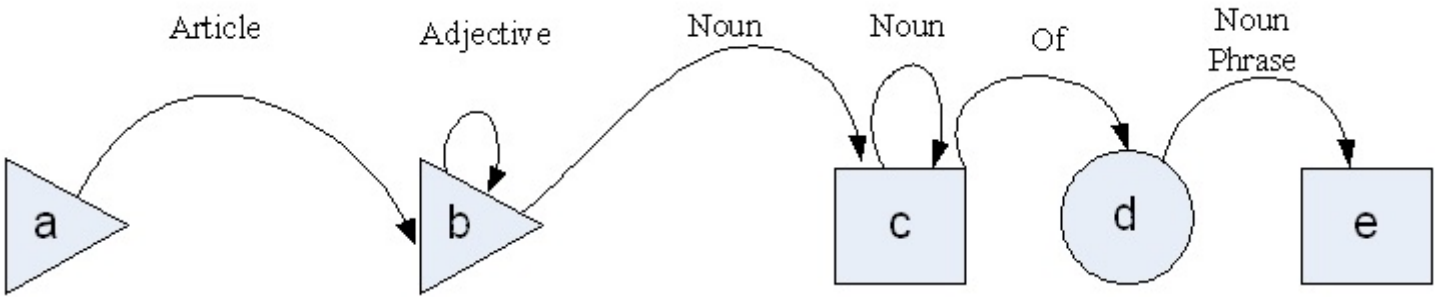
Sample noun phrases (by general linguistic convention, * means syntactically incorrect)

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

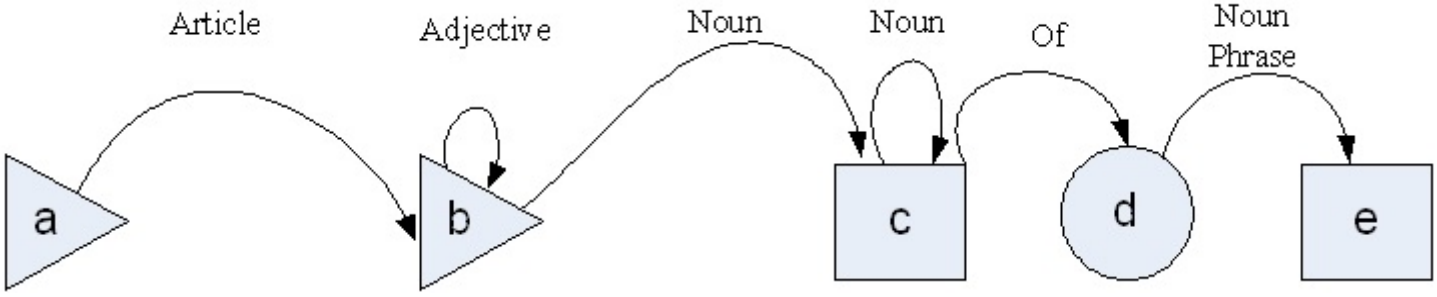
2 $_0$ the $_1$ growing $_2$ skeleton $_3$ parts $_4$ of $_5$ healthy $_6$ small $_7$ children $_8$

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

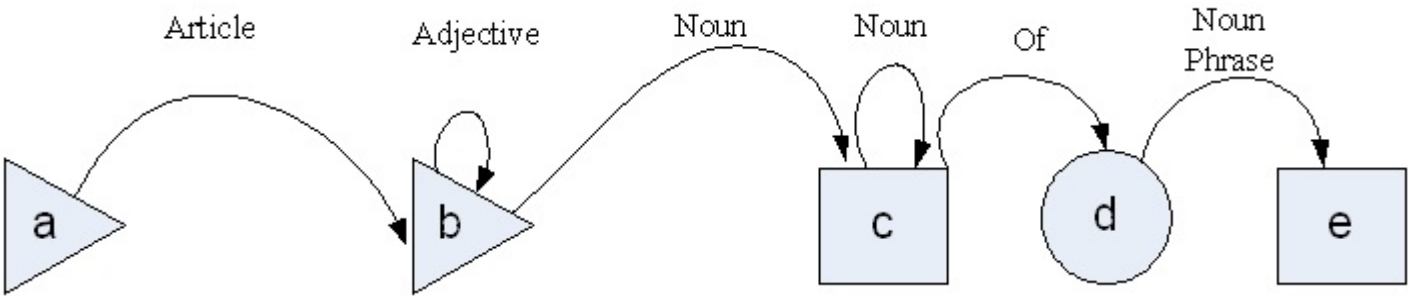
Noun phrase network (2)



Noun phrase network (3)



Noun phrase network (1)



OSTEOPOROSIS

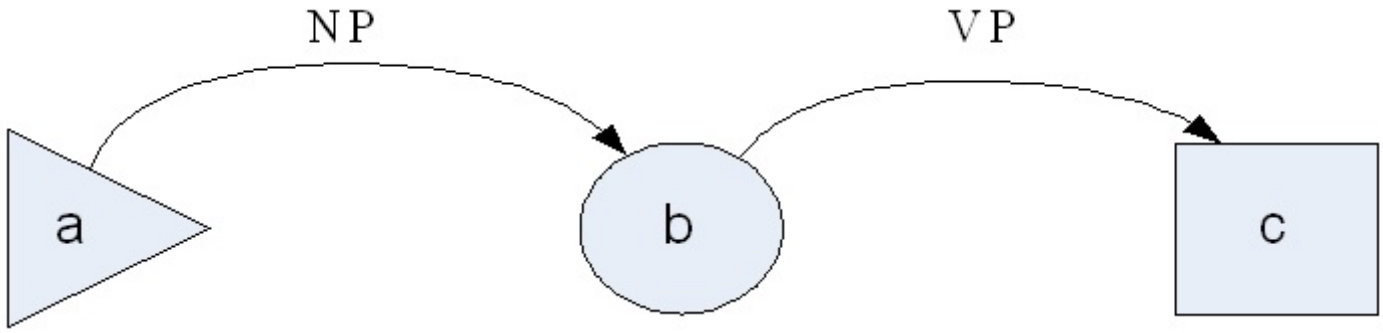
BONES NEED CALCIUM to maintain their strength, hardness, and to stay healthy. Milk, the main source of calcium in the diet, is important for the growing skeletons of children and adolescents as well as the bone-forming cells of adults. Regular daily consumption of at least 1 cup of skim or low-fat milk 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.

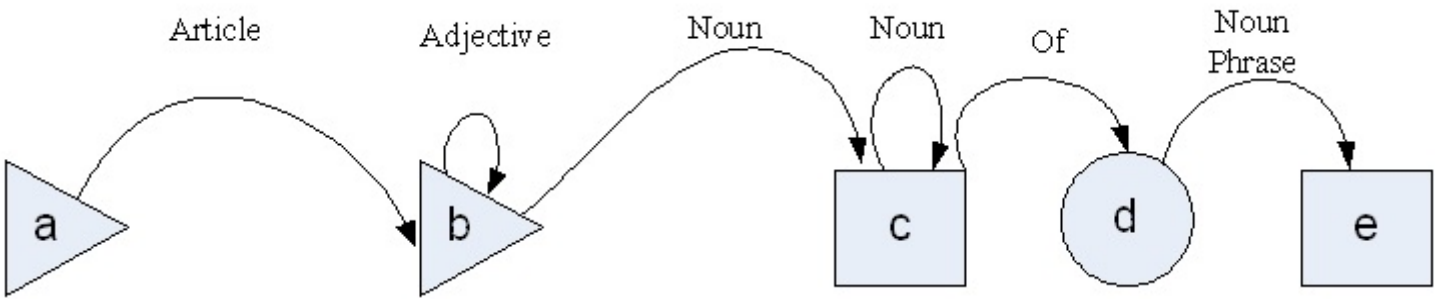
It is not known why calcium loss occurs. That postmenopausal women tend to get osteoporosis points in the direction of a hormonal disorder as estrogen in women of this age falls off sharply. Estrogen therapy is one treatment but its ability to decrease calcium loss may last only several years. Increased 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.

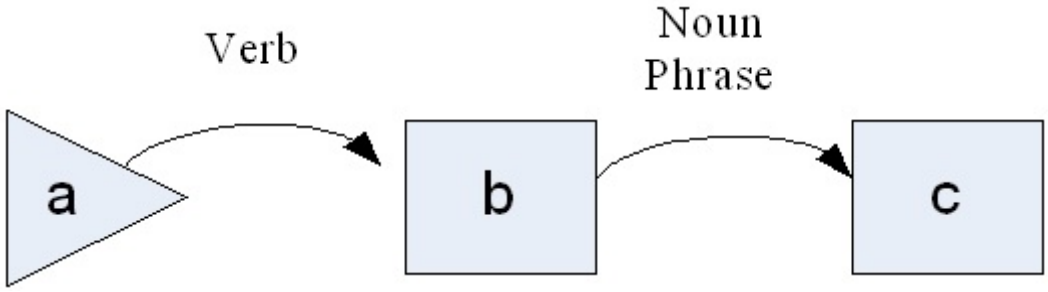
Sentence network



Noun phrase network (1)



Verb phrase network

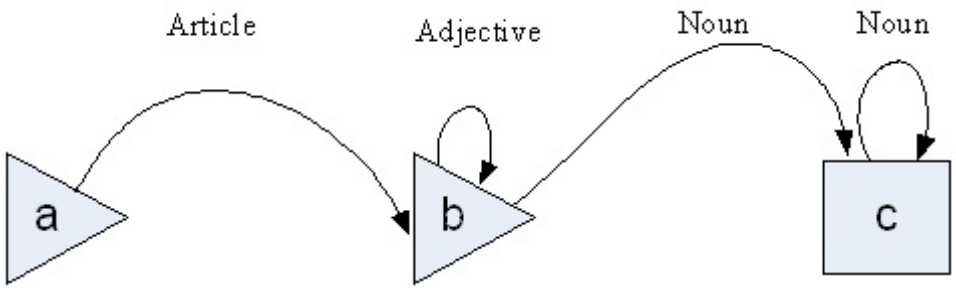


1 $_{0}$ **The** $_{1}$ **green** $_{2}$ **vegetables** $_{3}$ **supply** $_{4}$ **calcium** $_{5}$**.**
2 The green vegetables supply calcium to the body. [Not recognized by our simplistic parser.]
3 The green vegetables supply digestible calcium.
4 The green vegetables supply determines sufficiency of calcium.

Go to next page

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 $_2$

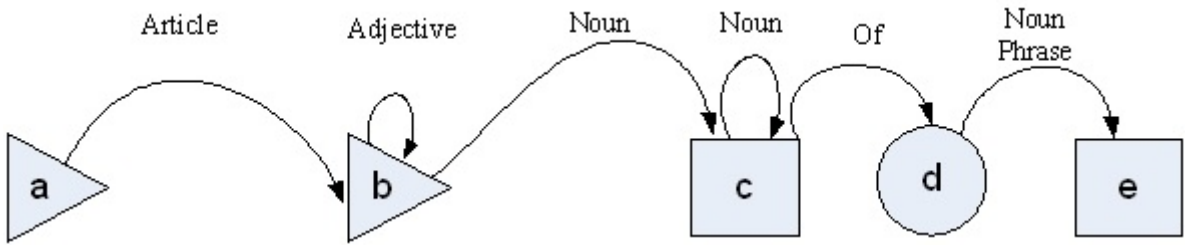
10 $_0$ a $_1$ deficient $_2$ calcium $_3$ supply

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



▷ possible start state ◯ intermediate state □ possible end state

(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**

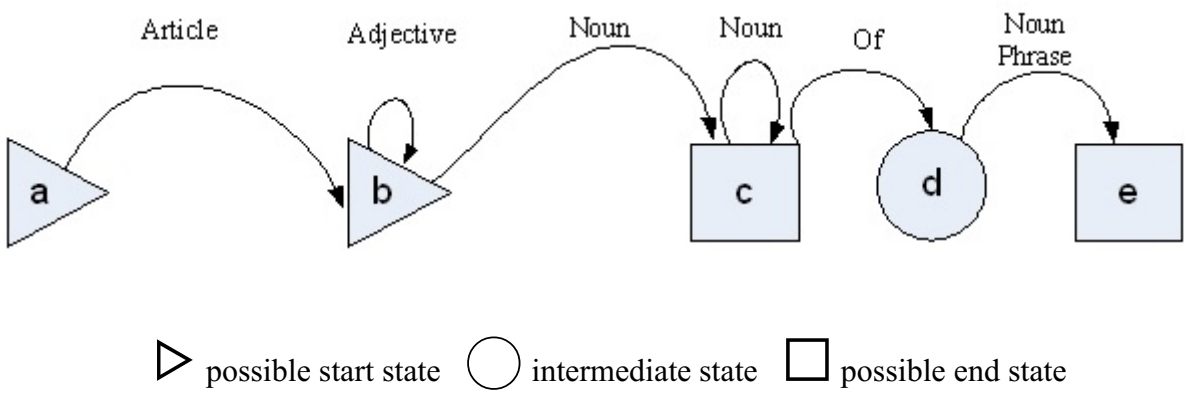
From pos	From state	Arc tried	segment (word) processed	To state	To pos	comment
	$_0$ the $_1$ growing $_2$ skeleton $_3$ parts $_4$ of $_5$ healthy $_6$ small $_7$ children $_8$					
0	a	ART	the	b	1	
1	b	ADJ	growing	b	2	
2	b	NOUN	skeleton	c	3	
3	c	NOUN	parts	c	4	
4	c	OF	of	d	5	
5	d	NP	healthy small children	**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.

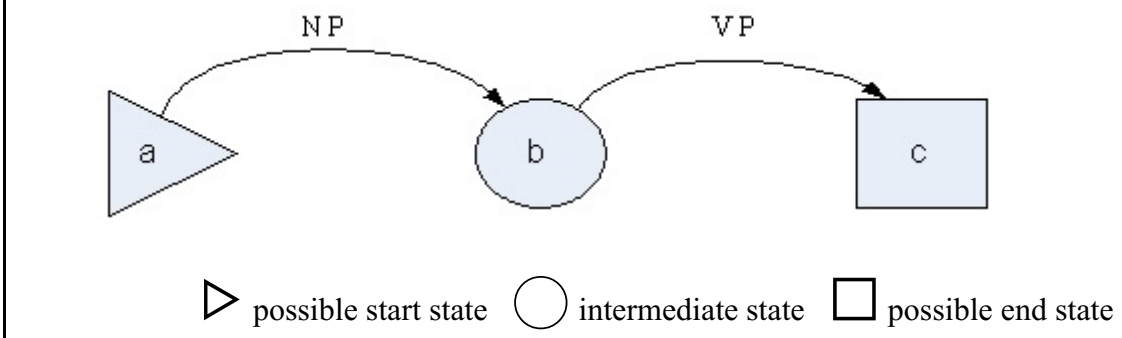
When calcium enters the body, it is absorbed into the bloodstream. If there is any excess, it is deposited in the end of the bone shafts where it is stored until the body needs to tap this reserve. (Some is also excreted via the kidneys.) When the calcium supply is deficient, the blood must take it back 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.

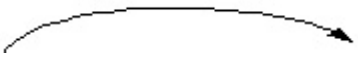
Parsing of sentences: The sentence network outlines a grammar for simple sentences.

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	

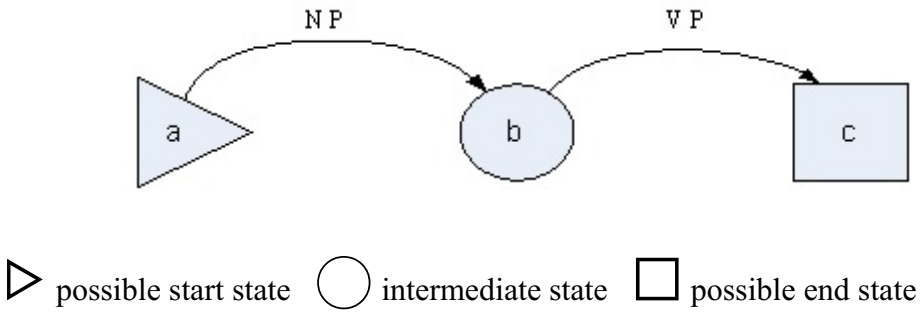
Result: An analysis of the sentence structure, a sentence diagram.

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

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

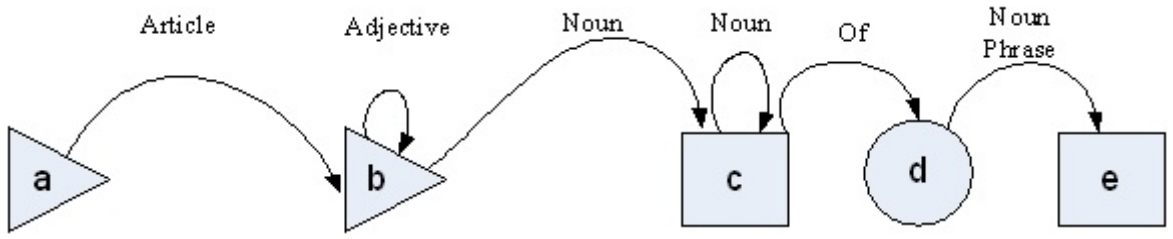
Sentence network

S



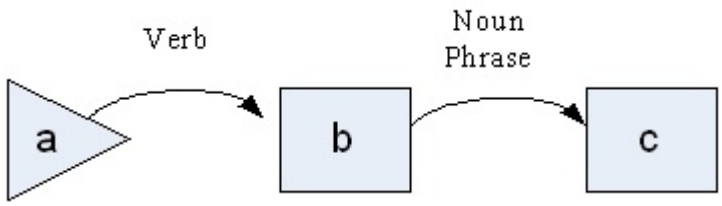
Noun phrase network

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)

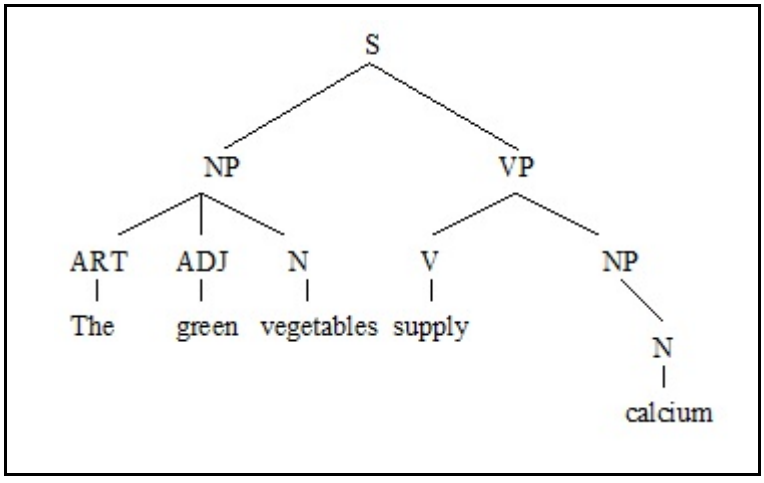
$_0$ The $_1$ green $_2$ vegetables $_3$ supply $_4$ calcium. $_5$

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)**]**
}

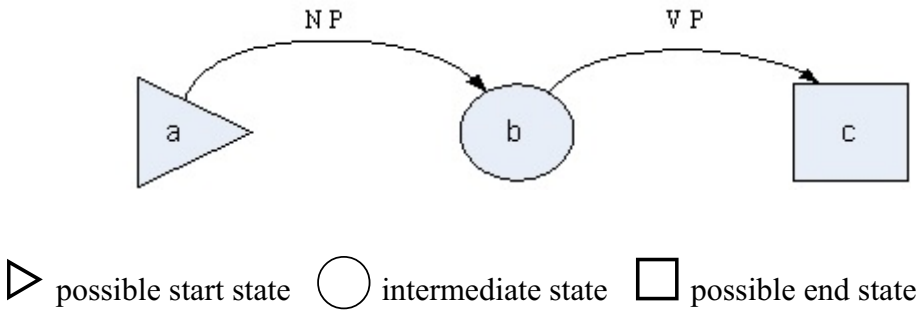
Parse tree representation:



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

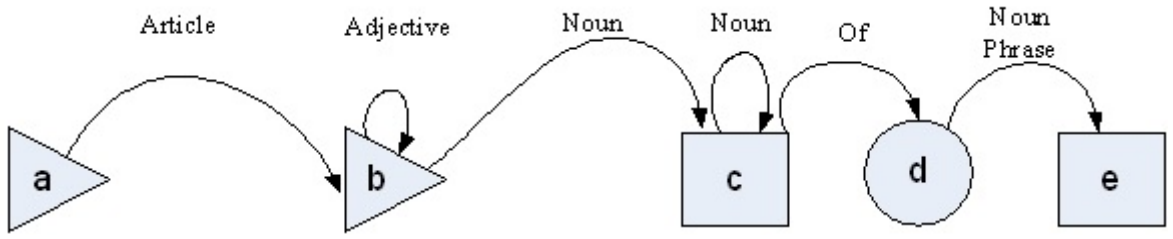
Sentence network

S



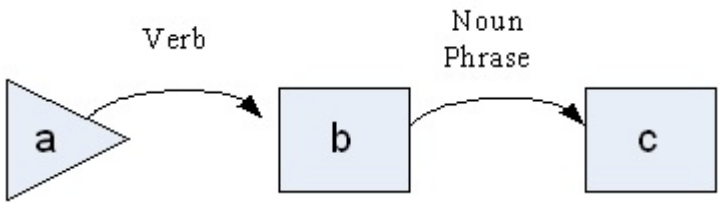
Noun phrase network

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		

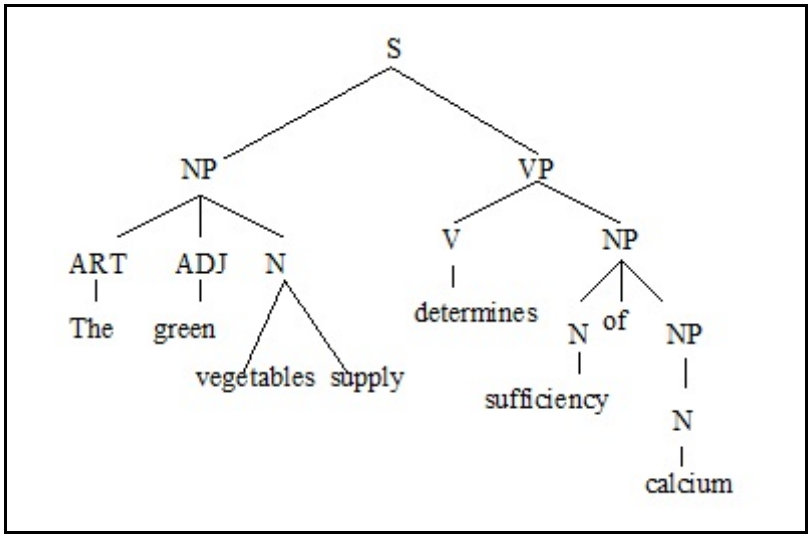
* Backtrack point Superscript indicates the nesting depth

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)**]**
}

Lecture 6.1b (40 min) ***February 23***

Document macrostructure, document templates Inter-document relationships

➜ LIS 506 Information Technology

Objectives (in addition to objectives inherited from Lect. 5.2-6.2)	1 Understand the importance of document structure in general and document templates in particular (see *practical significance*); 2 Understand document type / document template systems with hierarchy and hierarchical inheritance; 3 Be able to design a document template.
Practical significance	• Document templates make document creation so much easier and thus save a lot of work; • Good document structure makes reading and understanding documents easier; • Good document structure allows for pinpoint retrieval of relevant document sections; • Well-structured hypertext / hypermedia allows for reader-directed / learner-directed selection and sequencing of material.

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?

Document/text macrostructure

Structure of a scientific text - a frame for structuring information (in a full article or in an abstract)

One possible outline

1 **Background** (could also be called Problem)

- 1.1 General problem area (often including a review of the literature)
- 1.2 Specific problem. Purpose of the study, question to be answered

2 **Methods**

- 2.1 Discussion of the methods used in the study
- 2.2 Description of the actual conduct of the study

3 **Results**

4 **Conclusions**: Relationship to existing body of knowledge. Implications for decision making and/or further research

Another list of journal article components
(from a study of the human indexing process)

Journal title	Introduction
Title	Statement of purpose
Author	Materials and methods
Author's affiliation	Results and discussion
Keywords	Conclusions
Abstract	Figures, tables, and plates with captions
Table of contents (sometimes)	Acknowledgments
	Literature cited

Next page: Structured abstract from *Alcohol Research*, an extremely well designed abstracting journal (on reserve in the Wasserman Library).

CONIGRAVE KM abstract 1049

Conigrave KM, Saunders JB, Reznik RB. Predictive capacity of the AUDIT questionnaire for alcohol related harm. Addiction 90 (1995) 1479-1485.

'AUDIT can predict a range of harmful consequences of alcohol consumption'

Background

Drinking problems often are not recognized. Most of the people who become alcohol-dependent do not seek help until their problems are obvious. Late diagnosis is of particular concern because effective and low-cost methods of treating problem drinking at an early stage are now available. In 1989, the WHO published a brief 10-item screening questionnaire, the Alcohol Disorders Identification Test (AUDIT) specifically designed to identify problem drinkers before physical dependence or chronic problems have arisen. AUDIT has been reported to have a sensitivity of 92% and a specificity of 94% in detecting hazardous or harmful alcohol use. This study examined the ability of the AUDIT questionnaire to predict which subjects experience medical or social harm from their drinking.

Methods

Subjects were 350 patients who attended a hospital emergency ward in 1984-1985. They underwent a comprehensive assessment of medical history, alcohol use, dependence and related problems in an interview schedule; the AUDIT questions were interspersed among other items. Biochemical variables measured included y-glutamyltransferase (GGT) and mean corpuscular volume (MCV). Twenty subjects refused to be contacted after 2-3 years or were excluded because of malignant disease. Thus, a cohort of 330 subjects (212 men, 108 women) was left for the longitudinal study; 250 subjects were interviewed again after 2-3 years. Interviewers were blind to the results of the initial assessment. The AUDIT questions were scored from 0 to 4. Subjects who scored 8 or more were classified as potentially hazardous drinkers. AUDIT was examined for its ability to predict a number of end-points including alcohol-related medical disorders, health care utilization, social problems and hazardous drinking at the time of follow-up.

Results

Of those who scored 8 or more on AUDIT at the initial interview, 61% experienced alcohol-related social problems compared with 10% of those with lower scores. They also reported more frequently alcohol-related medical disorders and hospitalization. The AUDIT score was a better predictor of social problems and of hypertension than laboratory markers. Its ability to predict other alcohol-related illnesses was similar to the laboratory tests, but GGT was the only significant marker of mortality.

Conclusions

AUDIT is a brief and convenient questionnaire which can readily be incorporated into the standard medical history. It can predict a range of harmful consequences of alcohol consumption. AUDIT should prove a valuable tool in screening for hazardous and harmful alcohol use so that intervention can be provided to those at particular risk of adverse consequences.

K.M. Conigrave, Centre for Drug and Alcohol Studies, Royal Prince Alfred Hospital, Missenden Road, Sydney, NSW 2050, Australia.

From *Alcohol Research*

Preview of document templates: A simple mail merge example

The main document: A form letter

LITTLE PEOPLE SCHOOL

February 3, 2011

«NamePrefix» «FirstName» «LastName»
«Street»
«City», «State» «ZipCode»

Dear «NamePrefix» «LastName»,

According to our records, «StudentFirstName» does not have a current Emergency Card on file at our school. Because this form is essential to «GenderPossessive» safety while at the Lourie Center, «StudentFirstName» will not be allowed to go on the field trip without it. I have enclosed a copy of this form for you to fill out and return as soon as possible. Please call me if you have any questions or need help with this in any way.

Sincerely,

Administrative Assistant
Little People School

Enclosure

Data source: A MS Access table

AddressTable									
ID	Name Prefix	FirstName	LastName	Street	City	State	Zip Code	StudentFirst Name	Gender Possessive
1	Mr.	Eric	Smith	504 Flower Ct	Springfield	VA	22151	Rebecca	her
2	Mrs	Elizabeth	Kain	4801 Thames St.	Springfield	VA	22151	Alexander	his
3	Dr.	Sylvia	Campbell	3708 Duke St.	Alexandria	VA	22304	Mary	her

Resulting letters: See facing page

LITTLE PEOPLE SCHOOL

February 3, 2011

Mr. Eric Smith
504 Flower Ct
Springfield, VA 22151

Dear Mr. Smith,

According to our records, Rebecca does not have a current Emergency Card on file at our school. Because this form is essential to her safety while at the Lourie Center, Rebecca will not be allowed to go on the field trip without it. I have enclosed a copy of this form for you to fill out and return as soon as possible. Please call me if you have any questions or need help with this in any way.

Sincerely,

Administrative Assistant
Little People School

Enclosure

LITTLE PEOPLE SCHOOL

February 3, 2011

Mrs Elizabeth Kain
4801 Thames St.
Springfield, VA 22151

Dear Mrs Kain,

According to our records, Alexander does not have a current Emergency Card on file at our school. Because this form is essential to his safety while at the Lourie Center, Alexander will not be allowed to go on the field trip without it. I have enclosed a copy of this form for you to fill out and return as soon as possible. Please call me if you have any questions or need help with this in any way.

Sincerely,

Administrative Assistant
Little People School

Enclosure

LITTLE PEOPLE SCHOOL

February 3, 2011

Dr. Sylvia Campbell
3708 Duke St.
Alexandria, VA 22304

Dear Dr. Campbell,

According to our records, Mary does not have a current Emergency Card on file at our school. Because this form is essential to her safety while at the Lourie Center, Mary will not be allowed to go on the field trip without it. I have enclosed a copy of this form for you to fill out and return as soon as possible. Please call me if you have any questions or need help with this in any way.

Sincerely,

Administrative Assistant
Little People School

Enclosure

Note

Mail merge per se is not a topic in 571, just used as an example of document templates.
If for some other purpose you are interested in learning about mail merge, here is a useful introduction:
http://extension.oregonstate.edu/esoc/ectu/services/lessons/documents/MailMerge_000.pdf

I am also happy to send you the files I used for this example upon request.

Example. A simple document system

A frame/object hierarchy of document templates and documents

A document template is a frame with a slot (or element) 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.

Lecture 6.2b deals with implementing document templates in XML using XML schemas or the older Document Type Definition (DTD)

The simple document system consists of just five document types arranged in a hierarchy:

Generic memo
. Sales report memo
. . Content management sales report memo
. . Customer relations management sales report memo
. Self assessment memo

.

For each document type, we give the template and a sample document, using the following conventions:

Bold	A template slot (or element)
Arial	An instruction to be carried out when the template is applied to produce a document. Usually these instructions are attached to a slot.
<variable>	A variable to filled in with the appropriate value by the system
Courier	Text or data filled in by the system or selected by the user from a menu of options displayed by the system
Times Roman	Text entered by the user
Italics	Comments/explanations (not part of the document)
[ ], [[ ]]	Inherited, from one level up, two levels up Inheritance is indicated separately for the slot and the content of the slot (the slot may be inherited from the level above, yet the content can be specified at the current level)
/* . . . */	Comment

Document template 1: Generic memo

Subtype of / child of / inherits from:	Top level
Has subtypes / children / inherits to:	Sales report memo, Self assessment memo
Metadata	
To:	
From:	<name of person signed on to system>, <title of person>
Subject:	
Date:	<today's date> /* *from computer's clock* */
Keywords:	
URI:	<Universal Resource Identifier> /* *to be filled in by system* */
MemoBody	
PlainText:	

Document example 1: Generic memo

To:	Sue Feldman, CIO (*Chief Information Officer*)
From:	Bob Boiko, content management specialist
Subject:	What XML (eXtensible Markup Language) can do for us
Date:	February 7, 2001
Keywords:	XML; content management; document structure; databases on the Web
URI:	www.jasca.com/bboiko/memo20010207-04

XML allows us to define document structures that will make it easier to create documents. Once a document is created, it can be displayed in many different ways (Web page in multiple formats, print, etc.) through applying style sheets (the simple Cascading Style Sheets, CSS2, or the more powerful eXtensible Stylesheet Language for document Transformation, XSLT). A table of contents can be created automatically. Moreover, the document can be displayed selectively using just the parts most appropriate for a given audience. Parts of one document can be reused in another document. In retrieval, specific parts of the document can be targeted; for example, a user could search for just the *results* section of scientific reports.

With XML we can also define documents that hold database records to present databases on the Web. The boundary between text documents and formatted databases becomes blurred.

Document template 2: Sales report memo

Subtype of / child of / inherits from:	Generic memo
Has subtypes / children / inherits to:	Content management sales report memo Customer relations management sales report memo

[**Metadata**]

[**To:**] <name of director of sales>, <value = "Director of Sales">

[**From:**] [<name of person signed on to system>, <title of person>]

[**Subject:**] /* *to be filled in by memo designer of child template* */ <last_month>

[**Date:**] [<today's date> /* *from computer's clock* */]

[**Keywords:**]

[**URI:**] [<Universal Resource Identifier> /* *to be filled in by system* */]

[**MemoBody**]

[**PlainText:**]

Sales data table: header <value = "Sales"> <last_month> <value = "in $1,000">
Run query /* *query to be filled in by designer of child templates* */

Data analysis:

Recommendations:

No document example. People just use this template to make more specific templates with values for their specific sales report already filled in, as in the template for
Content management sales report memo. Making these specific templates is much easier if one can start from the more general sales memo template .

Document template 3: Content management sales report memo

Subtype of / child of / inherits from:	Sales report memo
Has subtypes / children / inherits to:	No children
[[**Metadata**]]	
[[**To:**]]	[<name of director of sales>, <value = "Director of Sales">]
[[**From:**]]	[[<name of person signed on to system>, <title of person>]]
[[**Subject:**]]	<value = "Content management sales report"> [<last_month>]
[[**Date:**]]	[[<today's date> /* *from computer's clock* */]]
[[**Keywords:**]]	<value = "content management software">
[[**URI:**]]	[[<Universal Resource Identifier> /* *to be filled in by system* */]]
[[**MemoBody**]]	
[[**PlainText:**]]	
[**Sales data table:**]	[header <value = "Sales"> <last_month> <value = "in $1,000">] [Run query] "monthly_CM_sales"
[**Data analysis:**]	
[**Recommendations:**]	

Underline: Added to the sales report memo template

Again:

Templates and inheritance

A slot defined in a broad template, such as the *generic memo* template, occur in all subordinate templates, such a the *sales report* and *self-assessment memo* templates. The slot may inherit just as a bare shell for content (only the slot name is enclosed in []) or it may inherit with some or all of its content specifications, such as default value, limitations on values, or a procedure to be used to get the content (slot content specification enclosed in []). For example, the From slot always inherits down with the attached procedure: put in the name of the person signed on to the computer. The To slot inherits as an empty shell; the *sales report* template and the *self-assessment memo* template each has its own procedure for filling in a value. However, from *sales report* to *content management sales report* the To slot inherits with the attached procedure.

Document example 3: Content management sales report

To: Joe Bush, Director of Sales

From: Cindy Weaver, Sales Associate

Subject: Content management sales report January 2001

Date: February 5, 2001

Keywords: Content management software

URI: www.jasca.com/rweaver/memo20010210-13

Sales January 2001 in $1,000

		Fed. Gov.	State& local	Fortune 500	Small comp.	Total
TeamSite	Dec. 2000	500	150	700	200	1,550
	Jan. 2001	**700**	**200**	**900**	**300**	**2,100**
Templating	Dec. 2000	250	30	350	50	680
	Jan. 2001	**350**	**40**	**450**	**75**	**915**
Metatagger	Dec. 2000	100	20	200	30	350
	Jan. 2001	**150**	**30**	**250**	**50**	**480**
Metafinder	Dec. 2000	100	10	130	30	270
	Jan. 2001	**80**	**0**	**90**	**20**	**190**
Total	Dec. 2000	950	210	1,380	310	2,850
	Jan. 2001	**1,280**	**270**	**1,690**	**445**	**3,685**

Data analysis:

Smaller organizations make proportionately less use of Templating. Conversations with some customers showed that they do not have the expertise to construct sophisticated templates that would bring great efficiency to their work.

Sales of metafinder are languishing.

Recommendations:

Offer training in the use of Templating and also a consulting service where the consultant would set up the templates for use by the organization's staff.

Promote Metafinder more aggressively through demonstrations of search improvements achieved through its spelling correction and thesaurus lookup features. Also offer a large generic thesaurus with the software so that an organization does not have the expense of constructing its own thesaurus form scratch.

Document template 4: Customer relations management sales report memo

Subtype of / child of / inherits from:	Sales report memo
Has subtypes / children / inherits to:	No children
[[**Metadata**]]	
[[**To:**]]	[<name of director of sales>, <value = "Director of Sales">]
[[**From:**]]	[[<name of person signed on to system>, <title of person>]]
[[**Subject:**]]	<value = " Customer relations management sales report"> [<last_month>]
[[**Date:**]]	[[<today's date> /* *from computer's clock* */]]
[[**Keywords:**]]	<value = "Customer relations management software">
[[**URI:**]]	[[<Universal Resource Identifier> /* *to be filled in by system* */]]
[[**MemoBody**]]	
[[**PlainText:**]]	
[**Sales data table:**]	[header <value = "Sales"> <last_month> <value = "in $1,000">] [Run query] "monthly_CRM_sales"
[**Data analysis:**]	
[**Recommendations:**]	

Underline: Added to the sales report memo template

Document example 4: Customer relations management sales report

To: Joe Bush, Director of Sales

From: James Barry, Sales Associate

Subject: Customer relations management sales report January 2001

Date: February 5, 2001

Keywords: Customer relations management software

URI: www.jasca.com/jbarry/memo20010210-13

Sales January 2001 in $1,000

		Fed. Gov.	State& local	Fortune 500	Small comp.	Total
Product 1	Dec. 2000	500	150	700	200	1,550
	Jan. 2001	**700**	**200**	**900**	**300**	**2,100**
Product 2	Dec. 2000	250	30	350	50	680
	Jan. 2001	**350**	**40**	**450**	**75**	**915**
Product 3	Dec. 2000	100	20	200	30	350
	Jan. 2001	**150**	**30**	**250**	**50**	**480**
Product 4	Dec. 2000	100	10	130	30	270
	Jan. 2001	**80**	**0**	**90**	**20**	**190**
Total	Dec. 2000	950	210	1,380	310	2,850
	Jan. 2001	**1,280**	**270**	**1,690**	**445**	**3,685**

Data analysis:

Smaller organizations make proportionately less use of Product 2. Conversations with some customers showed that they do not have the expertise or training that would allow them to utilize the software.

Sales of Product 4 are very low.

Recommendations:

Offer customer training and a consulting service.

Promote Product 4 through demonstrations, perform user studies and solicit feedback.

Document template 5: Self assessment memo

Subtype of / child of / inherits from:	Generic memo
Has subtypes / children / inherits to:	No children

[**Metadata**]

[**To:**]	<supervisor of person signed on to the system>, <title of supervisor>
[**From:**]	[<name of person signed on to system>, <title of person>]
[**Subject:**]	<value = "Self assessment for year"> <last_year>
[**Date:**]	[<today's date> /* *from computer's clock* */]
[**Keywords:**]	<some subject keywords filled in from job description in database>
[**URI:**]	[<Universal Resource Identifier> /* *to be filled in by system* */]

[**MemoBody**]

[**PlainText:**]

Accomplishments:	header <value = "Accomplishments in year"> <last_year>
Goals:	header <value = "Goals for year"> <this_year>
Training needs:	header <value = "Training needs for year"> <this_year>

Document example 5: 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.

Started the development of logical templates for the most important document types and implementation through XML document type definitions.

Developed specifications for the acquisition of content management software and selected a vendor.

Goals for year 2001:

Begin implementation of the content management master plan.

Install software and train staff in intranet-based document creation, deployment, and search.

Redesign the company Web site and use new software to streamline deployment of content on the Web

Training needs:

A course in information architecture

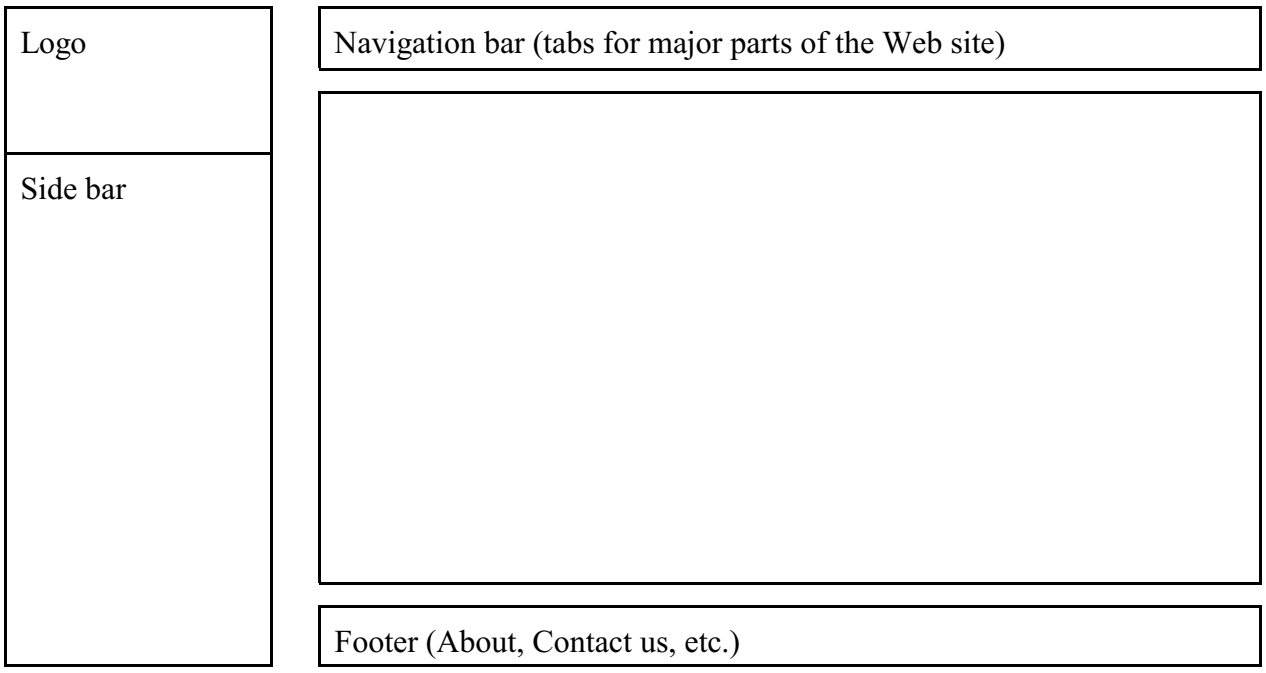
A course in advanced methods in XML, including XLink, XPointer, XPath, and XSLT (eXtensible Stylesheet Language for document Transformation)

Web templates

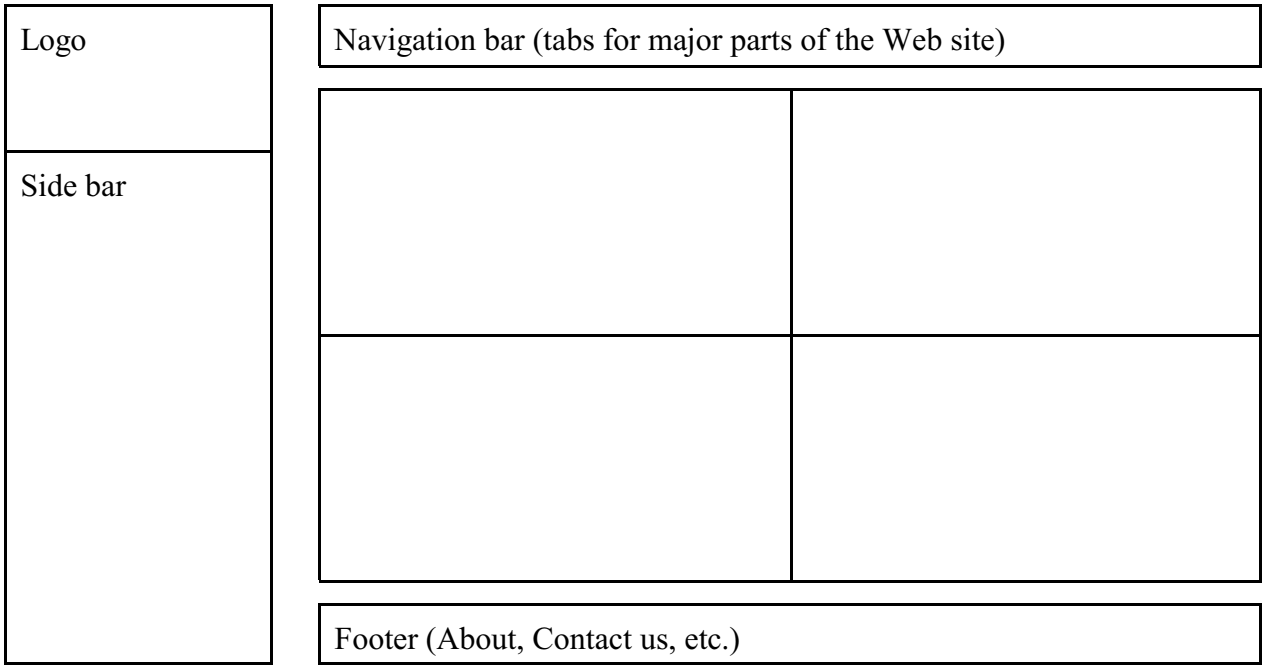
Web templates are very useful for creating and maintaining Web pages. The templates shown have slots and fillers just like the memo templates, but focus here is on the display area allocated to the data in each slot. The underlying definition is not shown.

When one of these templates is used to create a Web page, each area of the template is linked to a document to be displayed in that area. This may be any multimedia document (text, image, or combination) or it may be report that is created dynamically from a database. When a document is updated, the update is reflected immediately in all Web pages that link to the document.

Web template 1



Web template 2



Hypermedia/hypertext ➜ LIS 506 Information Technology

Linear text vs. hypertext	**Typical text is linear in a sequence set by author**: "Begin at the beginning," the King said, very gravely, "and go on till you come to the end: then stop." Lewis Carroll, Alice in Wonderland, Chapter XII **Hypertext / hypermedia is a collection of text pieces with links; the reader can and often must establish her own order through the text** (if indeed the reader goes through the text); this is accomplished by treating the text in blocks (or at least by establishing nodes/locations within the document) and by supplying/permitting links between nodes by which the reader can navigate the text in his or her own order. One could also say that the reader constructs his or her own text. A hypertext can include suggested linear sequences, often indicated by <next> and <previous>.
Major features of hypertext	• fragmented non-linear text form whose components can be rapidly accessed via machine-supported links/relationships under direction of user • interactive • malleable, modular: it is easy to add or revise small pieces • no strong document boundaries (at least in large hypertexts)

Hypertext examples	
World Wide Web	Primary example of hypertext: the **World Wide Web**, in which documents/sites typically have links to other documents/sites; it is the presence of these links that gives the web metaphor. The functionality of hypertext has existed long before, for example, in the form of the research paper with footnotes/bibliography/tables/figures, although WWW makes these links convenient to use.
Bible in hypertext format	Links from chapter, group of verses, verse, or word to "Original" version(s); manuscript image(s) Alternative translations Other Bible passages Commentary passages Sermons about the passage (published or own) Entry or subentry in Hebrew/Greek dictionary/grammar Map Archaeological evidence
Fiction examples	"interactive fiction," "Choose your own adventure"

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?

Inter-document structures

Relationships between works (from Dr. Green)	Continuations and sequels Answer key Parodies Critical reviews Concordances	Abstracts Indexes Bibliographies Guides to literature Translation
	These are often mentioned in cataloging rules). More examples were presented in the earlier reading Soergel, *Integrated information structure interface*.	

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.

Lecture 6.2a (50 min) ***February 23***

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

Objectives	Inherited from Lectures 5.2 - 6.2 Gain a feel for good document design where the external form conveys the internal structure well.
Practical significance	Inherited from Lectures 5.2 - 6.2 To provide users with documents, Web sites, screens, and other information representations that optimally support understanding, you must be able to select or create such documents or other representations.

Some principles for good document design

Know the reader	Problem to be solved / task to be accomplished Information need Background knowledge
Content	Select the information carefully - only what the reader needs to know. Avoid redundancy or use it purposefully
Structure	**Elaborate in your own mind the intrinsic structure of the topic** / the phenomena to be presented - good document design is grounded in a thorough understanding and structuring of the topic. **Choose the external representation structure that best facilitates the assimilation of the intrinsic structure — form follows function.** Examples of external representation structures: Plain text, typographically structured text (such as a list or a linear arrangement of a hierarchy), table, diagram, picture. **A general structure that is often useful** Make schemas explicit. Provide advance orienters ("Tell them what you are going to tell them"). Give the detail ("Tell them"). Provide opportunity for rehearsal or application to fix the new information in the reader's mind. ("Tell them what you told them.")
Layout	Provide guideposts that indicate the overall context . (for example, running heads, navigation chains as in Yahoo). Point out relationships (cross-references, links). In documents intended for looking things up, such as a dictionary: Provide guiding headings at the top of each page. (Counterexamples: Library of Congress Subject Headings, Dewey Decimal Classification index, MeSH.)

High-level means of expression, media modalities

	Non-linguistic (depicting)	**Linguistic** (text, verbal, convention)
Auditory Hearing/audio/ sound	Sound Music "Audons"	Spoken language Speech
Visual Sight/vision/ graphical	Images, pictures (photos, paintings, drawings, charts, diagrams) "Visuals" (including real objects and models) Still Moving Icons Pictograms Sign language	Written (printed) language "Text" in Information Retrieval
Touch/tactile	Tactile representations, for example a three-dimensional map	Braille
Other senses	Smell is of little practical significance here. The kinesthetic senses do not apply (except perhaps in virtual reality applications)	
Audiovisual	Sound and images simultaneously (as experienced in real life), often includes speech, may include written text.	
Multimedia, hypermedia	Combination of pieces of presentations in different modalities: Display of written text may be followed by still image, or a series of still images with explanation in speech, which in turn may be followed by an audiovisual segment. Multimedia kit. Hypermedia — interlinked segments in several modalities	

Consider	Medium Arsenal of artistic expression, visual vocabulary (icons, symbols in comic strips) Image structure Style	Language (Chinese, English, etc.) Vocabulary Text structure Style

Low-level means of expression

Typography: Type face, type size
Highlight or lead symbol (triangle, bullet, square, pointing hand, etc.)
Graphical means for highlighting or de-emphasizing (often used to distinguish between options that are available at the moment and those that are not) Bold, blinking, reverse on different background, black vs. gray
Color (but 8% of the population are color blind) (also for emphasis)
Boxes and other means of grouping

Methods for indicating parts of a document (large or small)	
Explicit labels Arrangement Type face	See the examples for different methods of displaying a catalog record in *Organizing Information*, p. 160 - 161.

Further elaboration of these principles through a series of examples

Document design examples

1	Two formats for salary data
2	Alphabetical vs. meaningful display (Art and Architecture Thesaurus)
3	Alphabetical vs. meaningful display (Art and Architecture Thesaurus)
4	Examples from the Longman Lexicon of the English Language
5	Display of information on buildings on a site in Perseus
6	Two displays of the same hierarchy
7	Winners and losers in the forecasting game (from Tufte)
8	Thermal conductivity of tungsten (from Tufte)
9	Napoleon's campaign to Russia (from Tufte)

Optional

10	Classified arrangement of descriptors in a document record for indexing test (Alcohol and Other Drug Thesaurus)
11	Contents page from *Alcohol Research*

Note

The syllabus and lecture notes are an example of document design, using boxes, labels, comparative columns, and color and striving for consistent format. For example, first pages of lectures follow a common format, so do first pages of assignments.

Example 1. Library Jobs by Level, ALA survey 2008

2008 *ALA-APA Salary Survey: Librarian – Public and Academic* (*Librarian Salary Survey*)

Job title	Average salary
Director/Dean/Chief Officer	
Public Libraries	86K
Academic Libraries	95K
Deputy/Associative/Assistant Director	
Public Libraries	73K
Academic Libraries	80K
Dept Head/Branch Mgr/Coordinator/Senior Mgr	
Public Libraries	61K
Academic Libraries	61K
Manager/Supervisor of Support Staff	
Public Libraries	52K
Academic Libraries	54K
Librarian Who Does Not Supervise	
Public Libraries	48K
Academic Libraries	55K
Beginning Librarian	
Public Libraries	43K
Academic Libraries	45K

http://www.ala-apa.org/salaries/SalarySummary2008.pdf (Data extracted from Tables 1 and 2)

Same data, different arrangement

Job title	Public	Academic
Director/Dean/Chief Officer	86K	95K
Deputy/Associative/Assistant Director	73K	80K
Dept Head/Branch Mgr/Coordinator/Senior Mgr	61K	61K
Manager/Supervisor of Support Staff	52K	54K
Librarian Who Does Not Supervise	48K	55K
Beginning Librarian	43K	45K

Examples 2 and 3. From the Art and Architecture Thesaurus (AAT)

<size: photograph formats> double whole plate half plate mammoth plate ninth plate quarter plate sixteenth plate sixth plate whole plate	**size: photograph formats** sixteenth plate ninth plate sixth plate quarter plate half plate whole plate double whole plate mammoth plate
Art and Architecture Thesaurus sequence	**Suggested meaningful sequence**

Figure 1. Alphabetical vs. meaningful sequence on same hierarchical level
(Art and Architecture Thesaurus)

In the **art genres** example on the next page, notice the advantage of having definitions / scope notes for related terms right next to each other.

<art genres>
- academic art
- amateur art
- apocalyptic art
- art brut
- children's art
- commercial art
- community art
  - SN Includes art undertaken in conjunction with particular communities, often socially deprived, usually with the idea of producing an effect or inspiring response specifically within those communities, with no reference to widely established standards. For art intended to beautify or enrich public places, use **public art**.
- computer art
- court art
- crafts
- cybernetic art
- didactic art
- dissident art
- ethnic art
- fantastic art
- figurative art
- folk art
- funerary art
- naive art
- nonrepresentational art
- primitive art
- public art
  - SN Use for art whose purpose is to beautify and enrich public places. For art undertaken in conjunction with particular communities, usually to produce an effect or inspire response specifically within those communities, use **community art**.
- rock art
  - cave art
- serial art
- sofa art
- street art

a. **AAT sequence**

art genres

. **art genres by content or other intrinsic characteristics**

. . figurative art
. . . fantastic art
. . . apocalyptic art
. . nonrepresentational art
. . cybernetic art
. . serial art
. . crafts

. **art genres by standard**

. . academic art
. . folk art
. . dissident art

. **art genres by type of artist or origin**

. . amateur art
. . naive art
. . art brut
. . children's art
. . computer art
. . ethnic art
. . primitive art

. **art genres by audience, purpose, or display context**

. . sofa art
. . court art
. . public art
. SN Art whose purpose is to beautify and enrich public places.
. . . community art
SN Public art undertaken in conjunction with particular communities, often socially deprived, usually with the idea of producing an effect or inspiring response specifically within those communities, with no reference to widely established standards.
. . . street art
. . rock art
. . . cave art [prehistoric, esp. paleolithic]
. . didactic art
. . commercial art
. . funerary art

b. **Suggested meaningful sequence**

Figure 2. Alphabetical vs. meaningful sequence.

Example from the Art and Architecture Thesaurus (AAT)

Example from Longman

Example 5.

Results of a search for architecture (buildings) whose site is "Amphiaraion" in the region of Attica (from an old version of Perseus)

Name	Summary	Period	Type
Amphiaraion, Earlier Temple of Amphiaraios	Small temple; on the western end of the Terrace of Dedications in the Sanctuary of Amphiaraios	Late Clas./Hell.	Temple
Amphiaraion, Klepsydra	Water clock and small annex; southeast of the Sanctuary of Amphiaraios, across the stream and east of the temple of Amphiaraios	Hellenistic	Klepsydra
Amphiaraion, Stoa	Stoa; on the east side of the Sanctuary of Amphiaraios, southeast of Theatre	Late Classical	Stoa
Amphiaraion, Temple of Amphiaraios	Temple; at the Western end of the Sanctuary of Amphiaraios	Hellenistic	Temple
Amphiaraion, Terrace of Dedications	Terrace with retaining wall; on the northwestern side of the Sanctuary of Amphiaraios	Late Classical	Terrace
Amphiaraion, Theater	Hellenistic theater; on the northwestern side of the Sanctuary of Amphiaraios, behind the west half of the Stoa	Hellenistic	Theater

Note:
In a display of all theatres, we could omit the column Type but should add a column Region.

Hierarchy from facets, bad example (replaces OCLC Prism)

Hierarchy from facets good example

Tufte, some winners

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

Optional.

Example 10

The **Alcohol and Other Drugs Thesaurus (AOD Thesaurus)** provides many examples of meaningful sequence (Wasserman Library, cataloging tools area).

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*

Contents page from Alcohol Research

More document design examples

Result display in the catalog of the Montgomery County Public Libraries using Sirsi Dynix
www.montgomerycountymd.gov/libtmpl.asp?url=/content/libraries/find/findbooks.asp

J 599.789 BRE 2006
Giant pandas up close
Bredeson, Carmen.
24 copies available at Aspen Hill Library, Chevy Chase Library, Damascus Library, Davis Library, Fairland Library, Gaithersburg Library, Germantown Library, Kensington Park Library, Noyes Children's Library, Olney Library, Poolesville Library, Potomac Library, Quince Orchard Library, Rockville Library, Silver Spring Library, Twinbrook Library, White Oak Library, and Longbranch Library

Revised design

J 599.789 BRE 2006
Giant pandas up close
Bredeson, Carmen.
24 copies available at

Aspen Hill	Chevy Chase	Damascus	Davis	Fairland	Gaithersburg
Germantown	Kensington	Park Noyes	Olney	Poolesville	Potomac
Quince Orchard	Rockville	Silver Spring	Twinbrook	White Oak	Longbranch

Lecture 6.2b (15 min) ***February 23***

(Very brief, see ➜506 for more detail)

Formatting documents for interpretation by computer programs. Document markup languages

HTML (Hypertext Markup Language) and XML (eXtensible Markup Language) / SGML (Standard Generalized Markup Language)

Objectives (for treatment of subject in ➜506)	1 Understand the principles of markup languages and their importance for the implementation of good document design as a basis of further study. 2 Be able to create simple Web pages using HTML markup.
Practical significance	Databases of machine-readable text are undergoing an unprecedented explosion, not only on the Web, but also in intranets and in efforts of creating large text corpora for linguistic and literary studies (the Text Encoding Initiative). Conventions for marking the structure of documents are a prerequisite for creating such databases and for common access and data exchange. Most students will find themselves in situations where they need to access such texts and assist users in the further processing such texts; some students might participate in the setup of text databases. Note: A text corpus is simply a (usually large) body of text in digital form, often with annotations, such as indicating the meaning of each homonym. Examples: The Brown corpus, *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*, originally created in 1964 at Brown University and updated several times, see http://helmer.aksis.uib.no/icame/brown/bcm.html Also see http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/corpus4/4fra1.htm Beyond marking up text, markup languages are now expanded to specify any kind of data structure, blurring the boundary between text and formatted data.

Outline

Brief introduction and basic principles

Definition and general introduction
Principles
HTML, XML, and SGML

Examples

HTML example
- a - c Template definition, XML document, stylesheet – not needed
- d Document with HTML tags ("under the hood"), done directly by the author
- e Document displayed

XML example
- a Template definition
- b Document with XML tags
- c Style sheet defining appearance
- d Document with HTML tags ("under the hood") produced from XML document
- e Document displayed

Brief introduction and basic principles

Definition and general introduction

Definition	Markup is the insertion of tags (codes) into a document text or other data stream to specify a structure which can then be used for further processing, in particular for controlling the appearance (or rendering) of a document when it is printed or displayed on a screen. Note: The term *markup* derives from typesetting. An editor put marks in a manuscript that specified for the typesetter the fonts to be used for a portion of text and other matters of appearance. The meaning of the term has much expanded since then, particularly in the last few years.
General introduction	HTML markup tags are designed primarily to direct the display of documents. HTML tags also specify links to other documents to be included automatically at display time (such as images) or available to the user by clicking on the link symbol. Tags defined through XML are much more powerful for expressing document and data structure.

Principles

Physical markup	Tags specify actual appearance properties, such as *Indent .3"*, *Center*, *Bold*, *Font Times Roman 12* Problem: What if display device cannot show Times Roman?
Logical markup	Tags specify the logical structure of the document, including importance of certain pieces of text. The display is done by a program, possibly in conjunction with style sheets, that renders logical elements in a format determined at output time. **Formal (or syntactic) logical elements**. Tags specify formal units such as *heading level 1*, *paragraph*, *numbered list*, *emphasize* **Content logical elements**. Tags specify content units such as *From*, *to*, *subject*, *Recommendations*, *warning*, *methods*, *conclusion* defining the content structure of a document. These tags can be used to define record formats even for highly structured data. (XML is used increasingly as a language to define the structure of data in Web-based database applications.) The **display program then determines the physical appearance** in accordance with the capabilities of the display device and the preferences set by the user. **Examples:** A *heading level 1* may appear in Times Roman 16 pt bold or in all caps. A new *paragraph* may start with a blank line and no indention (block style) or without a blank line with the first line indented. The document element *warning* may be displayed in a box with light gray background and a heading **Warning**. Since logical content markup makes the logical structure of a document explicit, it can be used for information organization and retrieval as well. It can be used to define record formats for straightforward data to be processed by a database management system or to define templates for complex documents (see the examples in Lecture 6.1b). Organizations use markup languages defined in XML to organize large databases of document content, including text and images.

HTML, XML, and SGML

HTML	HTML is a markup language; all tags are predefined. HTML emphasizes logical markup, but the logical elements are primarily formal, and HTML includes an increasing number of physical markup tags (but still not enough to provide tight control over the appearance of a page). An author uses HTML tags to describes the way she wishes the page to display, but parsing and interpretation of the HTML tags is dependent on the Web browser used to display the page. The browser may or may not implement all the features in the same way. For example, look at a complex web page side-by-side with Internet Explorer and Netscape/Firefox/Mozilla.
XML	XML (see the main XML Web site at www.w3.org/XML/) is not a markup language but a language that can be used to define one's own tags, one's own markup language; XML is a **markup metalanguage**: there are no predefined tags; authors and system administrators define their own tags. Many specific markup languages can be defined using XML This makes it possible to represent more of a document's semantic structure than HTML does. HTML is one of many markup languages that can be defined in terms of XML. **Standards expressed in terms of XML**. There are many domains where multiple users have similar kind of documents. They need a format for structuring these documents and for metadata describing them. There are many communities that use XML to define markup languages (domain-specific tagging schemes) for their own domain (with discussion in the whole user community) as a standard to be used by the community; examples are MathML, NewsML, HR-XML for human resource data, etc., financial documents or biological processes (tags for structuring data). This saves thousands of people from having to "reinvent the wheel" for their domain.
SGML	SGML (Standard Generalized Markup Language) is a markup metalanguage that was developed primarily by the publishing industry so they could deal more easily with electronic manuscripts. It was created to allow sophisticated and detailed markup for every need of book publishing. The feature richness (or feature excess) made it very difficult for programmers to write practical software for processing SGML documents, so XML started out as a simplified subset of SGML (20% of the complexity, 80% of the functionality). XML has since added features of its own, especially the definition of many data types (such as date and currency) to support databases encoded using XML. Strictly speaking, HTML was defined using SGML as the defining language.

Examples

HTML example (With XML, more steps are required, see XML process diagram)

Document with HTML tags ("under the hood") (File d in the XML process diagram)

```
<HTML>
   <HEAD>
      <TITLE>What XML can do for us</TITLE>
      <META NAME="creator" CONTENT="Bob Boiko">
      <META NAME="keywords" CONTENT="XML; content management; document structure;
      databases on the Web">
      <META NAME="GENERATOR" CONTENT="" >
   </HEAD>
   <BODY>
      <H1><Center> Memorandum </Center></H1>
      To: Sue Feldman, CIO <BR>From: Bob Boiko<BR>
      Date: February 7, 2003<BR><BR>
      Subject: <EM> What XML can do for us</EM>
      <P>XML allows us to define document structures that will make it easier to create
      documents.  Once a document is created, it can be displayed in many different ways (Web
      page in multiple formats, print, etc.) through applying style sheets (the simple Cascading
      Style Sheets, CSS2, or the more powerful eXtensible Stylesheet Language for document
      Transformation, XSLT).  A table of contents can be created automatically.  Moreover, the
      document can be displayed selectively using just the parts most appropriate for a given
      audience. Parts of one document can be reused in another document. . . .</P>
   </BODY>
</HTML>
```

Document displayed by the Web browser under the control of HTML tags (File **e** in the diagram)

Memorandum

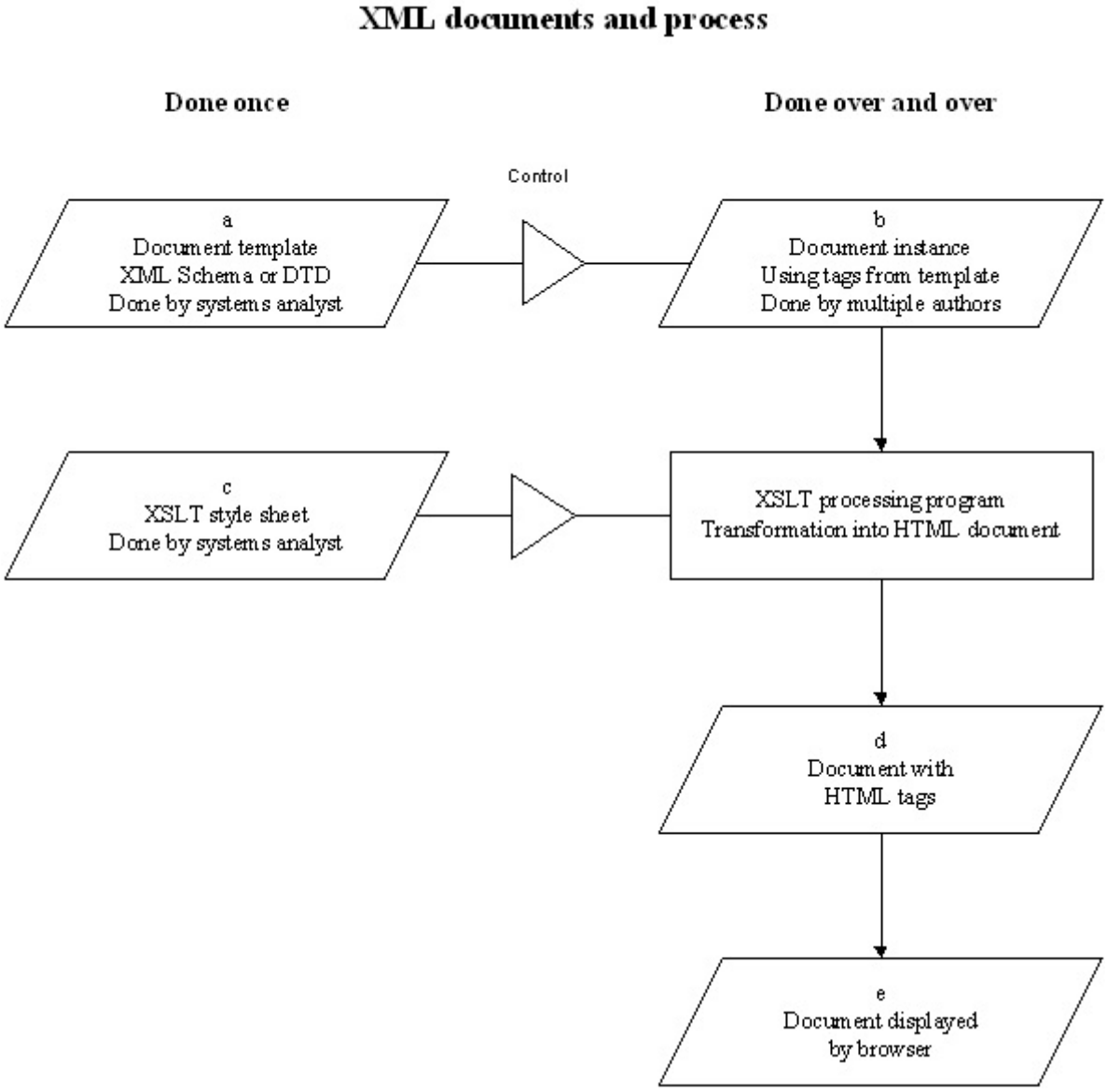
To: Sue Feldman, CIO
From: Bob Boiko
Date: February 7, 2003
Subject: **What XML can do for us**

XML allows us to define document structures that will make it easier to create documents. Once a document is created, it can be displayed in many different ways (Web page in multiple formats, print, etc.) through applying style sheets (the simple Cascading Style Sheets, CSS2, or the more powerful eXtensible Stylesheet Language for document Transformation, XSLT). A table of contents can be created automatically. Moreover, the document can be displayed selectively using just the parts most appropriate for a given audience. Parts of one document can be reused in another document. . . .

XML example

In XML one must first create an **XML template or schema** which specifies tags for the parts of a document (and thus document structure), in the example for a document type (or class) called *memo*. The example assumes that the memo schema is stored at www.jasca.com/cm/memo.xsd An XML schema is itself a document that follows XML syntax and tags defined by the W3C at the URL www.w3.org/2001/XMLSchema These tags form a **name space**. To make sure that these tags do not conflict with tags by the same name defined by somebody else, they are prefixed by xsd: in the example (xsd: is declared as the prefix for the name space defined at the URL).

Most document structure definitions still use a document type definitions (DTD), but XML schemas are more powerful and will replace DTDs. The XML schema syntax is defined in the W3C Recommendation XML Schema (appr. May 2, 2001).



a Definition of template for document *memo* (done once by systems analyst, defines tags)

An XML schema defines a document structure and identifies each element of the structure by a tag. This XML code creates a **memo template or schema**. The documents in the memo class must contain one top-level element, *memo*, which in turn consists of two subordinate elements, *metadata* and *memoBody* (exactly one of each in this order), which in turn contain subordinate elements.

```
<?xml version="1.0"?>
<xsd:schema xmlns:xsd="www.w3.org/2001/XMLSchema">
<!–- w3 schema file defines an XML name space; we use prefix xsd. –>
  <xsd:element name="memo" type="memoType"/>
  <xsd:complexType name="memoType">
      <xsd:sequence>
          <xsd:element name="metadata" type="metadataType"/>
          <xsd:element name="memoBody" type="memoBodyType"/>
      </xsd:sequence>
  </xsd:complexType>
  <xsd:complexType name="metadataType">
      <xsd:sequence>
          <xsd:element name="to" type="xsd:string"/>
          <xsd:element name="from" type="xsd:string"/>
          <xsd:element name="subject" type="xsd:string"/>
          <xsd:element name="date" type="xsd:date"/>
          <xsd:element name="keywords" type="xsd:string"/>
      </xsd:sequence>
  </xsd:complexType>
  <xsd:complexType name="memoBodyType">
      <xsd:sequence>
          <xsd:element name="plainText" type="xsd:string"/>
      </xsd:sequence>
  </xsd:complexType>
</xsd:schema>
```

b A document instance of type *memo* (done over and over by authors, uses tags defined in memo template)

```
<?xml version="1.0"?>
<?xml:stylesheet type=”text/XSLT”
      xlink:href=”www.jasca.com/cm/memo.xslt”?>
<memo xmlns=”www.jasca.com/cm/memo.xs”>
      <metadata>
            <to>Sue Feldman, CIO</to>
            <from>Bob Boiko</from>
            <subject>What XML can do for us</subject>
            <date>February 7, 2003</date>
            <keywords>XML; content management; document structure; databases on the Web"</keywords>
      </metadata>
      <memoBody>
            <plainText>XML allows us to define document structures that will make it easier to create
            documents. Once a document is created, it can be displayed in many different ways ...</plainText>
      </memoBody>
</memo>
```

c XSLT style sheet (Controls the display of the document.) An XSLT processor program uses XML tags to identify pieces of data.. Determines selection of data to be displayed, their arrangement, and the appearance of each element. In the example, the output is an HTML document. But many other types of formatting are possible, e.g., to Wireless Markup Language (WML) for display on a handheld device.

```
<xsl:stylesheet
  xmlns:xsl="www.w3.org/TR/WD-XSLT"
  xmlns="www.w3.org/TR/REC-html40">
  <xsl:template match="/">
     <HTML>
        <HEAD>
        <TITLE><xsl:value-of select="memo/metadata/subject"/></TITLE>
           <META NAME="creator" CONTENT="{<xsl:value-of select="memo/metadata/from"/>}"/>
           <META NAME="keywords"
              CONTENT="{<xsl:value-of select="memo/metadata/keywords"/>}"/>
           <META NAME="GENERATOR" CONTENT=""/>
        </HEAD>
        <BODY>
           <H1><Center> Memorandum </Center></H1>
           To: <xsl:value-of select="memo/metadata/to"/><BR/>
           From: <xsl:value-of select="memo/metadata/from"/><BR/>
           Date: <xsl:value-of select="memo/metadata/date"/><BR/><BR/>
           Subject: <EM><xsl:value-of select="memo/metadata/subject"/></EM>
           <P><xsl:value-of select="memo/memoBody/plainText"/></P>
        </BODY>
     </HTML>
  </xsl:template>
</xsl:stylesheet>
```

d HTML document (same as above) (done over and over, produced by an XSLT processor program)

```
<HTML>
   <HEAD>
      <TITLE>What XML can do for us</TITLE>
      <META NAME="creator" CONTENT="Bob Boiko">
      <META NAME="keywords" CONTENT="XML; content management; document structure;
      databases on the Web">
      <META NAME="GENERATOR" CONTENT="" >
   </HEAD>
   <BODY>
      <H1><Center> Memorandum </Center></H1>
      To: Sue Feldman, CIO <BR>From: Bob Boiko<BR>
      Date: February 7, 2003<BR><BR>
      Subject: <EM>What XML can do for us</EM>
      <P>XML allows us to define document structures that will make it easier to create
      documents. Once a document is created, it can be displayed in many different ways. . . .
      </P>
   </BODY>
</HTML>
```

e Document displayed by the Web browser under the control of HTML tags, see on p. 218

XML schema for a self-assessment memo (see p. 190/191)

Since a s*elf-assessment memo* is a specific type of memo, we can define its schema by adding to the *memo* schema; the *memo* schema is **reused**.

```
<?xml version="1.0"?>
<xsd:schema xmlns:xsd="www.w3.org/2001/XMLSchema"
   xmlns=”www.jasca.com/cm/memo.xsd”>

   <include xsd:schemaLocation=”www.jasca.com/cm/memo.xsd”/>
      <!--This schema includes a definition of the type metadataType, which is used below. -->

   <xsd:element name="selfAssessmentMemo" type="selfAssessmentMemoType"/>

   <xsd:complexType name="selfAssessmentMemoType">
      <xsd:sequence>
         <xsd:element name="metadata" type="metadataType"/>
         <xsd:element name="memoBody" type="memoBodyType"/>
      <xsd:sequence>
   </xsd:complexType>

   <!– redefinition of memoBodyType –>
   <xsd:complexType name="memoBodyType">
      <xsd:complexContent>
         <xsd:extension base="memoBodyType">
            <xsd:sequence>
               <xsd:element name="accomplishments" type="xsd:string"/>
               <xsd:element name="goals" type="xsd:string"/>
               <xsd:element name="trainingNeeds" type="xsd:string"/>
            </xsd:sequence>
         </xsd:extension
      </xsd:complexContent>
   </xsd:complexType>
</xsd:schema>
```

Reuse is a big theme in the application of XML; reuse can be implemented in several ways.. There are whole collections of defined data types, such as a data type for US states, with the values restricted to a list of two-letter abbreviations of US states, or data types for US address, UK address, France address, Germany address, etc. (all derived from a generic address as a common parent). These type definitions are collected into *vocabularies*, from which they can be included in any XML document schema, saving the schema creator a lot of work.

Note: The syntax of the XML examples may not be correct in every detail, but it does give the general idea.

There is supplemental material on XML and RDF at www.dsoergel.com/670/SYL2003FaLecturesAppendixNew.pdf

In particular a fully worked out example of using RDF to represent the food data from Lecture 2.2 and using style sheets to create several different outputs (a table of contents, a detailed listing, and an alphabetical index) from this database.

Document example 5: Self assessment memo (p. 190/191)

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*.

February 23. 2011 Name (optional)

Free-write 6

Lecture 6.1a. Natural language processing. Syntactic and semantic parsing

Lecture 6.1b. Document macrostructure & inter-document relationships

Lecture 6.2a. Document design (information design)

Lecture 6.2b. HTML and XML

- **Reflect** – what you learned, what was most important, what was most interesting, what was extraneous;
- **Ask questions** – ask for more explanation, how is a concept connected to other concepts, why is a concept important, how can it be applied, why is a reading important;
- Offer **critique and suggestions**;
- Say anything else you want to.

O
v
e

r

Lectures 7.1- 7.2.
Cataloging and metadata. Bibliographic and record control

(no text chapter) ***March 2***

Objectives	1. Understand the use of metadata for finding and interpreting or using any kind of data source (in some interpretations: any kind of object). 2. Understand the fundamental problems of bibliographic control as an application of general principles of Organization of Information. 3. Understand the problems of defining "document", and the problems of defining the relationships between several versions of a document. 4. Be able to apply this understanding to the analysis and design of cataloging codes and to actual cataloging (consulting the appropriate code for details). 5. Understand the complexities in determining the useful entries for a document. 6. Be able to apply some AACR2 rules for entry. 7. Have a general idea of the use of XML and RDF (Resource Description Framework) for expressing and implementing metadata schemas.
Practical significance	• Good catalogs are more important than ever, just look at the World Wide Web. It is now fashionable to call data about documents (i.e., data about data-carrying objects) *metadata*. • The question "what is a document" is important for library catalogs but even more important for electronic records, where different versions of the same document proliferate rapidly, especially on the Web. A Uniform Resource Locator or URL does not identify a document but a file storing a document. Many files in different locations can store the same document, creating a burden on the user. There are efforts to define a **Uniform Resource Identifier** or **URI** that would identify a **document** no matter where it is stored (like an ISBN). A URI identifies a *document, intellectual work,* a URL identifies a *document, physical volume* (as defined in Chapter 3). However, a system of URIs, while beneficial to users, introduces many difficulties: Who would assign URIs, using what rules? Who would maintain the database(s) with links from each URI to all its physical volumes (URLs)? What happens if one of these files is slightly modified? • Controlling different versions of a document is important for managing document production and access afterwards, as well as for reasons of legal and historical evidence. • Good catalogs that are widely usable and that can share data require standardized cataloging rules. In the World Wide Web domain, there are efforts to agree on a metadata standard. (A minimal standard, the *Dublin Core* has already found wide acceptance.) In the domain of geographic data there is the *Content Standard for Digital Geospatial Metadata* issued by the Federal Geographic Data Committee.

Note: Component lectures do not have pink sheets.

General introduction to metadata

Metadata – Synonymous with cataloging data or pointer data as defined in Organizing Information, Chapter 2.
Data used to **describe other data** and **give context for other data** for the following purposes:

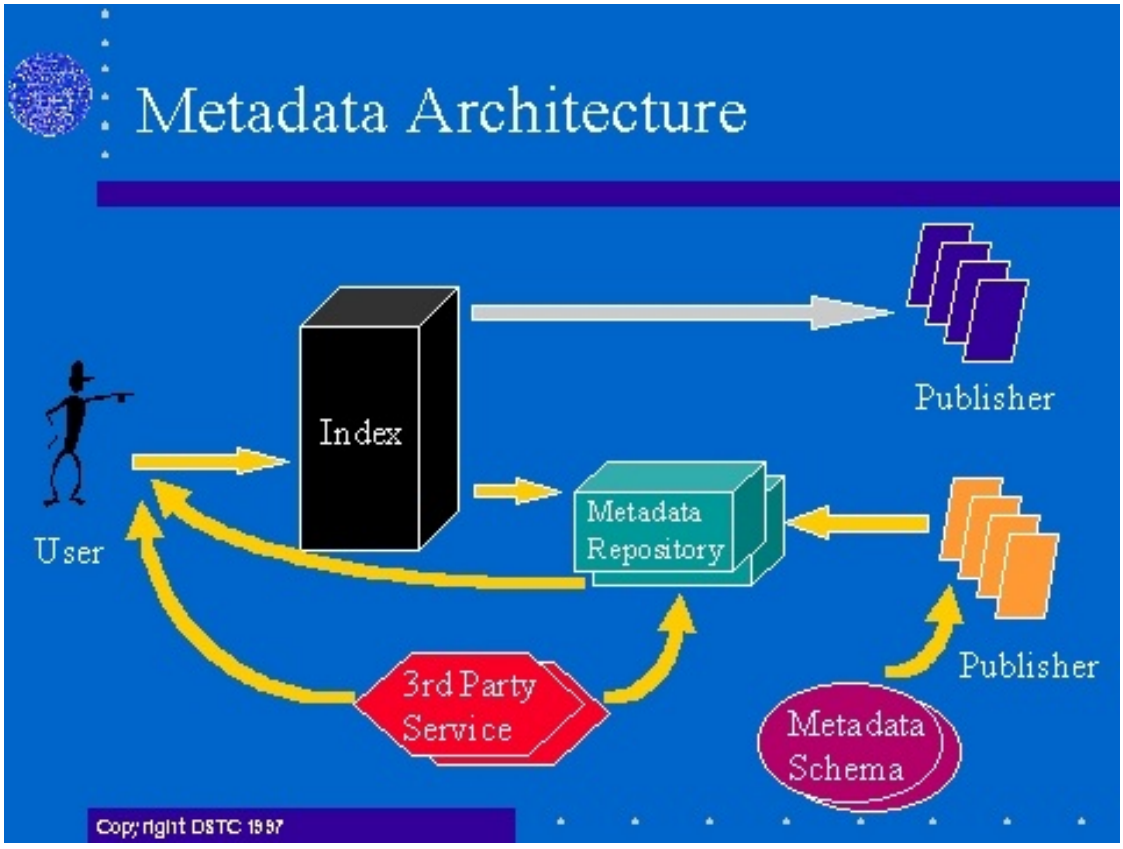
- retrieval
- assessment
- interpretation and use.

Now a hot topic in the context of the Web. Metadata schemas exist or are being developed for many types of "data containers" in the context of many user communities. See examples below.

Note: There is no intrinsic difference between data and metadata. If data are used for the purpose of retrieving, assessing, interpreting, or using other data, they are used as metadata. The data modeling mechanisms are the same no matter what the data modeled are used for.

Two slides of Renato Iannella adapted by DS.

- **Metadata will improve information discovery across digital repositories**
  - metadata schema registries (metadata schema = conceptual schema or record format)
  - metadata transmission
  - metadata repositories
  - metadata mappings (for example, from Dublin Core to MARC)
  - semantic interoperability (for example, mapping from DDC to LCC)



Examples of metadata schemas (many expressed in XML)
After the lecture, look at **one** of the examples marked with * ; use URL given in the email I sent.

Bibliographic data: **MARC / AACR2. Dublin Core** http://dublincore.org/ (see below)

*** Text**: **The Text Encoding Initiative (TEI)**
A comprehensive standard for describing (literary) texts, both metadata and the structure of the actual text, www.tei-c.org/

***Archival data**: **Encoded Archival Description (EAD)**
www.dlib.org/dlib/november99/11pitti.html
www.loc.gov/ead/
For an example repository using EAD see www.cdlib.org/inside/projects/oac/

***Museum data**:
The CIDOC Conceptual Reference Model (CRM), related to FRBR
Introduction: http://cidoc.ics.forth.gr/docs/cidoc_crm_meeting_Prato-1.ppt
Full specification: http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.0.pdf

***Learning objects (instructional materials)**:
The Gateway to Educational Materials (GEM)
www.thegateway.org/about/documentation/metadataElements
Learning Technology Standards Committee of the **IEEE**:
http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf
IMS Global: IMS learning resource meta-data information model. (September 2001)
www.imsproject.org/metadata/
Dublin Core Metadata Initiative. DCMI Education Working Group,
http://dublincore.org/groups/education/ (not much concrete to see there)
CRP Henri Tudor-CITI: Training Exchange Definition: TED.
www.xml.org/xml/schema/8dbca03a/trainingExchangeDefinition.pdf (July 2002)

Geospatial data: will be covered later. Ask me if you are interested

TV programs: TV Anytime Forum. TV Anytime is a set of specifications for the controlled delivery of multimedia content to a user's personal device (Personal Video Recorder (PVR))
www.tv-anytime.org (includes dealing with metadata);
http://portal.etsi.org/radio/TVAnytime/TVanytime.asp
BBC: www.bbc.co.uk/rd/pubs/whp/whp-pdf-files/WHP050.pdf
European Broadcasting Union (EBU): http://www.ebu.ch/metadata/pmeta/v0102/xml/
www.ebu.ch/metadata/pmeta/WIP/ESCORT/ESCORT2006.htm (A faceted classification)

Multimedia: **MPEG-7**,
"Multimedia Content Description Interface", is a standard for describing features of multimedia content: catalog data (e.g., title, creator, rights), semantic data (e.g., the who, what, when, where information about objects and events) and structural data (e.g., the color histogram - measurement of the amount of color associated with an image or the timbre of a recorded instrument). Builds on AV data representation defined by MPEG-1, 2 and 4. www.mpeg.org/MPEG/starting-points.html

PICS (Platform for Internet Content Selection) properties. www.256.com/gray/docs/pics/

Lecture 7.1b (40 min) ***March 2***

Bibliographic control. General issues

Introduction	This lecture deals with control of all kinds of documents (all kinds of materials): Regular books and reports, serials, journal and newspaper articles, organizational records, images, sound documents. New dimension of problem: Electronic documents. Ease of copying and modification, cryptic filenames, and online accessibility of electronic documents create special difficulties. A number of general principles of Organization of Information are applicable to the control of any kind of concrete object or "thought object." Each type of material presents its own challenging problems in applying these general principles. Parts of the thinking on descriptive cataloging and the resulting practices are still valid. Other parts have been made obsolete through the greater power of automated systems. Control is mainly access, but also inventory control, including preservation. A distinction is generally made between description and access, but the two are more closely intertwined than many people realize.
What is a catalog?	A catalog is a database that contains identifying/descriptive data about objects, such as books (or, more broadly, documents) or data sets (such as geospatial data sets) or merchandise. The coverage of a catalog may be limited to a given physical collection (the books for which physical copies are held in the library, the merchandise items available from a catalog store); that is, the catalog contains only data referring to objects in the given collection. Often the term *catalog* is defined in this sense of being tied to a physical collection as distinguished from a *bibliography*, which may include data about documents no matter where physical copies are held. A *union catalog* refers to objects in multiple collections. If the objects referred to in the catalog are information sources, the catalog data are used primarily as *pointer data* (see Section 2.5 of Organizing Information,), now commonly called *metadata*. However, remember, that the distinction between substantive data (data contributing directly to the problem solution) and pointer data lies not in the nature of the data themselves but rather in their use. If data are used to find other data/information sources, they are used as pointer data. If data are directly applied to problem solution, they are used as substantive data.

Objectives of the library catalog according to Cutter

1. To enable a person to find a book for which either
   A. the author
   B. the title
   C. the subject
   is known

2. To see what a library has
   D. by a given author
   E. on a given subject
   F. in a given kind of literature

3. To assist in the choice of a book
   G. as to its edition (bibliographically)
   H. as to its character (literary or topical)

Problems with Cutter's objectives

(a) Is the user interested in a particular *book* or in the *work* that is embodied in the book? And what is a *book* anyhow?

(b) In today's world of electronic access, what is a library?

Need to address these before restating objectives.

Fundamental problem in bibliographic control: What are the units we are dealing with?

Look at the examples on the following pages to get a feel for the problem.

Sample documents illustrating problems in defining bibliographic units

Next page

Sample documents illustrating problems in defining bibliographic units

(1) *The man I killed*, by Michael Halliday (i.e. John Creasey). London: Marx Brothers; 1935.

(2) *The man I killed*, by Michael Halliday (i.e. John Creasey). Large print edition. London: Society for Assistance to the Blind; 1938.

(3) *The man I killed*, by Michael Halliday (i.e. John Creasey). Audiotape, read by Sir Lawrence Olivier. New York: Books on Tape; 1966.

(4) *The man I killed*, play by Christopher Wern, based on the novel by Michael Halliday.

(5) *The man I killed*, a movie version of the play by Christopher Wern, based on the novel by Michael Halliday. On videotape.

(6) An individual copy of (1) as originally printed.

(7) An individual copy of (1), produced by making a copy of (6).

(8) An individual copy of (1), owned by Sir Lawrence Olivier, with many marginal notes in ink.

(9) A facsimile edition of (8), published by Marx Brothers.

(10) *The man I killed, completely revised and with a new ending* , by Michael Halliday (i.e. John Creasey). London: Marx Brothers; 1941.

(11) A legal document with original signatures

(12) A copy of the same

(13) A notarized copy of the same

Table II. Publishing History for *Guide to Reference Books* (adapted and updated by DS)

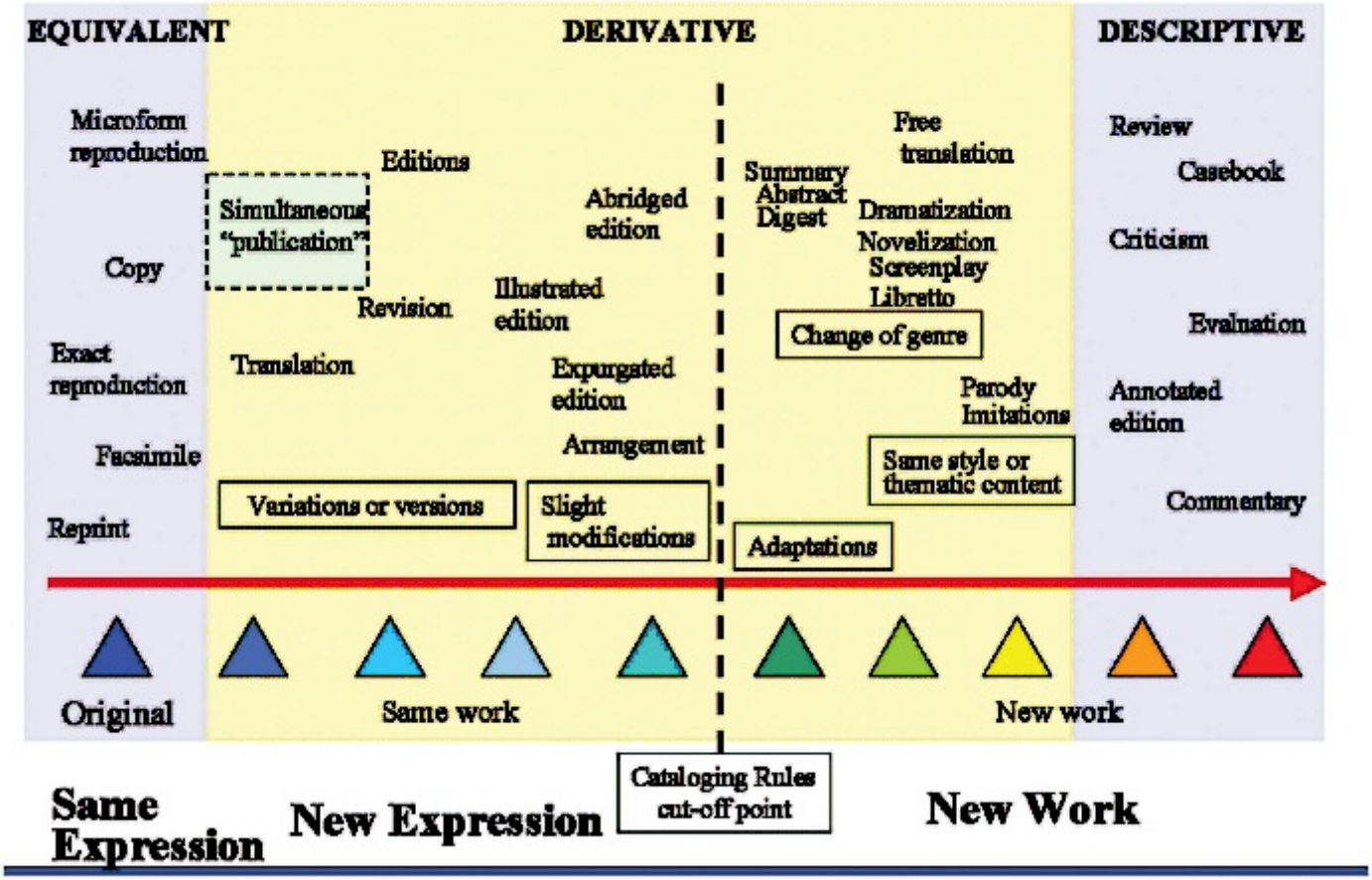
Edition	Date	Authors	Publisher	Title
	1902	Kroeger	Houghton, Mifflin & Company	Guide to the study and use of reference books; a manual for librarians, teachers, and students
Title edition	1904	Kroeger	American Library Association Publishing Board	Guide to the study and use of reference books; a manual for librarians, teachers, and students
2d ed., rev. and enl.	1908	Kroeger, Mudge	American Library Association	Guide to the study and use of' reference books
3d ed., rev. throughout and much enlarged	1917	Kroeger, Mudge	"	Guide to the study and use of reference books
[4th ed.]	1923	Mudge	"	New guide to reference books
5th ed.	1929	Mudge, Kroeger	"	Guide to reference books
6th ed.	1936	Mudge, Winchell	"	"
7th ed.	1951	Winchell, Mudge	"	"
8th ed.	1967	Winchell, Mudge, Sheehy	"	"
9th ed.	1976	Sheehy, Keckeissen, Mcllvaine, Winchell	"	"
10th ed.	1986	Sheehy (ed.	"	"
11th ed.	1996	Balay (ed.), Carrington, Martin	"	"
12th ed.	in prep.	Krieft	"	Guide to Reference Sources, GRS12 [Online; will also include Web sources]

Notes: This listing does not include supplements issued between editions. Up to edition 9, the person(s) associated with the work are listed thus: *compiled by*, from edition 10 onward *edited by*.

Definition of units in bibliographic and record control

<table>
<tr><th>Soergel draft
As the most inclusive term that is superordinate to all of the types defined here we will use document (in the broadest sense) or, even broader, resource.</th><th>FRBR
Functional Requirements for Bibliographic Records</th></tr>
<tr><td rowspan="2">Work
Intellectual or artistic entity, as the abstract essence or as a text, image, or piece of music.
Range:
A basic story or theme
the story of Faust
the myth of the Great Flood
A text telling the story, such as
Goethe's Faust
the account of the Great Flood in the Bible (original Hebrew)
A specific version of that text, a Latin version, the account of the same myth in another culture.</td><td>Work
A distinct intellectual or artistic creation.</td></tr>
<tr><td>Expression
The specific intellectual or artistic form that a work takes each time it is ‘realized'</td></tr>
<tr><td rowspan="2">Mannifestation (also called edition in one meaning of edition)
A specific expression or rendering of a work by means of a graphical image or sound, taken in the abstract; the idea of such an expression.
Examples:
The text of Goethe's Faust presented in a particular typeface and layout. (A performance at which the text is recited also renders the text but is a separate, but related, work.)
A specific score of a given version of Schubert's Fifth.
(A performance of that version of Schubert’s Fifth also renders the piece of music but is a separate, but related, work)
Also the expression or rendering of a work in the form of digital storage that can be transformed to a graphical image or sound, again taken as the abstract pattern of digital signals.</td><td>Manifestation
The physical embodiment of an expression of a work. As an entity, manifestation represents all the physical objects that bear the same characteristics, in respect to both intellectual content and physical form.</td></tr>
<tr><td>Printing (not in FRBR, but used)
A set of books printed at the same time or printed at different times containing no more than slight variations</td></tr>
<tr><td>Item (also individual copy or simply copy or physical copy)
The embodiment of a manifestation in a physical object.
We can perceive the content of a manifestation only through an individual item (copy) of it (unless we have memorized the contents of a manifestation and can conjure it up from memory). There are works that have only one manifestation of which there is only one item.</td><td>Item
A single exemplar of a manifestation. The entity defined as item is a concrete entity</td></tr>
</table>

FRBR Functional Requirements for Bibliographic Records	
Work	A work is an abstract entity; there is no single material object one can point to as the work. We recognize the work through individual realizations or expressions of the work, but the work itself exists only in the commonality of content between and among the various expressions of the work. When we speak of Homer's Iliad as a work, our point of reference is not a particular recitation or text of the work, but the intellectual creation that lies behind all the various expressions of the work.
Expression	An expression is the specific intellectual or artistic form that a work takes each time it is "realized." Expression encompasses, for example, the specific words, sentences, paragraphs, etc. that result from the realization of a work in the form of a text, or the particular sounds, phrasing, etc. resulting from the realization of a musical work. . . . excludes aspects of physical form, such as typeface and page layout, that are not integral to the intellectual or artistic realization of the work as such. When an expression is accompanied by augmentations, such as illustrations, notes, glosses, etc. that are not integral to the intellectual or artistic realization of the work, such augmentations are considered to be separate expressions of their own separate work(s). Such augmentations may, or may not, be considered significant enough to warrant distinct bibliographic identification. Inasmuch as the form of expression is an inherent characteristic of the expression, any change in form (e.g., from alpha-numeric notation to spoken word) results in a new expression. Similarly, changes in the intellectual conventions or instruments that are employed to express a work (e.g., translation from one language to another) result in the production of a new expression. If a text is revised or modified, the resulting expression is considered to be a new expression.
Manifestation	The physical embodiment of an expression of a work. . . . encompasses a wide range of materials, including manuscripts, books, periodicals, maps, posters, sound recordings, films, video recordings, CD-ROMs, multimedia kits, etc. . . . . represents all the physical objects that bear the same characteristics, in respect to both intellectual content and physical form. When a work is realized, the resulting expression of the work may be physically embodied on or in a medium such as paper, audio tape, video tape, canvas, plaster, etc. That physical embodiment constitutes a manifestation of the work. . . . Whether the scope of production is broad (e.g., in the case of publication, etc.) or limited (e.g., in the case of copies made for private study, etc.), the set of copies produced in each case constitutes a manifestation. All copies produced that form part of the same set are considered to be copies of the same manifestation. [Changes in physical form result in a new manifestation; examples:] changes affecting display characteristics (typeface, size of font, page layout, etc.), changes in physical medium (e.g., a change from paper to microfilm), changes in the container (e.g., a change from cassette to cartridge as the container for a tape).
Item	A single exemplar of a manifestation. . . . a concrete entity. It is in many instances a single physical object (e.g., a copy of a one-volume monograph, a single audio cassette, etc.). There are instances, however, where an item comprises more than one physical object (e.g., a monograph issued as two separately bound volumes, a recording issued on three separate compact discs, etc.). . . . variations may occur from one item to another, even when the items exemplify the same manifestation, where those variations are the result of actions external to the intent of the producer of the manifestation (e.g., damage occurring after the item was produced, binding performed by a library, etc.).



This diagram from Tillet 2004 illustrates the boundary between work and expression in FRBR

The relationships between the bibliographic entities (Group 1) in FRBR are

Work	<*isRealizedThrough*>	Expression	(1:N)
Expression	<*isEmbodiedIn*>	Manifestation	(1:N)
Manifestation	<*isExemplifiedBy*>	Item	(1:N)

(A work can have many expressions, but an expression is always of one work)

FRBR includes other entity types, namely

Group 2. person (an individual) and **corporate body** (an organization or group of individuals and/or organizations).

Person or Corporate Body <*creates*>	Work	(N:M)
Person or Corporate Body <*realizes*>	Expression	(N:M)
Person or Corporate Body <*produces*>	Manifestation	(N:M)
Person or Corporate Body <*owns*>	Item	(N:M)

Group 3. Entities that serve as the **subjects of works**. The group includes **concept** (an abstract notion or idea), **object** (a material thing), **event** (an action or occurrence), and **place** (a location).

Work	<*hasAsSubject*>	Concept, Object, Event,Place	(N:M)
Work	<*hasAsSubject*>	Work, Expression, Manifestation, Item, Person, Corporate Body	(N:M)

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.

How to design a better catalog system

The key to designing an efficient database structure for a catalog lies in analyzing and applying the relationships between bibliographic entities. The root cause for the complexity of many cataloging rules is the attempt to force data with very complex relationships into a simple-minded data structure.

Elements of a conceptual data schema for a database with data about documents.

<*isVersionOf*>, more specific <*isTranslationOf*>

<*isPartOf*>

These relationships may hold between works, between manifestations, or between items.

Work <*isRenderedIn*> Manifestation or (inverse) Manifestation <*isRenderingOf*> Work

FRBR chain: Work <*isRealizedThrough*> Expression,
Expression <*isEmbodiedIn*> Manifestation

Manifestation <*generatedFrom*> (Manifestation, RenderingProcess)

(Examples of rendering processes: different screen renderings from same HTML source text through different browsers; facsimile; Optical Character Recognition. The distinction between a rendering process and a reproduction process is fluid.)

Manifestation <*isInstantiatedIn*> Item or (inverse) Item <*isInstantiatedIn*> Manifestation

FRBR: Manifestation <*isExemplifiedBy*> Item

Item <*reproducedFrom*> (Item, reproduction process)

Some problems in a conceptual data schema for bibliographic and record control

Records, originals vs. copies

Permanent copy vs. fleeting copy

Specific printing may use different paper - preservation!

Performance of a work may be more than a mere emanifestation since it brings separate creative elements. Perhaps a performance should be considered a work of its own, with the tape (or audio file) on which it is captured being an item of a manifestation of that work (remastering such a tape would create another manfestation).

This ruling is a good example of the importance of discussing the problem of different versions of a document. Emphasis added.

Public Citizen v. John Carlin, Archivist of the United States Oct. 1997
Overturned by Court of Appeals for the District of Columbia Circuit August 1999

Washington Post, Thursday, October 23, 1997, p. A21

Judge Nullifies Rule on Computer Data

Archivist Criticized for Letting Agencies Eliminate Electronic Records

By George Lardner Jr.
Washington Post Staff Writer

A federal judge held yesterday that the head of the National Archives ignored his duties and acted illegally in issuing a regulation that authorizes all government agencies to wipe out their electronic mail and other computerized records regardless of content.

In a 36-page ruling sharply critical of Archivist John W. Carlin, U.S. District Judge Paul L. Friedman declared the controversial rule "null and void" and called the government's defense of it "irrational on its face."

The two-year-old regulation, known in bureaucratic jargon as "GRS [General Records Schedule]-20," permitted all agencies, from the Executive Office of the President on down, to destroy e-mail and wordprocessing records once they have been copied on paper or some other format and deemed "no longer needed for updating and revision."

Historians, researchers and journalists represented by the nonprofit advocacy group Public Citizen denounced the provision as an "electronic shredder" and filed suit, accusing Carlin of abdicating his responsibilities to appraise the value of the records on an agency-by-agency basis.

Friedman agreed. "Simply put," he said, "**electronic communications are rarely identical to their paper counterparts; they are records unique and distinct from printed versions of the same record.**"

Citing an example from the Iran-contra scandal, the judge pointed out that so-called PROF notes-computerized messages between national security adviser John M. Poindexter and White House aide Oliver L North-played an important role in the trials of both men.

"Admiral Poindexter, a computer expert set up a special channel known as "Private Blank Check," which allowed North and Poindexter to relay messages to each other without those messages being accessible to other NSC staff," noted Friedman, who was once an Iran-contra prosecutor. "The communication itself was clearly important to investigators, but the mode of communication and the special channel through which it was sent, which would not have been reflected in paper printouts of the messages, was also important."

In promulgating GRS-20 in 1995, the judge said, Carlin categorically determined that electronic records possess no administrative, legal, research or historical value beyond paper printouts of the same document. In doing this, "the Archivist has absolved both himself and the federal agencies he is supposed to oversee of their statutory duties to evaluate specific electronic records as to their value."

Carlin, the judge said, also exceeded his authority in giving agencies "carte blanche" to destroy electronic versions" whenever agency officials believe they are no longer needed.

The government had argued that GRS-20 was soundly based because such government-wide rules were meant for records of common form, such as "electronic" media.

Lawyers for Carlin had also protested that most federal agencies are not yet equipped to preserve records in electronic format. Friedman said this was " an important concern" but observed that "computers have now become a significant part of the way the federal government conducts its business" and the government must adapt to that reality.

The Archives had no immediate comment.

Definition of "catalog" - elaboration

A work is *covered by a catalog* if the catalog contains data about the work, or any manifestation of the work, or any item (individual copy) of any manifestation of the work. The collection linked with a catalog may be either a collection of items or merely a list of works, manifestations, or copies; a manifestation is said to be *represented in a collection* of items if any item of the manifestation is in the collection. Note that in the electronic age the concept of *collection* becomes more and more fluid. Is the whole World Wide Web one collection, or is a collection confined to the documents (files) stored on one Web server? Likewise, the concept of *library* becomes more and more fluid; there are now digital libraries whose "collections" may be distributed over many sites (whence the term *virtual library*). In fact, there is no sharp distinction between a digital library and any computerized information system. The functional distinctions made in Section 2.6 of the text are useful to clarify some of the issues here.

Objectives of the library catalog - restatement by D. Soergel

The catalog (of a library, a book seller, ...) should be an efficient instrument for ascertaining

(1) **Criterion search (intellectual access)** [4.1.2]
Which works, manifestations, or copies are helpful to a given user for a given purpose, to wit

(1a) which works, manifestations, or copies covered by the catalog meet a combination of criteria relating to provenance (including authorship), subject, artistic characteristics, and/or other criteria (**retrieval** or **identification**) (in some cases only certain manifestations or individual copies may meet the user's search criteria) (**find** a set of rsources) [4.1.2];

(1b) whether a work, manifestation, or item meets the needs of the user and how several suitable works, manifestations, or copies should be ranked (**selection**) [4.3];

(1c) how a work, manifestation, or item relates to (another) work, manifestation, or item (**relation**) (for example, <*is revision of*>, <*is reprint of*>, <*is based on*>);

(2) **Search for a known** work, manifestation, or item (**find** a single resource) [4.1.1]
(Confusingly, this is called *known-item search*, a term coined before FRBR)
To ascertain that a resource given in the catalog is the same as the resource in hand (**identity**) [4.2]

(2a) whether a known work is covered by the catalog and, if so, which are the manifestation(s) of the work that are covered in the catalog (**coverage**);

(2b) whether a known manifestation is covered in the catalog;

(2c) whether a known item (specific copy) is covered in the catalog (important for rare books)

(3) how the user can get **physical access** to some item (copy) of the work (method of access, time, cost) (**acquire** or **obtain)** [4.4].

The objectives are arranged by decreasing complexity and increasing concreteness, not by importance. The user who has achieved objective 1 must then achieve objective 2 and finally objective 3.

The Statement of International Cataloging Principles (a reading), Section 4, presents a somewhat different organization of objectives of the catalog; their numbers given in []

The Statement lists 4.5 to **navigate**, but navigation is a means for achieving any of the objectives, just as query-based search, so it does not belong in this list. Of course, a catalog should support both

Note: Many of these objectives apply to searching for people, organizations, or any other *resource* (in the sense used on the Web).

Lecture 7.1c (25 min) ***March 2***

Bibliographic and record control: Description

Describing texts and documents in a more general context

General principles of description; their application to bibliographic and record control; their implementation in ISBD/AACR2; relationship to the MARC format. User-oriented analysis of elements of description needed.

Description: What needs to be known about an entity?
Relates to catalog objective 1b, ascertaining relevance. Also relates to objective 3, ascertaining whether a given manifestation is indeed the same manifestation that is covered in a catalog record. (The "given manifestation" may be the manifestation requested by the user or the manifestation of which the item in hand is an instance.) Data about bibliographic entities - conceptual data schema. Peg each piece of data to the correct bibliographic level (work, manifestation, item).

Sources of information for cataloging data
Primary: Title page and verso (back of title page), Secondary: Preface, last page, cover page of a journal issue, etc. (data shown in []) Which is source is most authoritative?

Arrangement of cataloging data
• **in a record** - record format such as MARC • **in a display** (printed or on a screen) Many different styles: AACR2, ANSI Standard, Turabian, American Psychological Association Let a computer program, such as ProCite or Library Master, do the work!

Lecture 7.2a (40 min) ***March 2***

Bibliographic and record control 2: Entries

General principles of access; their application to bibliographic and record control; their implementation in AACR2 choice and form of entry. Authority files. User-oriented analysis of access points needed.

Definition	An *entry* is an element, such as an author name, a title, a series title, or a subject descriptor under which a document (or another object) can be found in a catalog or index. (The term comes from book or card catalogs, where an entry for a document is made by writing or filing a card.) Determining entries is a problem of data structure and access.
Two issues:	A Which of the data in the description should be made access points for lookup searching? (The answer to that question might have repercussions for description if a data element is important for access but not for ascertaining the relevance of an item.) B What form should each entry take? (Rules for entity values)
Main entry **Added entry**	A document may have many authors / contributors. Most of the time, this does not present a problem: just list them all and provide access from all ("make an entry," as in a card catalog, for all). But sometimes we want to list a document record only once: In a printed bibliography, in a listing of search results arranged by author, in a card catalog before reproduction equipment when every card had to be written or typed by hand. In that case, we need to select the most significant author / contributor, the one under which the one entry should be made. This is the main entry. (This concept was quite important in the age of card catalogs; it is less important now but still has applications.)
In-class exercise	**Problems of determining author entry analyzed according to Lubetzky** Lubetzky was the foremost thinker on bibliographic cataloging rules.
Advanced	Thinking about rules for corporate entry

The Author approach: Conditions and cases (Lubetzky after Needham)
See next page. Lubetzky's discussion of possible solutions is found in the readings.

Condition	Issue A: What entries to make	Issue B: What form the entry should take
1. Documents having more than one Author	1.1 Document prepared by an author with the aid of collaborators or contributors.	
	1.2 Document composed by an editor or compiler from the writings of several other people	
	1.3 Document by several authors with no one author more responsible for it than any of the others.	
	1.4 Document in which the writer reports the communication of another person (real or fictitious).	
2. Authors having more than one name		2.1 The author has changed his or her name in consequence of marriage, adoption of new citizenship, joining a religious order, or for any other reason.
		2.2 The author always writes under an assumed name different from his real name, or under his title of nobility, or under part of his name.
		2.3/4 Author uses more than one name in successive documents
		2.5 Authors whose names appear in translation in varying forms.
3. Dependent Documents		
4. Corporate authors	4.1 The reports and statements of a corporate body are usually prepared by one of its officers or by another person engaged to prepare the statement for it.	4.2 Many corporate bodies have no proper identifying names of their own but only generic names describing their type and common to most bodies of that type e.g. public library, historical society, dramatic club, etc.
		4.3 Change of name in corporate bodies.
		4.4 An organization may act or speak as a whole or through one of its branches, divisions, offices, etc.

Sample documents for analyzing author entry according to Lubetzky

(a) *The record guide* by Edward Sackville-West and Desmond Shaw-Taylor, with Andrew Porter and William Mann.

(b)* *Studies in the social psychology of adolescence*, by J. E. Richardson, J. F. Forrester, J. K. Shukla, and P. J. Higginbotham; edited with a foreword by C. M. Fleming.

(c) *The tropics*, by Edgar Aubert de la Rue, Francois Bourliere, Jean-Paul Harroy.

(d) *Ambit* (a periodical), edited by M. C. O. Bax and Edwin Brock.

(e)* *Chisholm's handbook of commercial geography*, entirely rewritten by L. Dudley Stamp and S. Carter Gilmour.

(f)* *Making magical apparatus*, by Jane Reid (i.e. Mrs. David Johnstone).

(g) *Lord Jim*, by Joseph Conrad (i.e. Josef Theodor Konrad Korzeniowski).

(h) *The far country*, by Neville Shute (i.e. Neville Shute Norway).

(i) *The trimmed lamp, and other stories*, by O. Henry (i.e. William Sydney Porter).

* designates items to be analyzed in Assignment 8.

(j) *The scene of the crime*, by John Creasey.

(k) *The man I killed*, by Michael Halliday (i.e. John Creasey). London: Marx Brothers; 1935

(l) *A branch for the baron*, by Anthony Morton (i.e. John Creasey).

(m)* *Schubert: thematic catalogue of all his works in chronological order*, by Otto Erich Deutsch in collaboration with Donald R. Wakeling.

(n)* *A concordance to the poems of William Wordsworth*, by Lane Cooper.

(o)* *The poetical works of Wordsworth*, edited by E. de Selincourt.

(p) *Oxford book of English verse, 1250-1918*, chosen and edited by Sir Arthur Quiller-Couch.

(q) *Shakespeare's 'Much ado about nothing'*, by N. T. Carrington (Notes on chosen English texts). Text and commentary.

(r) *The man I killed*, by Michael Halliday (i.e. John Creasey). Audiotape, read by Sir Lawrence Olivier. New York: Books on Tape; 1966.

(s) *The man I killed*, play by Christopher Wern, based on the novel by Michael Halliday.

(t) *The Aeneid of Virgil*, retold by N. B. Taylor.

(u) *Billy Budd* (libretto), adapted from the story by Herman Melville by E. M. Forster and Eric Crozier.

(v) *Iban agriculture: a report on the shifting cultivation of hill rice*, prepared for the Colonial Office by John Derek Freeman.

(w) *Essays and studies, 1962: being volume fifteen of the new series of Essays and studies*; collected for The English Association by Beatrice White.

(x) *The Library Association Record*, edited by Edward Dudley.

(y) Ministry of Education. 15-18. (The Crowther report).

(z) *Yearbook* of the Institution of Agricultural Engineers (originally the Institution of British Agricultural Engineers).

(aa) National Physical Laboratory. *Mathematical tables*. (The National Physical Laboratory is a branch of the Department of Trade and Industry.)

(bb) *Farm business statistics for south-east England*. Wye College (London University).

(cc) *Annual report* of the Association of Assistant Librarians (a division of the Library Association).

Problems of entry for works emanating from corporate bodies

What is a corporate body?

The **definition of a corporate body** according to AACR2:

"21.1B Entry under corporate body

21.1Bl. Definition. A corporate body is an organization or a group of persons that is identified by a particular name and that acts, or may act, as an entity. Consider a corporate body to have a name if the words referring to it are a specific appellation rather than a general description. If, in a script and language using capital letters for proper names, the initial letters of the words referring to a corporate body are consistently capitalized, and/or if, in a language using articles, the words are always associated with a definite article, consider the body to have a name. Typical examples of corporate bodies are associations, institutions, business firms, nonprofit enterprises, governments, government agencies, projects and programs, religious bodies, local churches, and conferences.

Note that some corporate bodies are subordinate to other bodies (e.g. the Peabody Museum of Natural History is subordinate to Yale University.)

Consider ad hoc events (such as athletic contests, exhibitions, expeditions, fairs, and festivals) and vessels (e.g., ships and spacecraft) to be corporate bodies."

Issue A. **When to make an entry under a corporate body** (corporate entry)?
(When to establish a relationship between a work and a corporate body?)
One question to be asked: When does a corporate body have a role similar to the responsibility of an author (corporate authorship)?

AACR2:

"21.1B2. Footnote 2. Consider a work to have emanated from a corporate body if it is issued by that body or has been caused to be issued by that body or if it originated with that body."

A corporate body can

- be fully responsible, as in a law enacted by a government or the official statement of an organization;
- have commissioned a work;
- have issued / published a work;
- provided the environment in which a work was created, such as a university or Rand providing an environment for a researcher (usually handled, if at all, as author affiliation).

Easiest solution: Make a good list of these relationship types and use the specific relationship when cataloging a work. This does not fit into the prevailing system of cataloging; there are just the MARC fields 110 and 710 for corporate names (corporate body in an author-like role). AACR2 gives rule under what circumstances such entries should be made.

Issue B. What form of name should be used for the corporate body?

See AACR2 Chapter 24

Problems:

- Corporate bodies change their names, for example
  Bureau of Foods became *Center for Food Safety and Applied Nutrition*.
- More complex problem: Corporate bodies change, cease to exist, or are merged with other corporate bodies.
- Corporate bodies are part of other corporate bodies, for example
  US. Department of Health and Human Services. Public Health Service. Food and Drug Administration. Center for Food Safety and Applied Nutrition. Technical Operations Branch.
  Some levels in such a hierarchy are often better known than others, for example,
  Food and Drug Administration is better known than
  Public Health Service
  (possibly a phenomenon similar to basic level concepts)
  There are other relationships, such as corporation B being a wholly-owned subsidiary of corporation A (interesting if you want to sue corporation B for damages).
- Corporate bodies often have generic names that are meaningful only in conjunction with a place, such as
  Metropolitan Museum of Art
  Some rules suggest to put the name of the place first:
  New York Metropolitan Museum of Art
  Cleveland Metropolitan Museum of Art
- Corporate bodies are often better known under an acronym or short popular name, for example
  FDA is better known than *Food and Drug Administration*

These problems occur in any information system that deals in any way with corporate bodies. The best solution is to have a database of corporate bodies including all their names, their life span, their relationships, and information on place. Such a database could be used in conjunction with a bibliographic catalog or other information systems.

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 This is a work 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

Lecture 7.2b (35 min) ***March 2***

Metadata, Resource Description Framework (RDF), Dublin Core (DC)

Resource Description Framework (RDF)

Definition	A general abstract data modeling method based on the entity-relationship approach, along the lines of what is described in Integrated Information Structure Interface. Developed with metadata in mind, but can be used for any kind of data for any kind of use. Entities / objects are called *resources* Relationship types are called *properties* A statement is made *about* a resource (the subject of the statement), using a property and giving another resource as the *value* Example: www.dsoergel.com/670 <*creator*> "Soergel"
Implementa-tion	Major implementation in XML using the Resource Description Framework (RDF) Model and Syntax Specification (www.w3.org/TR/REC-rdf-syntax/) The RDF syntax is specified using *XML Schema*. By convention, RDF syntax elements are identified by the prefix rdf:. RDF is merely a syntax; it does not specify any particular *properties* (relationship types). These properties can and must be defined by each user community.
Difference from XML	Enhancements of RDF syntax over plain XML: The meaning (the semantics) of syntactical constructs are precisely defined, while in XML they are deliberately left to the specific application. In particular, for each property the types of the participating resources can be specified. RDF has many additional features.
Example	See next page. More examples in discussion of Dublin Core
Name spaces in XML	RDF syntax as a name space Dublin core properties as a name space

Example of an RDF resource description: simple document description

Here is a simple example of RDF syntax used to describe a resource. This example uses properties defined in the Dublin Core (DC). The RDF syntax and the DC syntax are each identified in separate XML name spaces within this description (rdf: and dc:, respectively). Name spaces ensure that there is no collision between tag names in the two syntaxes (imagine what happens when two language syntaxes, each defined in XML, use the same tag name, like "description", to define different entities. Without the use of distinct name spaces, things would get horribly confusing, both to document creators and the automated systems parsing XML documents).

```
<?xml version="1.0"?>
<rdf:RDF
    xmlns:rdf="www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.0/">
    <rdf:description  rdf:about="www.ukoln.ac.uk/metadata/resources/dc/datamodel/WD-dc-rdf/">
            <dc:title> Guidance on expressing the Dublin Core within RDF </dc:title>
            <dc:creator> Eric Miller </dc:creator>
            <dc:creator> Paul Miller </dc:creator>
            <dc:creator> Dan Brickley </dc:creator>
            <dc:subject> Dublin Core; RDF; XML </dc:subject>
            <dc:publisher> Dublin Core Metadata Initiative </dc:publisher>
            <dc:contributor> Dublin Core Data Model Working Group </dc:contributor>
            <dc:date> 1999-07-01 </dc:date>
            <dc:format> text/html </dc:format>
            <dc:language> en </dc:language>
    </rdf:description>
</rdf:RDF>
```

Dublin Core (DC) (http://dublincore.org/)

The Dublin Core is a minimal standard for the description of "document-like objects"

- Title
- Creator
- Subject
- Description
- Publisher
- Contributor
- Date
- Type
- Format
- Identifier
- Source
- Language
- Relation
- Coverage
- Rights

The "plain" Dublin Core has just 15 properties (relationship types), simplicity both good and bad. The Dublin Core list of properties can be implemented in many ways, among them:

In the *meta* section of an HTML document (see next page and the model catalog)

In plain XML

In XML, using the RDF enhanced syntax

Here are a few lines of an HTML document with Dublin Core metadata

```
<HTML>
    <HEAD>
            <META name="dc.creator" content="Renato Iannella">
            <META name="dc.creator.affiliation" content="DSTC">
            <META name="dc.subject" content="Cats, Fur, Purr">

            Specification of the scheme from which subject descriptors were taken, here LCSH
            <META name="dc.subject" content="(scheme=LCSH) Animals~Felines">

            Language of title or subject descriptors can be specified in like manner
    </HEAD>
    <BODY>
            . . .

    </BODY>
</HTML>
```

More complex example illustrating more features of RDF and refinements of the Dublin Core.

```
<?xml version='1.0'?>
<rdf:RDF
    xmlns:rdf="www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc = "http://purl.org/dc/elements/1.0/"
    xmlns:dcq = "http://purl.org/dc/qualifiers/1.0/"
    xmlns:vcard = "www.imc.org/vcard/3.0/">

    <rdf:description  rdf:about="www.ukoln.ac.uk/metadata/resources/dc/datamodel/WD-dc-rdf/">
            . . .
            <dc:creator>
                    <rdf:Description>
                            <rdf:type rdf:resource =
                            "http://purl.org/dc/terms/1.0/creator/class/Person"/>
                            <dcq:creatorType rdf:resource =
                            "http://purl.org/dc/terms/1.0/creator/type/Editor"/>
                            <rdf:value rdf:resource = "http://411.com/EricMiller"/>
                    </rdf:Description>
            </dc:creator>
            . . .
    </rdf:description>

    <rdf:description rdf:about = "http://411.com/Eric Miller">
            <vcard:fn>Eric Miller </vcard:fn>
            <vcard:email> emiller@oclc.org </vcard:email>
            <vcard:org>OCLC.</vcard:org>
    </rdf:Description>
 </rdf:RDF>
```

Dublin Core elements

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.

RDF Schema	Before properties can be used in RDF descriptions, they must be defined, just as tags must be defined before they can be used in valid XML documents. For this purpose, RDF provides RDF Schema, which is turn specified using XML Schema. . Below is a very simple piece of an RDF schema definition for bibliographic data. Note the specification of hierarchical relationships among classes and hierarchical relationships among relationship types (properties) XML and RDF schemas for the food database from Lecture 2.2,is given in www.dsoergel.com/670/SYL2003FaLecturesAppendixNew.pdf (see Lecture 6.2b)

RDF schema definition example: Dublin Core

Definition of an entity-relationship conceptual schema in lengthy syntax

```
<?xml version='1.0'?>
 <rdf:RDF
   xmlns:rdf="www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:rdfs="www.w3.org/2000/01/rdf-schema#"
   xmlns:dc="http://purl.org/dc/elements/1.1/"/>
```

Definition of entity types (classes)

```
 <rdfs:Class rdf:ID="Book">
    <rdfs:label>Book</rdfs:label>
    <rdfs:comment>The class of books</rdfs:comment>
    <rdfs:subClassOf rdf:resource="www.w3.org/2000/01/rdf-schema#Resource"/>
 </rdfs:Class>

 <rdfs:Class rdf:ID="LegalEntity">
    <rdfs:label>Legal entity/rdfs:label>
    <rdfs:comment>The class of person or organizations</rdfs:comment>
    <rdfs:subClassOf rdf:resource="www.w3.org/2000/01/rdf-schema#Resource"/>
 </rdfs:Class>

 <rdfs:Class rdf:ID="Person">
    <rdfs:label>Person</rdfs:label>
    <rdfs:comment>The class of persons</rdfs:comment>
    <rdfs:subClassOf rdf:resource="#LegalEntity"/>
 </rdfs:Class>

 <rdfs:Class rdf:ID="Organization">
    <rdfs:label>Person</rdfs:label>
    <rdfs:comment>The class of organizations</rdfs:comment>
    <rdfs:subClassOf rdf:resource="LegalEntity"/>
 </rdfs:Class>
```

Definition of relationship types (properties)

```
  <rdf:Property rdf:ID="title">
```

```
    <rdfs:label>Title</rdfs:label>
    <rdfs:comment>The name given to the resource</rdfs:comment>
    <rdfs:domain rdf:resource="#Book"/>
    <rdfs:range
       rdf:resource="www.w3.org/2000/01/rdf-schema#Literal"/>
  </rdf:Property>

  <rdfs:Property rdf:ID="creator">
    <rdfs:label>Author</rdfs:label>
    <rdfs:comment>A person or organization responsible for the content of a book</rdfs:comment>
    <rdfs:subPropertyOf rdf:resource="http://purl.org/dc/elements/1.1/dcmes.rdf#Creator"/>
    <rdfs:domain rdf:resource="#Book"/>
    <rdfs:range
       rdf:resource="#LegalEntity"/>
  </rdfs:Property>

  <rdfs:Property rdf:ID="editor">
    <rdfs:label>Editor/rdfs:label>
    <rdfs:subPropertyOf rdf:resource="#creator"/>
  </rdfs:Property>

  </rdf:RDF>

  <rdfs:Property rdf:ID="affiliatedWith">
    <rdfs:label>person affiliation</rdfs:label>
    <rdfs:comment>The organization a person is affiliated with</rdfs:comment>
    <rdfs:domain rdf:resource="#Person"/>
    <rdfs:range
       rdf:resource="#Organization"/>
  </rdfs:Property>
</rdf:RDF>
```

Assume this schema is stored in a file with the URL

www.dsoergel.com/DSBibSchema.rdf

We can then refer to that file to use the definitions of entity types (classes) and relationship types (properties). See the following pages for some sample data.

Data on some books structured according to the rdf schema just given

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:bib="http://www.dsoergel.com/DSBibSchema">

/* Entity values assigned to entity types (Define class membership) */

<rdf:Description rdf:about="www.oclc.org/cat#ISBN0126542619">
     <rdf:type rdf:resource="http://www.dsoergel.com/DSBibSchema#Book"/>
</rdf:Description>

<rdf:Description rdf:ID="ISBN081086007">
     <rdf:type rdf:resource="http://www.dsoergel.com/DSBibSchema#Book"/>
</rdf:Description>

<rdf:Description rdf:about="www.oclc.org/cat#ISBN083893529X">
     <rdf:type rdf:resource="http://www.dsoergel.com/DSBibSchema#Book"/>
</rdf:Description>

<rdf:Description rdf:ID="Soergel">
     <rdf:type rdf:resource="http://www.dsoergel.com/DSBibSchema#Person"/>
</rdf:Description>

<rdf:Description rdf:about="www.simmons.edu/faculty#Chan">
     <rdf:type rdf:resource="http://www.dsoergel.com/DSBibSchema#Person"/>
</rdf:Description>

<rdf:Description rdf:ID="ALA">
      <rdf:type rdf:resource="http://www.dsoergel.com/DSBibSchema#Organization"/>
</rdf:Description>

/* Data on books (relating books to persons and organizations) */

<rdf:description  rdf:about="www.oclc.org/cat#ISBN0126542619">
         <bib:title> “Organizing information”</bib:title>
         <bib:creator>#Soergel</bib:creator>

<rdf:description  rdf:about="#ISBN081086007">
         <bib:title> “Cataloging and classification”</bib:title>
         <bib:creator>www.simmons.edu/faculty#Chan</bib:creator>

<rdf:description  rdf:about="www.oclc.org/cat#ISBN083893529X">
         <bib:title> “Anglo-American cataloging rules. 2. rev. ed.”</bib:title>
         <bib:creator>#ALA</bib:creator>

<rdf:Description rdf:about="www.oclc.org/cat#ISBN083893529X">
     <rdf:type rdf:resource="http://www.dsoergel.com/DSBibSchema#Book"/>
</rdf:Description>

<rdf:description  rdf:about="www.oclc.org/cat#ISBN083893529X">
```

```
        <bib:title> “Information retrieval”</bib:title>
**      <bib:creator>#Lancaster</bib:creator>
</rdf:Description>
</rdf:RDF>
```

** This line will be rejected since Lancaster has not been defined as a person

Midterm exam sample questions

There will be 5 questions. 15-20 minutes = number of points each, for a total of 90 minutes.

(20) 1. A congregation has a calendar database, which lists for each event (religious service, committee meeting, etc.) the title, time and place, and main participants (celebrant, speaker, etc.). They also have a separate calendar with the dates and names of holy days. They need to produce the following documents.

1a A congregational bulletin that lists the week's events.

1b A Web page that is meant to attract prospective members and that gives the week's events.

2 For each service a leaflet with the program (date and name of holy day, scripture passages, hymns, main participants).

Sketch a document template for document 1a or 1b and for document 2.

(15) 2. Describe the process of developing a conceptual data schema for an employment service (an information system on jobs/positions and on people seeking jobs with the purpose of finding matches between jobs and people). Illustrate your discussion with examples, including some of the key entity types and relationship types.

(15) 3. Assume you have

- a database of recipes that gives ingredients (basic foods) and their amounts for many prepared dishes, and
- a nutrient database that gives for each basic food the amount of each nutrients (proteins, carbohydrates, fats, vitamins and minerals) it contains.

You want to find all prepared dishes that contain both vitamin A and vitamin D. Describe a process that could be followed by an automated retrieval system to accomplish such a search.

(20) 4. There will be a question on restructuring a semantic network with application to bibliographic records.

(20) 5. This question has to do with possible improvements in retrieval through linguistic techniques. Below is an example consisting of a query and some short passages of text. Assume a straightforward free-text search system that searches for **words**; all passages that contain all query words joined by AND are retrieved. As a refinement, the system provides the proximity operator **ws**, which means **within the same sentence**. Thus, the query formulation *forest* **ws** *fire* retrieves all passages in which the two words occur in the same sentence; such passages are considered **retrieved**. See the instructions on the next page.

In the table of passages below, check all passages that are relevant to the user's need as expressed in the query. Then check all passages that are retrieved by the query formulation. Based on these checks, fill in the 3x3 grid and compute recall and discrimination. What refinements could be used to improve retrieval performance? For each refinement, list the passages whose retrieval is affected (give passage numbers). Which of these refinements improve recall, which improve discrimination, and which improve both?

Query topic: Forest fires

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

Passage	Relevant	Retrieved
30 Forest fires in Indonesia cause serious air pollution in South East Asia.		
31 The fire in Yellowstone Park destroyed 25% of the forest.		
32 The fire station is located behind the city forest.		
33 With fire in her eyes she chased him through the forest.		
34 The soldiers opened fire into the forest.		
35 The fire went out of control. It reached the forest and destroyed many acres.		
36 The animal got scared by the fire burning in the field. It ran into the forest.		
37 He asked whether he should fire the forest workers.		
38 Many square miles of forest in the West are burning.		
39 The dry wooded area went up in flames.		

	Relevant	Not relevant	All
Retrieved			
Not retrieved			
All			

March 2, 2011 Name (optional)

Free-write 7

Lecture 7.1a General introduction to metadata
Lecture 7.1b. Bibliographic and record control. General issues
Lecture 7.1c. Bibliographic and record control. Description
Lecture 7.2a. Bibliographic and record control. Entries
Lecture 7.2b. Metadata, RDF, Dublin Core (DC)

- **Reflect** – what you learned, what was most important, what was most interesting, what was extraneous;
- **Ask questions** – ask for more explanation, how is a concept connected to other concepts, why is a concept important, how can it be applied, why is a reading important;
- Offer **critique and suggestions**;
- Say anything else you want to.

Over

Part 4 **March 8 - April 27**

Part 4. Classification and subject access

Objectives	1 Understand the problems of vocabulary control. Be able to apply vocabulary control principles to indexing and searching. 2 Be able to extend vocabulary control principles to entity types other than Subject, such as names of organizations. 3 Understand the pervasive role of classification throughout the human endeavor. 4 Understand the functions of classification in information retrieval systems, especially request-oriented indexing and inclusive searching. 5 Understand the principles of the structure of subject classification, in particular facet organization and hierarchy, and be able to apply these principles to the analysis of existing schemes and to indexing and query formulation. 6 Be able to discern the facet structure of a domain. (There are facets everywhere.) 7 Be aware of the variety of classification schemes and other subject access vocabularies and have an acquaintance with the major subject access vocabularies used in American libraries, namely the Dewey Decimal Classification, Library of Congress Classification, and Library of Congress Subject Headings. 8 Be able to extend classification principles entity types other than Subject, for example to a hierarchy of organizations and organizational units.

over

Practical significance	• The practical significance of vocabulary control in indexing and, more importantly, in free-text searching is detailed in Lecture 8.1. • The multiple and pervasive uses of classification have been detailed in the reading for Lecture 8.1. Also remember Lectures 2.1-2.2, The nature of knowledge and knowledge representation. • For IR systems specifically: The index language – the set of subject descriptors used in an IR system and their interrelationships – underlies all activities in subject retrieval. Understanding index language functions and structure - facet structure and hierarchy - is, therefore, at the heart of understanding IR systems.
Cross-reference	Lectures 2.1 and 2.2. The nature of knowledge and knowledge representation

Lecture 8.1 / Small Groups 1 **March 8 - March 10**

Explorations in subject access (based on Assignment 10) (120 min)

Objectives	1 Through your own analysis and discussion, you should arrive at an appreciation of the complexities of subject access and identify the major problems. You do this by working through realistic examples. 2 This practical experience and problem awareness form the basis for the treatment of solutions at a more theoretical level in lectures and readings. Note: We have not yet discussed nor have you read about what solutions might exist for these problems. **The whole point of this small group exercise is for you to think on your own and figure out solutions to subject access problems yourself**. In the remainder of the semester, we will address each problem in turn.
Tasks	There are 5 tasks, all exploring problems of the index language in an ISAR system. Tasks a - c deal with a sample collection of transportation documents constructed for this assignment. Tasks d and e deal with the Library of Congress Subject Headings (LCSH) and the Library of Congress Classification (LCC), respectively. Task a Formulate queries to search the rough alphabetical index Task b Design a good alphabetical index Task c Establish an index language for a computerized IR system (no more than 100 descriptors) Task d From the Library of Congress Subject Headings (LCSH) extract a list of headings dealing with Transportation and Traffic. Task e From the Library of Congress Classification (LCC) extract a list of classes dealing with Transportation and Traffic.
Materials (attached)	• Rough alphabetical subject index to the sample collection • Excerpts from the Library of Congress Subject Headings (LCSH) • Excerpts from the Library of Congress Classification (LCC) Note: LCSH and LCC are two different systems for subject access. The Library of Congress Subject Headings are an alphabetical list of descriptors (subject headings) used in online searching or to arrange a card catalog. The Library of Congress Classification is a systematic arrangement of descriptors for all areas of knowledge; it is used to arrange books on shelves or links on a subject directory web page. At the Library of Congress, one cataloger assigns one or more LC subject headings to a book, another the one LC class that best fits.

In this small groups exercise emphasis is on identifying the problem and possible solutions. There is not enough time to actually implement a solution. We will do just enough practical work to understand the problem.

Tasks a-c: **Dealing with the sample collection**

Tasks a-d deal with subject access problems in a sample collection created by other students assigning subject descriptors to about 200 documents (six for each student) without any guidance whatsoever. The rough alphabetical index is simply an alphabetized list of the terms assigned in this "instructionless" indexing, with document numbers following each term. (Many terms have been culled to make the index shorter.)

Task a: **Identify the terms that should be used in the rough alphabetical index to search for the following queries** (no more than 30 terms for each query): (15 min.)

1. Harbors for large tankers.
2. Air cushion craft.
3. The consequences of the development of new types of vehicles for terminal design.
4. Simulation of passenger flow over transportation networks

Task b: **Design a good alphabetical index** to the sample collection (20 min). Your experience with Task b revealed problems with the rough alphabetical index. How would you design an alphabetical index that would address these problems and make searching easier? Start making some revisions to the index, just enough to get some experience that enables you to address the following points:

- What should a good alphabetical index look like?
- How would you go about transforming the rough index into a good index?

Task c: Establish an **index language** for a computerized ISAR system in the field of transportation (35 min). Assume that the list of terms in the rough alphabetical index is representative of the topics to be searched. So that users can remember all descriptors, the index language is limited to 100 descriptors, yet must allow searching for most of the concepts expressed in the rough alphabetical index without loss of specificity. Remember from Chapter 11 how a computerized ISAR system can be searched (remember its manipulative power.) See whether you can come up with an idea on the nature of such an index language and apply it to a few examples.

Result of the discussion: A sketch of what the index language might look like

Tasks d-e: **Dealing with existing systems used in American libraries.**

The Library of Congress, and many libraries in the US and around the world (esp. academic and research libraries), use **two separate schemes for subject access**.

1 The first is an alphabetical list of subject terms, the **Library of Congress Subject Headings (LCSH)**. These are used to index books, originally for an alphabetical subject catalog on cards and now for search in an **O**nline **P**ublic **A**ccess **C**atalog (OPAC). Usually several subject headings are assigned to a book to provide multiple access points. There are about half a million subject headings, listed in four large red volumes (to be found in Baldy 14A, older editions are fine)

2 The second is a systematically arranged scheme of subject classes, the **Library of Congress Classification (LCC)**; each class is identified by a class number marking its place in the classified arrangement, for example, BJ 2139 *Etiquette for airplane travel*. LCC is used for the systematic subject arrangements of books on the shelves. Since books are customarily shelved in only one place (even if there are multiple copies), only one LCC class is assigned to a book, providing only one access point. For example, consider a book titled *The history of State Street in Boston, 1870-1930*. This book could be classed under F73.5 *History of Boston > 1865-1950* or under F73.67.P3 *History of Boston > Streets. Bridges. Railroads > Park Street*, but the cataloger has to choose one of these. The 400,000 LCC classes are listed in 30 volumes of *classification schedules* (McKeldin Library Government Documents SU Docs LC 26.9 and McKeldin Reference Z696.U5)

The task is assembling a list of headings from LCSH and a list of class numbers from LCC that deal with transportation and traffic. These lists can serve the following purposes:

- A list of subject headings or a classification, respectively, for a transportation library.
- A query formulation to regularly search the OCLC WorldCat database for new records in the area of transportation as the basis for book selection in a transportation library. All elements of your list would be connected by OR and the resulting query formulation would retrieve all items on transportation either based on the subject headings or based on the class numbers assigned.
- An aid to a user who wants to search a general catalog for transportation topics. Such a user can find the appropriate subject heading(s) or class numbers much more quickly in a selected list than in the full LCSH or full LCC.

The task is the same as that of any user approaching the subject catalog or the shelves with a question, only magnified by the breadth of your topic.

You can get a feel for both schemes from the sample pages in the assignment materials, but you should also look at the actual schemes. Look at the LCC Outline (a thin separate volume).

Task d: Think about what you would do to put together a list of LCSH headings dealing with **Transportation and traffic**. (25 min) (Start with *Transportation*; you might also try *Ships*, *Railroads*, and other broad terms. Explore from there.)

A list of 15 relevant subject headings, at least 5 of which are not simply taken from the cross-references listed under a very broad heading such as *Transportation*, *Ships*, *Railroads*, or *Air transport.* (Cross-references are the Broader Term (BT), Narrower Term (NT), and Related Term (RT) cross-references given in LCSH, as well as the USE and UF cross-references, which are not of interest here.) You may not include subject headings that start with either *transportation* or *traffic*.

Task e: Think about what you would do to put together a list of classes dealing with **Transportation and traffic** (25 min). Can you restrict your efforts to one or two volumes (After all, a classification is supposed to bring all related subjects together)? Why not?

In a given branch of the hierarchy always list the broadest class that falls under transportation, for example, do not list *TF840-851 Technology > Railroads > Elevated railways and subways*, but go up to the broader level still included in transportation, *TF?? Technology > Railroads*.

Rough alpha index

based on "instructionless" indexing of a sample collection on transportation by students in previous classes

For some terms, the document number got lost but the terms are still important

Airport access

Commercial aircraft survey

Commercial airliners

Freight ferry

Freight stations

International passenger transport

International railroads

Pipelines

Planning

Planning 15
Planning -rural transportation 165
Port accommodations and facilities 44
Port authorities 90
Port dues and charges 44
Port exports and imports 44
Port remodeling 131
Ports 344, 349, 90
Ports - accommodations and facilities 344
Ports - directories 344
Ports - dues and levies 344
Ports - exports 344
Ports - imports 344
Potential 76, 80
Private train equipment 320
Prototype-trains 379
Rail installations 41
Rail transit 130
Rail transport 88, 206
Rail transportation 374, 383, 386, 387
Railroad car design 64
Railroad construction 41
Railroad freight 111
Railroad freight equipment 122
Railroad industry 149
Railroad market guide 206
Railroad personnel 206
Railroad signals 41
Railroad suppliers 149
Railroad travel 86
Railroads 64, 149, 156, 166, 184, 206, 341
Railroads - construction 341
Railroads - traffic control 341
Railroads-timetables, maps, schedules 74
Railway crossings 97
Railway directories 149
Railway engineering 34, 62, 362
Railway finance 34
Railway finances 334
Railway locomotives 324
Railway operating results 334
Railway schedules and fares 374
Railway statistics 34, 149
Railway, underground--design 50

Seadrome design

Railway, underground--Stockholm 50
Railways 166, 334
Railways-electric 149
Railways, electric 99
Railways, high speed 97
Ramp entry systems 171
Rapid rail transit 129
Rapid tramway-Europe 213
Rapid transit 56, 86, 87, 130, 171, 327, 356
Rapid transit construction system 130
Rapid transit systems 99, 386, 387
Rapid transit--design and construction 50
Recent developments 76
Regional highway planning 123
Regulations 12
Research 132, 133
Rivers 14
Road capacity 177
Road design 159
Road planning 348
Road planning--condition rating factors 48
Road planning--rural 48
Road planning--service rating factors 48
Road ratings 348
Road traffic surveys 135
Road transportation 370
Road way functions 351
Roads 158
Rolling stock 126, 41
Rotary wing aircraft 118
Rotary-wing aircraft 116
Route networks 96
Runway design 136
Rural arterial roads 348
Rural highways-planning 16
Rural transportation-planning 165
Safety 91, 112, 326
Satellite 133
School transportation 151
Sea carriers 205
Sea freight 205
Sea transport 205
Sea transportation 32, 371, 374
Seadrome design 113

Shared facilities

Shared facilities 174
Ship builder 77
Ship building 161,336
Ship data 36
Ship engineering 161
Ship information 380
Ship models 36
Ship names 77
Ship names list 71
Ship owner 77
Ship owners 32
Shipbuilders 71
Shipbuilding 157
Shipping 12,178
Shipping lines in the United States 32
Shipping register 380
Shipping schedules 32
Shipping-Great Lakes 198
Ships 90, 104
Ships-Canada 212
Ships-motor 210
Ships-U.S. 212
Ships--Merchant 36
Short haul air transportation 316
Short haul airliner 193
Short take off and landing 325
Short-haul traffic explosion 216
Short-haul utility aircraft design 216
Short- range transportation 58, 62
Size limits 11
Slow trains 140
Small commercial aircraft 134
Southern Inland Waterway 355
Space vehicles 188
Specialized carriers 218
Specialized trucking 202
Speed 326
Speed problems 79
Speed-volume-density-relationship 179
SST Flight Simulations 194
SST proposals 347
State-of-the-art 62
Statistics 141
Statistics for passenger liners 371

Traffic problems

Steam ships 192
Steamship schedules and fares 374
Steamships - timetables, maps; schedules 74
STOL 121, 186
Streets and highways 335
Subsonic aeronautics 109
Subsonic aircraft developments 216
Subways 129, 327, 356
Super express train 346
Super jet 170
Super sonic transport program 47
Supersonic aircraft 40, 194, 340
Supersonic transport 180, 194
Surface effects ships 12
Swing-way 40
Swing-wing aircraft 340
System analysis 132, 135
Systems analysis 113, 128
S.S.T. 136
Tank truck carriers 53
Tanker 131
Tanker owners 368
Tanker tonnage 368
Tankers 205, 368
Tankers; Sea transport 68
Technical data 188
Technological developments 383, 384
Terminal areas 215, 219
Terminal design 170, 29
Terminal facilities 203
Tilt wing 384
Timetable networks 140
Titanium alloy 347
Tokyo 101
Traffic 105, 141, 369
Traffic aids 179
Traffic aids 151, 153, 177, 186
Traffic control 56, 57, 69, 130, 151, 171
Traffic dynamics 179, 351
Traffic flow theory 369
Traffic lane detectors 1 71
Traffic overload 124
Traffic patterns 33
Traffic problems 153

Traffic projections

Traffic projections 196
Traffic service quality 351
Traffic variables 179
Traffic-intercity 196
Train 79
Train design 362
Train operation 41
Trains 62, 140
Transit systems 177
Transport aircraft 175
Transport capacity 177
Transport safety 83
Transport systems 28
Transport, individual 21
Transport, inter-city 21
Transportation 58, 60, 62, 113, 142, 143, 144, 145, 146, 147, 157, 166, 357
Transportation - innovation 345
Transportation engineering 351
Transportation lines 14, 19, 319, 322
Transportation network 189, 316
Transportation networks 115, 16
Transportation operations 322
Transportation research, government 21
Transportation research, private 21
Transportation Systems - U.S. 343
Transportation-air 211
Transportation-highways 208
Transportation-rapid tramway-Europe 213
Transportation-urban 213
Transportation-water 210, 210
Travel (mode of)--passenger preference 46
Trolley cars 327
Truck freight 111,187
Trucking 66, 202
Trucks 105, 153, 202
Tube vehicles 89
Turbine powered trains 364
Turbine-powered train 10
Turbo trains 142
Turbojet type 80
Turbo-train propulsion 191
Turbo-train structures 191
Turnpikes 141

Wingless aircraft

Underground railway 350
United States
United States-Freight stations 164
Unloading facilities 323
Urban freeway routes 185
Urban freeways 171
Urban highway planning program 123
Urban parking demand 207
Urban parking facilities 207
Urban rail systems 345
Urban renewal 185
Urban traffic engineering 207
Urban traffic flow 207
Urban transit, California 99
Urban transportation 15, 39
Urban transportation systems 195
Urban planning 32
U.S. ports 217
U.S. 13, 59
U.S. Air Service 365
U.S. Air Transport 365
U.S. Rapid Transit 129
Vehicle characteristics 381
Vehicle design 43
Vehicle detectors 179
Vehicle operating costs 381
Vehicle performance model 381
Vehicular parking spaces 321
Vertical-lift aircraft 181
Vessels 198, 322
VSTOL 186
VSTOL jet 316
VSTOL metroport 189
VSTOL 118
V-STOL 80
V-STOL aircraft 17
V/STOL 118, 384, 58
V/STOL aircraft 358, 377
Water transport 82, 205
Water transportation 95
Waterway guide 61
Waterway routes 32
Wind tunnel 379
Wingless aircraft 163

Pages from LCSH

Library of Congress Classification

Broad Outline (Main classes)

A General works

B Philosophy. Psychology. Religion

C Auxiliary sciences of history
D History: General and outside the Americas
E-F History of America
  E History: America General and United States General
  F History: United States local, Canada, and Latin America

G Geography

H Social sciences
J Political science
K Law

L Education

M Music and books on music
N Fine arts
P Language and literature

Q Science
R Medicine
S Agriculture
T Technology

U Military science
V Naval science

Z Bibliography and library science

The following pages give first a detailed outline and then examples of classes dealing with or relevant to *transportation and traffic*.

Library of Congress Classification. Detailed Outline

A General works

- AC Collections. Series. Collected works
- AE Encyclopedias (General)
- AG Dictionaries and other General reference works
- AI Indexes (General)
- AM Museums (General). Collectors and collecting (General)
- AN Newspapers
- AP Periodicals
- AS Academies and learned societies (General)
- AY Yearbooks. Almanacs. Directories
- AZ History of scholarship and learning. The humanities

B Philosophy. Psychology. Religion

B-BJ Philosophy, incl. BF Psychology

- B Philosophy (General)
- BC Logic
- BD Speculative philosophy
- BF Psychology. Parapsychology. Occult sciences
- BH Aesthetics
- BJ Ethics. Social usages. Etiquette

BL-BX Religion

- BL Religions. Mythology. Rationalism
- BM Judaism
- BP Islam. Bahaism. Theosophy
- BQ Buddhism

BR-BX Christianity

- BR Christianity
- BS The Bible
- BT Doctrinal theology
- BV Practical theology
- BX Christian denominations

C Auxiliary sciences of history

- C Auxiliary sciences of history (General)
- CB History of civilization
- CC Archaeology (General)
- CD Diplomatics. Archives. Seals
- CE Technical chronology. Calendar
- CJ Numismatics
- CN Inscriptions. Epigraphy
- CR Heraldry
- CS Genealogy
- CT Biography [General]

D History: General and Old World

- D History (General). Europe (General)
- DA Great Britain
- DAW Central Europe
- DB Austria, Hungary, Czech Republic, Slovakia
- DC France
- DD Germany
- DE Mediterranean region. Greco-Roman World
- DF Greece
- DG Italy
- DH Netherlands (low Countries). Belgium, Luxemburg
- DJ Netherlands (Holland)
- DJK Eastern Europe
- DK Russia and former Soviet republics. Poland
- DL Northern Europe. Scandinavia
- DP Spain. Portugal
- DQ Switzerland
- DR Balkan peninsula
- DS Asia
- DT Africa
- DU Oceania (South Seas) [Australia. New Zealand]
- DX Roma (Gypsies)

E-F History of America

- E1-143 America (General)
- E151-857 United States (Gen.)
- F1-957 United States: States and Local
- F1001-1140 Canada
- F1201- Other individual countries [mostly Latin America]

G Geography

G - GF Geography

- G Geography (General). Atlases. Maps
- GA Mathematical geography. Cartography
- GB Physical geography
- GC Oceanography
- GE Environmental sciences
- GF Human ecology. Anthropogeography

HB-HJ Economics

- GN Anthropology
- GR Folklore
- GT Manners and customs (General)
- GV Recreation. Leisure

H Social sciences

- H Social sciences (General)
- HA Statistics
- HB Economic theory. Demography
- HC- Economic history and
- HD conditions
- HE Transportation and communication
- HF Commerce
- HG Finance
- HJ Public finance

HM-HX Sociology

- HM Sociology (General and theoretical)
- HN Social history. Social problems. Social reform
- HQ The family. Marriage. Woman
- HS Societies: secret, benevolent, etc. Clubs
- HT Communities. Classes. Races
- HV Social pathology. Social and public welfare. Criminology
- HX Socialism. Communism. Anarchism

J Political science

- J General legislative and executive papers

JA-JC Political science

- JA Collections and general works
- JC Political theory
- JF-JQ Political institutions and public administration
- JF General works. Comparative works
- JK United States
- JL Brit. America. Latin America
- JN Europe
- JQ Asia. Africa. Australia. Oceania
- JS Local government
- JV Colonies and colonization. Emigration and Immigration
- JX International law. International relations

No longer used at LC

K Law

- K Law (General)
- KD United Kingdom and Ireland
- KDZ America. North America
- KE Canada
- KF United States
- KG Central America, Caribbean
- KH South America
- KJ-KK Europe

L Education

L Education (General)
LA History of education
LB Theory and practice of educ.
LC Special aspects of education

LD-LG Individual institutions
LD United States
LE America except United States
LF Europe
LG Asia, Africa, Oceania

LH College and school magazines and papers
LJ Student fraternities and societies in the United States
LT [Multi-subject] Textbooks

M Music and books on music

M Music [instrumental and vocal]
ML Literature of music
MT Musical instruction and study

N Fine arts

N Visual arts (General)
NA Architecture
NB Sculpture
NC Drawing. Design. Illustration
ND Painting
NE Print media
NK Decorative arts. Applied arts. Decoration and ornament
NX Arts in General

P Language and literature

P Philology and linguistics (Gen.)
PA Classical languages and lit.

PB-PH Modern European lang.
PB Celtic languages and literature
PC Romance languages

PD-PF Germanic languages
PD Scandinavian. North Germanic
PE English
PF West Germanic
PG Slavic. Baltic. Albanian languages and literature
PH Finno-Ugrian. Basque l & l

PJ-PL Oriental languages & lit.
PJ Oriental. Semitic
PK Indo-Iranian
PL Languages and literatures of Eastern Asia, African, Oceania
PM Hyperborean, Indian, and artificial languages

PN-PZ Literature
PN Literary history and collections
PQ Romance literature
PR English literature
PS American literature
PT Germanic literature
PZ Children's literature

Q Science

Q Science (General)
QA Mathematics. [Computer science]
QB Astronomy
QC Physics
QD Chemistry
QE Geology

QH-QR Biology
QH Natural history (General). Biology (General)
QK Botany
QL Zoology
QM Human anatomy
QP Physiology
QR Microbiology

R Medicine

R-RL Medicine
R Medicine (General)
RA Public aspects of medicine
RB Pathology
RC Internal medicine. Practice of Medicine
RD Surgery
RE Ophthalmology
RF Otorhinolaryngology
RG Gynecology and obstetrics
RJ Pediatrics
RK Dentistry
RL Dermatology

RM-RZ Allied disciplines
RM Therapeutics. Pharmacology
RS Pharmacy and materia medica
RT Nursing
RV Botanic, Thomsonian, and eclectic medicine
RX Homeopathy
RZ Other systems of medicine [Chiropractic. Osteopathy. Mental healing]

S Agriculture

S Agriculture (General)
SB Plant culture
SD Forestry
SF Animal culture
SH Aquaculture. Fisheries. Angling
SK Hunting

T Technology

T Technology (General)

TA-TH General engineering and civil engineering
TA General
TC Hydraulic and ocean eng.
TD Environmental technology, sanitary engineering
TE Highway engineering
TF Railroads
TG Bridge engineering
TH Buildings

TJ-TL Mechanical group
TJ Mechanical engineering
TK Electrical engineering. Nuclear engineering
TL Motor vehicles. Aeronautics. Astronautics

TN-TR Chemical group
TN Mining, metallurgy
TP Chemical technology
TR Photography

TS-TX Composite group
TS Manufactures
TT Arts and crafts. Handicrafts
TX Home economics

U Military science

U Military science (General)
UA Armies: Organization, description, facilities, etc
UB Military administration
UC Maintenance and transportation
UD Infantry
UE Cavalry, armor
UF Artillery
UG Military engineering. Air forces. Air warfare
UH Other services

V Naval science

V Naval science (General)
VA Navies: Org., descr., fac., etc
VB Naval administration
VC Naval maintenance
VD Naval seamen
VE Marines
VF Naval ordnance
VG Minor services of navies
VK Navigation. Merchant marine
VM Naval engineering. Ship-building. Marine Engineering

Z Bibliography and library science.

Z4-115 Books (General). Writing. Paleography
Z116-659 Book industry & trade
Z662-1000 Libraries. [Library science. Information science]
Z1001-8999 Bibliography

Library of Congress Classification
Sample classes dealing with or relevant to *transportation*

The following list gives a sampling of LC classes dealing with or relevant for *transportation and traffic*. The example classes are in *italics*. For each, the hierarchical chain leading to it is given to provide a sense of context, but neighboring classes are shown only in a few cases for illustration. Some full pages from the classification are also included with examples underlined (unless the entire section is about transportation). The examples have been chosen to make it easy for you to detect patterns on your own.

B Philosophy. Psychology. Religion

BJ Ethics
BJ1801-2195 . Social usages. Etiquette
BJ 2137 . . *Etiquette of travel*
BJ2139-2156 . . . *Special topics*
BJ2139 . . . . *Airplane travel*
BJ2140 . . . . *Bus travel*

BS The Bible
BS1-680 . General (Whole Bible)
BS410-680 . . Works about the bible
BS620-672 . . . Auxiliary topics
BS647-649 . . . . Prophecy
BS649 . . . . . Prophecy of special future events, A-Z
BS649.S8 . . . . . *Steam engines*

BV Practical theology
BV5-530 . Worship (public and private)
BV205-287 . . Prayer
BV229-283 . . . Prayers
BV283 . . . . Other special prayers, A-Z
BV283.A4 . . . . . *Air pilots' prayers*
BV283.T7 . . . . . *.Traveller's prayers*
BV590-1652 . Ecclesiastical theology
BV900-1450 . . Religious societies, associations, etc.
BV950-1280 . . . Religious societies of men, brotherhoods, etc.
BV955-1280 . . . . By period
BV960-1280 . . . . . 19th-20th centuries
BV1000-1220 . . . . . Young Men's Christian Associations
BV1160-1220 . . . . . Work with special classes
BV1175 . . . . . *Commercial travellers*
BV1200 . . . . . *Railroad employees*

BV2002-3705 . Missions
BV2610-2695 . . Special types of missions
BV2660-2695 . . . Work among special classes, by occupation
BV2695 . . . . Other classes, A-Z
BV2695.R3 . . . . . *Railroad men*
BV4000-4470 . Pastoral theology
BV4200-4317 . . Preaching. Homiletics
BV4239-4317 . . . Sermons
BV4309-4316 . . . . Sermons and talks to special classes of persons
BV4316 . . . . . Other classes, A-Z.
BV4316.R3 . . . . . *Railroad men*
BV4316.S3 . . . . . *Sailors and seamen*
BV4400-4470 . . Practical church work. Social work. Work of the layman
BV4435-4470 . . . Church work with special classes
BV4457-4459 . . . . Soldiers and sailors
BV4458 . . . . . *Sailors and seamen*
BV4485-5099 . Practical religion. The Christian life
BV4527-4596 . . Religious works for special classes
BV4588-4591 . . . Soldiers. Soldiers and sailors
4590-4591 . . . . *Sailors and seamen*
4596 . . . Other, A-Z
4596.R3 . . . . *Railroad men*

C Auxiliary sciences of history

CB History of civilization
CB156 . *Terrestrial evidence of interplanetary voyages*
CB440-481 . Relation to special topic
CB440 . . *Astronautics and civilization*

CJ Numismatics
CJ1-4625 . Coins
CJ161 . . Symbols. devices, etc., A-Z
CJ161.B2 . . . *Bridges*
CJ 161.S5 . . . *Ships*
CJ 161.T73 . . . *Transportation*

3 pages from F

G Geography

GN Anthropology

GN301-673 . Ethnology. Social and cultural anthropology
GN406-517 . . Cultural traits, customs, and institutions
GN406-442 . . . Technology. Material culture
GN438-442 . . . . *Transportation*
GN438 . . . . . *General works*
GN438.2 . . . . . *General special*
GN439 . . . . . *Routes of communication*
Including trails, roads, bridges, etc.
GN440-440.2 . . . . . *Transportation by water. Navigation*
GN440 . . . . . *General works*
GN440.2 . . . . . *Boats*
GN440.2 . . . . . *.Canoes*
GN441 . . . . . *Vehicles. Wheels*
GN442 . . . . . *Snowshoes. Skis*
GN448-450 . . . . Economic organization. Economic anthropology
. . .

GT Manners and customs

GT3400-5090 . Customs relative to public and social life
GT5010 . . Official ceremonies of royalty, nobility, etc.
. . .
GT5220-5285 . *Customs relative to transportation and travel*
Cf, BJ2137+Etiquette of travel
G149+ Voyages and travels
G540 Seafaring life
GT490 Customs relating to wayfarers
HE Transportation
GT5220 . . *General works*
. . *By period*
GT5230 . . . *Ancient*
GT5240 . . . *Medieval*
GT5250 . . . *Modern, through 1800*
GT5260 . . . *1801-*
. . *Vehicles. Chariots. Cars*
GT5280 . . . *General works*
GT5285 . . . *Sleighs and sledges*
Horses, see GT5885

H Social sciences

HD **Economic history and conditions** [See sample pages following]

HE **Transportation and communications** [See sample pages following]

HF **Commerce**
HF5001-6182 . Business
HF5601-5689 . . Accounting. Bookkeeping
HF5686 . . . By business or activity, A-Z
A list of seven pages, including
HF5686.A38 . . . . *Air transportation. Airlines*
Railways, see HE2241 [Accounting under Railways]
HF5686.T6 . . . . Tobacco
HF5686.T7 . . . . *Transportation*
Transportation, air, see HF5686.A38
Transportation, automotive,
see HE5618 Automotive transportation > Finance, accounting, etc.
Transportation, local,
see HE4351 Street railways. Subways. Rapid transit systems > Finance. Accounting. Auditing
HF5686.T73 . . . . Travel agents
HF5686.T8 . . . . Trustees

HJ **Public finance**
HJ2240-5957 . Revenue. Taxation. Internal revenue
HJ3231-3696 . . Taxation. Administration and procedure
[Note: Transportation taxes used to be here, but have been moved to HE: HE196.9 [Taxation under Transportation in general] or HE384+ [Control, taxation, tolls, etc. under Water transportation]

HQ **The family. Marriage. Woman**
HQ503-1064 . The family. Marriage. Home
HQ1060-1064 . . Aged, Gerontology (Social aspects). Retirement
HQ1063.5 . . . *Transportation*

HV **Social pathology. Social and public welfare. Criminology**
HV697-4959 . Protection, assistance, and relief
HV1551-3024 . . Handicapped
HV1568.6 *Transportation and travel*
For transportation of persons with specific handicaps, see the specific class of handicapped persons. [But not all have a subclass transportation.]
HV3011-3024 . . . Physically handicapped
HV3022 . . . . *Transportation and travel*
HV3025-3174 . . Special classes. By occupation
HV3025-3164 . . . *Seamen*

Sample pages from HD and HE, an even number, now 10 (2002)

Make this number of pages even

J Political science

JF-JQ	**Political institutions and public administration**
JK	. United States
JK401-1685	. . Government. Public administration
JK468	. . . Other special, A-Z
JK468.T7	. . . . *Transportation*

K Law

KF	**United States — general**
KF1600-2940	. Regulation of industry, trade and commerce. Occupational law
KF2161-2849	. . *Transportation and communication* [10 pages, divided by type of transportation, for example]
KF2271-2379	. . . *Railroads* [with much detail]
KFC	**United States — California**
KFC390-547	. Regulation of industry, trade and commerce. Occupational law
KFC469-543	. . *Transportation and communication* *[3 pages, divided by type of transportation, but less detail.]*
KFC490-499	. . . *Railroads [only three classes under here]*

L Education

LC	**Special aspects of education**
LC65-245	. Social aspects of education
LC189-214.53	. . Educational sociology
LC213-214.53	. . . Educational equalization. Right to education
LC214-214.53	. . . . School integration
LC214.5-53	. . . . . Special means of integration
LC214.5-53	. . . . . *Transportation. Busing*

M Music and books on music

M	**Music**
M1497-5000	. Vocal music
M1497-1998	. . Secular vocal music
M1900-1980	. . . Songs (part and solo) of special character
M1977-1978	. . . . By topic, A-Z [A three-page list, including]
M1977.R3	. . . . . *Railroads*
M1977.T87	. . . . . *Truck drivers*

N Fine arts

NA Architecture

NA4100-8480 . Special classes of buildings
NA4170-8480 . . Classed by use
NA4170-7010 . . . Public buildings
NA6290-6370 . . . . *Transportation and storage buildings*
NA6290 . . . . . *General works*
NA6300-6307 . . . . . *Airport buildings*
Divided like NA4410-4417
NA6310-6317 . . . . . *Railway stations*
NA6320-6327 . . . . . *Bus terminals*

NC Drawing. Design. Illustration

NC760-825 . Special subjects (Technique, history, and collections)
NC825 . . Other subjects, A-Z
NC825.A4 . . . *Airplanes*
NC825.A8 . . . *Automobiles*
NC825.B6 . . . *Boats*
NC825.B7 . . . *Bridges*
. . .

P Language and literature

PN Literary history and collections

PN6147-6231 . Wit and humor, satire
PN6231 . . Collections on special topics
[A ten-page list, including]
PN6231.T68 . . . *Traffic regulations*

Q Science

QC Physics

QC251-338.5 . Heat
QC290-297 . . Calorimeters and calorimetry
QC293 . . . Special types of calorimeters, A-Z
QC293.F8 . . . . *Fuel* [related to transportation]

QH Natural history (General). Biology (General)

QH540-549.5 . Ecology
Qh545.A1-Z . . Influence of special factors in the environment
QH545.A2-Z . . . Special, A-Z
QH545.A3 . . . . *Air pollution* [related to transportation]
[Note: In Germany, speed limits were introduced to cut emissions thought harmful to forests]

R Medicine

RA	**Public aspects of medicine**
RA1-1270	. Public health. Hygiene. Preventive medicine
RA772	. . Other subjects of public health, A-Z
RA772.T7	. . . *Traffic accidents*
	Cf. HE5613.5+, Motor vehicles
RC	**Internal medicine**
RC952-1245	. Special situations and conditions
RC970-986	. . Military medicine. Naval medicine
RC981-986	. . . *Naval medicine*
	Including merchant marine
	. . .
RC1030-1160	. . *Transportation medicine*
RC1040-1045	. . . *Automotive medicine (and classes under it)*
RC1050-1097	. . . *Aviation medicine (and classes under it)*

S Agriculture

SF	**Animal culture**
SF277-359.7	. Horses
SF311-312	. . *Draft horses*
SH	**Aquaculture. Fisheries. Angling**
SH209-399	. Fisheries
SH337	. . *Packing, transportation, and storage*
SH337.5	. . *Fishing port facilities*

T Technology

TE	***Highway engineering***
TF	***Railroads***
TG	***Bridge engineering***
TL	***Motor vehicles, aeronautics, astronautics***

[See two sample pages following. Note difference in perspective from class HE.]

Two pages from T

U Military science

UC **Maintenance and transportation**
UC270-360 . *Transportation*
[One-page broad classification of all modes of transportation, e.g.]
UC310-315 . . *Railroads*

V Naval science

[Almost all of this is relevant, see the detailed LCC outline. Especially]

VK ***Navigation. Merchant marine*** *and*
VM ***Naval engineering. Ship-building. Marine Engineering***
[Both refer to civilian water transport]

Z Bibliography and library science.

Z662-1000 **Libraries**
Z665-718.8 . Library science. Information science
Z675 . . Classes of Libraries, A-Z [Three-page listing, including]
Z675.N3 . . . *Naval*
Z675.T7 . . . *Transportation libraries*
Z687-718.8 . . The collections. The books
Z693-Z695.83 . . . Cataloging
Z695.1 . . . . By subject, A-Z
[four-page list, including]
Z695.1.N3 . . . . . *Naval art and science*
Z695.1.R34 . . . . *Railroads*
Z695.1.T73 . . . . . *Transportation*
Z696-697 . . . Classification and notation
Z697 . . . . By subject or form, A-Z [Two-page list, including]
Z697.T7 . . . . . *Transportation*

Z1001-8999 **Bibliography**
Z1001-1121 . General bibliography
Z1201-4980 . National bibliography
Z5051-7999 . Subject bibliography
Subjects arranged in alphabetical sequence [sic!]
Z5811-14 . . Education
Z5814 . . . Special topics, A-Z [Four-page list, including]
Z5814.T7 . . . . *Transportation of pupils*
Z7231-7234 . . *Railroads*
Z8001-8999 . Personal bibliography
Names of individuals arranged in alphabetical sequence

Lecture 8.2a (in Small Groups 1) ***March 8 - March 10***

Vocabulary control (terminological control)

Objectives	1 Understand the retrieval problems caused by terminological variety– synonymy and homonymy – in language, including any kind of names. 2 Understand and be able to apply vocabulary control to remedy these problems, either through vocabulary control in indexing or through query term expansion in searching. 3 Understand the structure of a thesaurus with its **synonym-homonym structure** (all terms), **classificatory structure** (concepts expressed by preferred terms), **index language** (concepts and corresponding preferred terms selected as subject descriptors), and **lead-in vocabulary** (all terms that are not subject descriptors).
Practical significance	• Authority control is applied to terms designating subjects, to names of persons and organizations, to titles of often cited or reprinted works, and in many other cases. It is a major principle underlying many information retrieval systems, especially those used in libraries. • Lack of vocabulary control and authority control more generally is one of the most serious problems impeding the success of end-user searching in free-text searching. The solution lies in the design of systems, including search thesauri, that can assist end users.

Wider applications	**Vocabulary control as a special case of authority control** Vocabulary control is the control of subject identifiers. Similar problems arise in the control of the identifiers of other types of entities, such as persons or organizations; thesaurus of organizational names. In the broader sense one speaks of authority control (see Sections 9.1.1 and 9.1.2). The purpose of authority control can also be stated as **referential integrity**, that is, assuring a one-to-one correspondence between entity values and the character strings or other symbols that refer to them.

Review of Organizing Information, Chapter 12 (20 min)

Lexical relationships (10 min.)

Paradigmatic relationships: Synonymy, antonymy, hyponymy

In linguistics: Relationships between terms based on their meanings, that is, on the concepts they designate. If a term has multiple meanings, only one of these meanings participates in the relationships discussed here.

In classification theory/knowledge representation: Relationships between concepts in a classificatory structure.

Paradigmatic relationships are contrasted with **syntagmatic relationships** that bind together words into phrases and sentences or elemental concepts into compound concepts, statements, or larger units of meaning.

Synonymy	Two terms designate the same concept. True synonyms can be used interchangeably in sentences without changing the meaning. Core meaning and connotations. Problem of shades of meaning and connotation.
Antonymy	Two terms designate opposite concepts. Opposites can be endpoints of a scale, such as *light* and *dark*, or exclusive categories, such as *male* and *female*
Hyponymy	Term A designates concept Concept A', term B designates concept Concept B', and Concept B' is more specific then Concept A'. Examples: *flute* (in one of its meanings) has as hyponym *recorder* (in one of its meanings); *keyboard instrument* has as hyponyms *harpsichord* and *cembalo*. Note: In a thesaurus with a controlled vocabulary we would select a preferred term, for example *harpsichord*, and have the relationships *keyboard instrument* Narrower Term *harpsichord* *harpsichord* Synonymous Term *cembalo* Antonymy and hyponymy are really concept relationships to be dealt with in Chapter 14; hyponymy is the relationship that defines a concept hierarchy. But all three relationships have in common that one term can be exchanged for the other in a sentence and still leave a sentence that has meaning.

Homonymy and polysemy

Note: The transition from homonymy to polysemy is gradual

<table>
<tr><td>Homonymy</td><td>Strict definition: two different words or phrases have the same spelling (homography) or the same pronunciation (homonymy in the narrowest sense).

Examples:
<table><tr><td>seal (marine mammal)
seal (document)</td><td>drill (bore a hole)
drill (furrow)
drill (fabric).</td></tr></table>
Note: While they are spelled the same, the words in each group have different etymological origin.</td></tr>
<tr><td></td><td>More expansive definition: The same word has two quite different meanings.

Examples:
<table><tr><td>drill (bore a hole)
drill (training)</td><td>seizure (disorder)
seizure (law enforcement).</td></tr></table>
Note: In each group, we have the same word (same etymological origin). The word acquired completely different meanings over time.</td></tr>
<tr><td>Polysemy</td><td>The same word has several meanings that can all be traced to a common core of meaning.

Example:
integration (mathematics)
integration (psychology)
integration (social groups)
integration (economic-political)
integration (curriculum)

All these meanings share a common core meaning: putting together pieces into whole where the pieces are held together in a larger structure.

Polysemy is often the result of metaphoric extension of the meaning of a term.

Example:
field (piece of land)
field (subject)
field (physics)
field (mathematics)

Even for drill (bore a hole) and drill (train) one can identify the core meaning of repetitive and persistent performance of an operation.</td></tr>
</table>

6 WordNet pages

Lecture 8.2b (in Small Groups 1) ***March 8 - March 10***

Index language functions (Organizing Information, Chapter 13) (60 min)

Subject analysis; abstracting and indexing; types and functions of abstracts

Objectives	1 Understand the principle of request-oriented (user-centered) indexing and the fundamental role of the index language to communicate users' interests to the indexer. 2 Be able to make intelligent decisions about the type of index language, indexing, and query formulation to be used in a given IR system, considering costs and benefits. 3 Be able to recognize search requests that are difficult to handle in a system that does not use request-oriented indexing and be able to compensate, as far as possible, through creative pursuit of different avenues for the search.
Practical significance	Request-oriented indexing (also called problem-oriented indexing or user-centered indexing) is a special case of the maxim that the design and operation of information systems should be based on a thorough understanding of user requirements. Request-oriented indexing is the key to good system performance for the questions that matter to users. Yet in practice it is rarely used. Understanding this will enable students to make the best of existing systems and, more importantly, to go out and change practice.
Discussion question	How could request-oriented indexing be implemented in a reference tool addressed to a general audience, such as the *Reader's Guide to Periodical Literature*?

Chapter highlights	• Derivation of the principle of **request-oriented indexing** from the problem-oriented approach to information systems introduced in Chapters 1 and 5 (Sections 13.1 and 13.2), • the role of index languages in searching and database organization (Sections 13.3 and 13.4), • design issues (Section 13.5), • review of index language functions (Section 13.6), • culminating in the recognition of an index language as a communication device from users to indexers, so that the indexers understand the users' interests (Section 13.7). Terminology: Filtering technique of indexing (Mooers 1958), Request-oriented indexing (DS 1974), problem-oriented indexing (DS), user-centered indexing (term in vogue now).
Questions	Your questions here
Discussion question (repeated)	How could request-oriented indexing be implemented in a reference tool addressed to a general audience, such as the *Reader's Guide to Periodical Literature*?

Document representation: purpose, structure, process of creation

Abstracts as a different form of document representation

Indicative abstract - merely indicates what the document is about or relevant for, pointer data.

Informative abstract - in addition, includes some of the substantive data given in the document or reports some generalization that can be derived from the document.

Both types of abstract assist the reader in deciding whether to pursue the document further (and incur any costs in doing so). An informative abstract often gives the substantive data needed and thus saves the user the trouble of having to consult the document itself.

Other categorization of abstracts: Reporting vs. analytical-critical. Book reviews

The structure of document representations (abstracts or lists of index terms) discussed in the lecture on document structure. Use of controlled vocabulary

Abstracting and indexing as a cognitive process
Empirical study of document-oriented indexing

Parts of the document considered

Method of information assimilation (reading, interpreting pictures)
Reading/scanning to identify subject matter of interest to users — request-oriented reading
Reading/scanning to fill slots of a frame
Building up mental image

Selecting topics to be included in the abstract or the index terms. Request-orientation comes into play here as well

Choosing a form of expression

Knowledge brought to bear on these operations - from own knowledge or tools (such as thesauri) consulted, for example

- General knowledge of the field
- Knowledge of user needs
- Frames for phenomena in the field
- Knowledge of terminology
- Knowledge of document structure, including knowledge of cue words

Automatic or computer-assisted abstracting and indexing

Conigrave KM, Saunders JB, Reznik RB. **Predictive capacity of the AUDIT questionnaire for alcohol related harm**. Addiction 90 (1995) 1479-1485.

Indicative abstract

This study deals with early identification of alcohol use disorders. It examined the ability of the Alcohol Disorders Identification Test (AUDIT) questionnaire published by the World Health Organization to predict which subjects experience medical or social harm from their drinking. Subjects were 350 emergency room patients who answered the AUDIT questions was part of a comprehensive medical assessment. 250 subjects were interviewed after 2-3 years to determine alcohol-related medical disorders, health care utilization, social problems and hazardous drinking at the time of follow-up. Audit is compared to biochemical indicators for its ability to predict these conditions.

Informative abstract

'**AUDIT can predict a range of harmful consequences of alcohol consumption**'

Background. Drinking problems often are not recognized. Most of the people who become alcohol-dependent do not seek help until their problems are obvious. Late diagnosis is of particular concern because effective and low-cost methods of treating problem drinking at an early stage are now available. In 1989, the WHO published a brief 10-item screening questionnaire, the Alcohol Disorders Identification Test (AUDIT) specifically designed to identify problem drinkers before physical dependence or chronic problems have arisen. AUDIT has been reported to have a sensitivity of 92% and a specificity of 94% in detecting hazardous or harmful alcohol use. This study examined the ability of the AUDIT questionnaire to predict which subjects experience medical or social harm from their drinking.

Methods. Subjects were 350 patients who attended a hospital emergency ward in 1984-1985. They underwent a comprehensive assessment of medical history, alcohol use, dependence and related problems in an interview schedule; the AUDIT questions were interspersed among other items. Biochemical variables measured included y-glutamyltransferase (GGT) and mean corpuscular volume (MCV). Twenty subjects refused to be contacted after 2-3 years or were excluded because of malignant disease. Thus, a cohort of 330 subjects (212 men, 108 women) was left for the longitudinal study; 250 subjects were interviewed again after 2-3 years. Interviewers were blind to the results of the initial assessment. The AUDIT questions were scored from 0 to 4. Subjects who scored 8 or more were classified as potentially hazardous drinkers. AUDIT was examined for its ability to predict a number of end-points including alcohol-related medical disorders, health care utilization, social problems and hazardous drinking at the time of follow-up.

Results. Of those who scored 8 or more on AUDIT at the initial interview, 61% experienced alcohol-related social problems compared with 10% of those with lower scores. They also reported more frequently alcohol-related medical disorders and hospitalization. The AUDIT score was a better predictor of social problems and of hypertension than laboratory markers. Its ability to predict other alcohol-related illnesses was similar to the laboratory tests, but GGT was the only significant marker of mortality.

Conclusions. AUDIT is a brief and convenient questionnaire which can readily be incorporated into the standard medical history. It can predict a range of harmful consequences of alcohol consumption. AUDIT should prove a valuable tool in screening for hazardous and harmful alcohol use so that intervention can be provided to those at particular risk of adverse consequences.

March 8 - March 10, 2011 Name (optional)

Free-write 8

Lecture 8.1 / Small Groups 1.
Explorations in subject access. Vocabulary control
Lecture 8.2. Index language functions

- **Reflect** – what you learned, what was most important, what was most interesting, what was extraneous;
- **Ask questions** – ask for more explanation, how is a concept connected to other concepts, why is a concept important, how can it be applied, why is a reading important;
- Offer **critique and suggestions**;
- Say anything else you want to.

Over

Lecture 9.1 ***March 22 - March 24***
Small Groups 2

Index language structure 1: conceptual (Organizing Information, Chapter 14)

Objectives of Lectures 9.1-11.1	Be able to use the entity-relationship approach, specifically facet analysis, to discern the conceptual structure of a subject. Put differently: Be able to discern the facet structure of a subject.
Practical significance	This understanding provides a basis for • constructing an index language, a task required in setting up specialized information systems and, more importantly, in developing expert systems; • evaluating an index language to determine whether it is suitable for a given application; • indexing, particularly making sure that all applicable facets have been covered; • query formulation, facet analysis of queries.

In-class exercises: Three steps in the **conceptual analysis and synthesis** in a subject:

Step 1.	Semantic factoring (results in a list of elemental concepts).
Step 2.	Arranging the elemental concepts in a well-structured hierarchy.
Step 3.	Fit compound concepts into the framework of the hierarchy (if compound concepts need to be dealt with explicitly)

In-class exercise: Semantic factoring

Semantic factor the concepts (Dewey classes) from the attached list.
Keep a running list of elemental concepts as they arise.

In-class exercise: Building a hierarchy of elemental concepts

Sort elemental concepts into entity types or facets.

Arrange values within each entity type or facet in a meaningful structure.

In-class exercise: Fitting compound concepts into a hierarchy

In-class exercise, Organizing Information, Chapter 14: Semantic factoring

Factor the following concepts (from Dewey Decimal Classification) **into their semantic components (semantic factors)**. If this is not possible, comment.
Keep a running list of the elemental concepts needed.

Note: A broader class is given in ( ) if necessary to specify the meaning of a term.

372.19	Curriculums of elementary schools
372.35043	Science in the elementary school curriculum
372.414	Methods of instruction for reading in elementary schools
372.72043	Arithmetic in the elementary school curriculum
373.19	Curriculums in secondary schools
373.243	Military schools (Secondary Education)
376.63	Secondary education of women
378.19	Curriculum of colleges and universities
378.33	Fellowships (Higher Education)
371.7	School health and safety
371.855	Men's social societies and fraternities (Generalities of Education)
371.856	Women's social societies and sororities
371.911	Blind and partially sighted students
371.912	Deaf and hard-of-hearing students
371.95	Curriculums for gifted students

In-class exercise: Building a hierarchy of elemental concepts

Sort elemental concepts into entity types or facets.

Arrange values within each entity type or facet in a meaningful structure.

Elemental concepts Running list	**Elemental concepts in a meaningful structure**

In-class exercise: Fitting compound concepts into a hierarchy

Lecture 9.2 ***March 22 - March 24***
Small Groups 2

Application of index language structure to searching

(Organizing Information, Section 14.4)

Objectives Inherited from Lect. 9.1-11.1 plus these	1 Understand inclusive searching (hierarchically expanded searching). 2 Be able apply this concept in searching any system.
Practical significance Inherited from Lect. 9.1-11.1 plus these	Inclusive searching is an essential technique for searches that emphasize recall.

In-class exercise: Retrieval of documents in a sample collection

The sample collection consists of about 200 documents on transportation and traffic and is indexed using the index language shown on the following pages (same as the index language used in Assignment 11, *Request-oriented indexing*).

Query statement: I need information on vehicles used in rail transport

Query formulation: E6 Vehicles AND B2 Rail transport

Search in a printed index: Look for document numbers listed for both E6 Vehicles and B2 Rail transport (they are marked with = in the entry for B2).

E6 Vehicles 10, 12, 13, 24, 25, 26, 30, 36, 40, 46, 47, 50, 53, 54, 58, 59, 62, 64, 70, 76, 77, 79, 80, 81, 85, 91, 92, 94, 95, 100, 101, 102, 103, 104, 105, 106, 108, 109, 110, 116, 118, 121, 122, 126, 127, 132, 133, 134, 138, 148, 150, 151, 153, 155, 168, 169, 170, 171, 173, 174, 176, 178, 180, 181, 186, 187, 188, 191, 192, 193, 194, 199, 202, 204, 205, 207, 210, 211, 212, 213, 214, 216, 218, 219, 322, 330, 332, 333, 336, 337, 340, 346, 347, 353, 354, 355, 356, 357, 358, 362

B2 Rail transport =10, =24, 34, 41, 42, 44, =50, 89, 114, =126, 140, 149, 166, 184, =191, 195, =213, 310, 334

B2.3 Intercity railroads =10, 30, =46, =62, =64, =79, 82, 84, 97, =102, =108, 114, =122, =132, 156, 177, =213, 341, 362

B2.7 Local rail transit 27, 56, 87, 99, =108, 111, 120, 123, 129, 130, =213, 327, 345, 350, =356

Question: Did this search find all relevant documents?

Additional index entry to solve the problem

B2 Rail transport, inclusive =10, =24, 27, =30, 34, 41, 42, 44, =46, =50, 56, =62, =64, =79, 82, 84, 87, 89, 97, 99, =102, =108, 111, 114, 120, =122, 123, =126, 129, 130, =132, 140, 149, 156, 166, 177, 184, =191, 195, =213, 310, 327, 334, 341, 345, 350, =356, =362

Search in a peek-a-boo file (some samples distributed)

Like a printed index, but with more manipulative power. Today one uses computers.

Each descriptor has its own card. Each document number has a position on the card. In a printed index, the applicable document numbers are listed after the descriptor. In a peek-a-boo file, the applicable document number positions are punched on the descriptor card. In this particular implementation, document numbers are read off as follows: Find the column number (printed in tiny print), for example **12**). Find the row number (large single digits printed across each row, punched out if the position is punched), for example **6**. The document number is **126**. (One of the first uses of peek-a-boo cards: A bird guide. Retrieval of birds based on their features.)

This peek-a-boo file makes provision for inclusive searching: Each descriptor that has narrower descriptors under it has two cards: An **inclusive card** that includes all the documents form the narrower descriptors as well, and a **general references card** that includes only the documents indexed by the descriptor itself.

To find documents for

E6 Vehicles AND B2 Rail transport, inclusive

superimpose the two cards and read off the document numbers from the holes that still appear (document numbers in common to both cards).

We will broaden and narrow the search to observe the effects of hierarchy.

File builders and searchers classification display

The descriptors shown in italics with numbers D1.xx are precombined descriptors. Each system using this index language would decide whether to use these precombined descriptors (such as *D1.20 Aircraft*) or whether to index with the corresponding elemental descriptors (in the example D1 Air transport and E6 Vehicles) instead. **In assignment 11 only elemental descriptors are used for indexing.**

Outline: Facets

B	**Division by mode of transportation**
E	**Transportation system elements**
F	Power supply for vehicles
G	Type of propulsion
H	Materials to build facilities or vehicles
J	**Passenger transport vs. freight transport**
K	Traffic operations
L	Transportation providers
M	Creation of traffic systems and components
N	Organization, administration
Q	General and other concepts
R	Geographic range
S	Geographic location

The three facets used for arrangement are shown in **bold**.

In the full display

+ signifies descriptors that have Narrower Terms under them

+A Transportation and traffic, inclusive

A . Transportation and traffic, gen. references

+B Mode of transport, inclusive

B . Mode of transport, general references *22
+B1 . Ground transport, inclusive
B1 . . Ground transport, general references 24
+B2 . . Rail transport, inclusive
B2 . . . Rail transport, general references 25
B2.3 . . . Intercity railroad BT R4 26
B2.7 . . . Local rail transit BT R2 27
B4 . . Road transport 28
B6 . . Pipeline transport
B7 . . Pedestrian mode
B8 . . Multi-modal ground transport
+C1 . Water transport, inclusive
C1 . . Water Transport, general references 30
C3 . . Inland water transport 31
C7 . . Ocean Transport 32
+D1 . Air transport, inclusive
D1 . . Air transport, general references 34
D1.10 . . Air transport facilities, inclusive BT E6
+D1.20 . . Aircraft, inclusive BT E6
D1.20 . . . Aircraft, general references 38
+D1.30 . . . Airplanes, inclusive
D1.30 . . . . Airplanes, general references 40
D1.35 . . . . Conventional airplanes 41
+D1.40 . . . . New generation airplanes, incl.
D1.40 . . . . . New generation airplanes, g. r.
D1.44 . . . . . VTOL
D1.45 . . . . . STOL
D1.46 . . . . . Variable geometry airplanes
D1.60 . . . Helicopters 43
D4 . . Supersonic air transport 35
D7 . Multi-modal transport
D98 . Other specific modes of transportation
D99 . Mode of transportation not applicable

+E Transportation system elements, incl

E . Transportation system elements, gen. ref. 15
+E1 . Traffic facilities, inclusive NT D1.10
E1 . . Traffic facilities, general references 17
E2 . . Traffic ways 18
E3 . . Traffic stations 19
E4 . . Stationary equipment
+E5 . Methods to move persons or freight, incl.
E5 . . Methods to move persons or freight, g. r.
+E6 . . Vehicles, inclusive NT D1.20, F, G
E6 . . . Vehicles, general references 20
E7 . . . Air cushion craft
E8 . . Containers
E9 . . Self-transport
E98 . Other concepts
E99 . Transp. system elements not applicable

+ Inclusive (hierarchically expanded, finds documents on all narrower terms as well)

+F Power supply for vehicles, incl. BT E6

F . Power supply for vehicles, gen. references
+F1 . Hydrocarbons, inclusive
F1 . . Hydrocarbons, general references
F2 . . Gasoline
F3 . . Diesel fuel
F4 . . Hydrocarbons from renewable sources
F5 . Electric power
F6 . Nuclear power
F7 . Animate power
F98 . Other power supply
F99 . Power supply not applicable

+G Type of propulsion, inclusive BT E6

G . Type of propulsion, general references
+G1 . Engine, inclusive
G1 . . Engine, general references
+G2 . . Combustion engine, inclusive
G2 . . . Combustion engine, general ref.
G3 . . . Cylinder engine
G4 . . . Rotary engine
G5 . . Steam engine
G6 . Turbines
G7 . Walking
G98 . Other type of propulsion
G99 . Type of propulsion not applicable

+H Materials to build facilities or vehicles, inclusive

H . Materials to build facilities or vehicles, g. r.
H12 . Soils, aggregates
H14 . Bitumen
H16 . Cement, Concrete
H18 . Metal
H22 . Marking or coating materials
H24 . Adhesives, seals
H26 . Ceramics, glasses
H28 . Fibers, textiles
H32 . Plastics
H34 . Rubbers
H36 . Wood, paper
H38 . Petroleum products
H98 . Other specific materials
H99 . Material not applicable

* Edge-notched card hole no.

+J Passenger transport vs. freight transport

J . Passenger transport vs freight transport, g.r. 8
J3 . Passenger transport 9
+J4 . Freight ransport, inclusive
J4 . . Freight ransport, general references 11
J6 . . Transport of material of heavy weight 12
J7 . . Transport of bulk material 13
J99 . Passenger vs. freight transport not applicble

+K Traffic operations, inclusive

K . Traffic operations, general references
+K1 . Traffic communication, control, safety, incl.
K1 . . Traffic comm., control, safety, gen. ref. 4
K2 . . Traffic communications
K3 . . Traffic control 6
K4 . . Traffic safety 7
+K5 . Routes and schedules, inclusive
K5 . . Routes and schedules, general references
K6 . . Routes, route systems, traffic networks 2
K7 . . Schedules
K8 . Handling, loading, unloading
K98 . Other specific traffic operations
K99 . Traffic operations not applicable

+L Transportation providers, inclusive

L . Transportation providers, general references
L2 . Organizations, companies
L6 . Personnel, operators
L99 . Transportation providers not applicable

+M Creation of traffic systems and components, inclusive

M . Creation of traffic systems & components, g
+M1 . Research, design, and evaluation, inclusive
M1 . . Research, design, and evaluation, g. ref.
M2 . . Research and development
M3 . . Planning
M4 . . Design
M5 . . Testing, demonstration, evaluation
M6 . Manufacturing, construction
M7 . Acquisition
M8 . Training
M9 . Maintenance
M98 . Other specific activities in system creation
M99 . System creation not applicable

+N Organization, administration, incl.

N . Organization, administration, gen. reference
N2 . Administration, management
N4 . Costs, financing
N6 . Marketing
N8 . Legal aspects
N98 . Other specific topics in organization
N99 . Organization, administration not applicable

+Q General and other concepts inclusive

Q . General and other concepts, gen. references
Q22 . Traffic flow
Q24 . Simulation 3
+Q40 . System characteristic, inclusive
Q40 . . System characteristics, general references
Q41 . . Noise, vibration
Q42 . . Pollution
Q43 . . Quality, performance
Q44 . . Durability, life, reliability
Q45 . . Demand, use
Q46 . . Human characteristics
Q47 . . Community characteristics
Q49 . . Other characteristics
+Q60 . Small vs large capacity, inclusive
Q60 . . Small vs. large capacity, gen. references
Q63 . . Small capacity
Q67 . . Large capacity
+Q70 . Civilian vs military, inclusive
Q70 . . Civilian vs military, general references
Q73 . . Civilian
Q77 . . Military
Q99 . Other concepts not applicable

+R Geographic range, inclusive

R . Geographic range, general references
+R1 . Local systems, inclusive
R1 . . Local Systems, general references
+R2 . . Urban systems, inclusive NT B2.7
R2 . . . Urban systems, general references
R3 . . Rural systems
+R4 . Beyond local systems, inclusive NT B2.3
R4 . . Beyond-local systems, general references
R5 . . Interurban systems
R6 . . State-wide systems
R7 . . National systems
R8 . . International systems
R98 . Other specific range
R99 . Geographic range not applicable

+S Geographic location, inclusive

S . Geographic location, general references
+S1 . North and Central America, inclusive
S1 . . North and Central America, gen. ref.
S2 . . U.S.
S3 . . Central America
S4 . South America
S5 . Europe
S6 . Asia
S7 . Australia
S8 . Africa
S98 . Other geographic locations
S99 . Geographic location not applicable

In-class exercise: Retrieval access and hierarchy

Below are six documents which were indexed in the request-oriented approach you used in Assignment 11. Each descriptor is on a separate line. Using the hierarchy of the index language in the *File builder's and searcher's display* (see preceding pages), do the following:

1 For each descriptor (index term), list the descriptor(s) under which the document should be found on the basis of this index term.

2 Give some query formulations retrieving the document. The query formulations should illustrate how a search for a combination of two broad concepts finds documents indexed by more specific concepts.

Document 1 is a filled-in example

Document 1

Automatic control of freeway ramp traffic, P.J.ATHOL. SAE—Analysis & Control of Traffic Flow Symposium—Conf Proc. Jan 9-10 1968 paper 680172 p 61-5.

Major problem in operating transportation system is traffic overloading demands at peak periods; expressway Surveillance Project was formed to improve efficiency of highway system through application of electronic automation and traffic engineering to problem of traffic congestion.; by providing means for quick response in case of accidents and fast removal of hindrances. Volume capacity of freeways was effectively increased during peak periods; use of ramp metering controls achieved reduction in delay, safer merging characteristics, and reduced freeway accidents.

Descriptors assigned		**Descriptors under which the document should be found**
B4	Road transport	B4, B1 inclusive, (B incl)
E2	Traffic routes	E2, E1 inclusive, (E inclusive)
K1	Traffic communication, control, safety	K1 gen ref, K1 inclusive, (K inclusive)
M3	Planning	M3, M1 inclusive, (M inclusive)
M4	Design	M4, M1 inclusive, (M inclusive)
M5	Testing, demonstration, evaluation	M5, M1 inclusive, (M inclusive)
Q22	Traffic flow	Q22, (Q inclusive)

Query formulations

B1 Ground tr., incl. AND E1 Tr. fac., incl.
B4 Road tr. AND K1 Tr.comm., contr., incl.

Document 2
Antwerp's new container dock, K.W.Flitcroft for the Antwerp Harper Committee. Dock & Harbor Authority v 49 n 571 May 1968 p 28-30.

Dock described is protected by locks from rise and fall of tides; spreader is employed in lifting of containers and is adaptable in spread to handle both long and short types; containers can be stored on quay and special connections for powering of plants of refrigerated containers are set in concrete paving every 10 ft.; set of rail tracks runs along quay between high legs of container cranes to bring rail-hauled containers directly for lifting off.

Descriptors assigned		**Descriptors under which the document should be found**
C7	Ocean transport	
E3	Traffic stations	
J4	Transport of freight, material, cargo	
K8	Handling, loading, unloading	
R8	International system	
S5	Europe	

Query formulations

Document 3. **Rolling Stock for London Transport's Victoria Line**

Descriptors assigned		**Descriptors under which the document should be found**
B2.7	Local rail transit	
E6	Vehicles	
F5	Electric power	
G1	Engine	
M4	Design	
M7	Acquisition	
Q49	Other characteristics (automation)	
S5	Europe	

Query formulations

Document 4. **Air Transp. 1975 and Beyond - Systems Approach**

Descriptors assigned		**Descriptors under which the document should be found**
D1	Air transport	
E	Transportation system elements	
J	Passenger transp. vs. freight transp.	
K	Traffic operations	
M2	Research and development	
N	Organization, administration	
Q70	Civilian vs. military	
R7	National systems	
S2	U.S.	

Query formulations

Document 5. **Technical and Economic Prospects of Air Cargo Traffic**

Descriptors assigned		**Descriptors under which the document should be found**
D1.20	Aircraft	
F1	Hydrocarbons	
G6	Turbines	
J4	Transport of freight, material, cargo	
K8	Handling, loading, unloading	
M1	Research, design, and evaluation	
Q22	Traffic flow	
Q45	Demand, use	
Q49	Other characteristics (automation)	
R8	International system	
S	Geographic location	

Query formulations

Document 6. **United States Subway Requirements 1968-1990**

Descriptors assigned		**Descriptors under which the document should be found**
B2.7	Local rail transit	
E1	Traffic facilities	
M5	Manufacturing, construction	
N4	Costs, financing	
N6	Marketing	
Q24	Simulation	
Q40	System characteristic	
S2	U.S.	

Query formulations

Free-write 9

Lecture 9.1. Index language structure 1

Lecture 9.2. Application of index language structure to searching

- **Reflect** – what you learned, what was most important, what was most interesting, what was extraneous;
- **Ask questions** – ask for more explanation, how is a concept connected to other concepts, why is a concept important, how can it be applied, why is a reading important;
- Offer **critique and suggestions**;
- Say anything else you want to.

Over

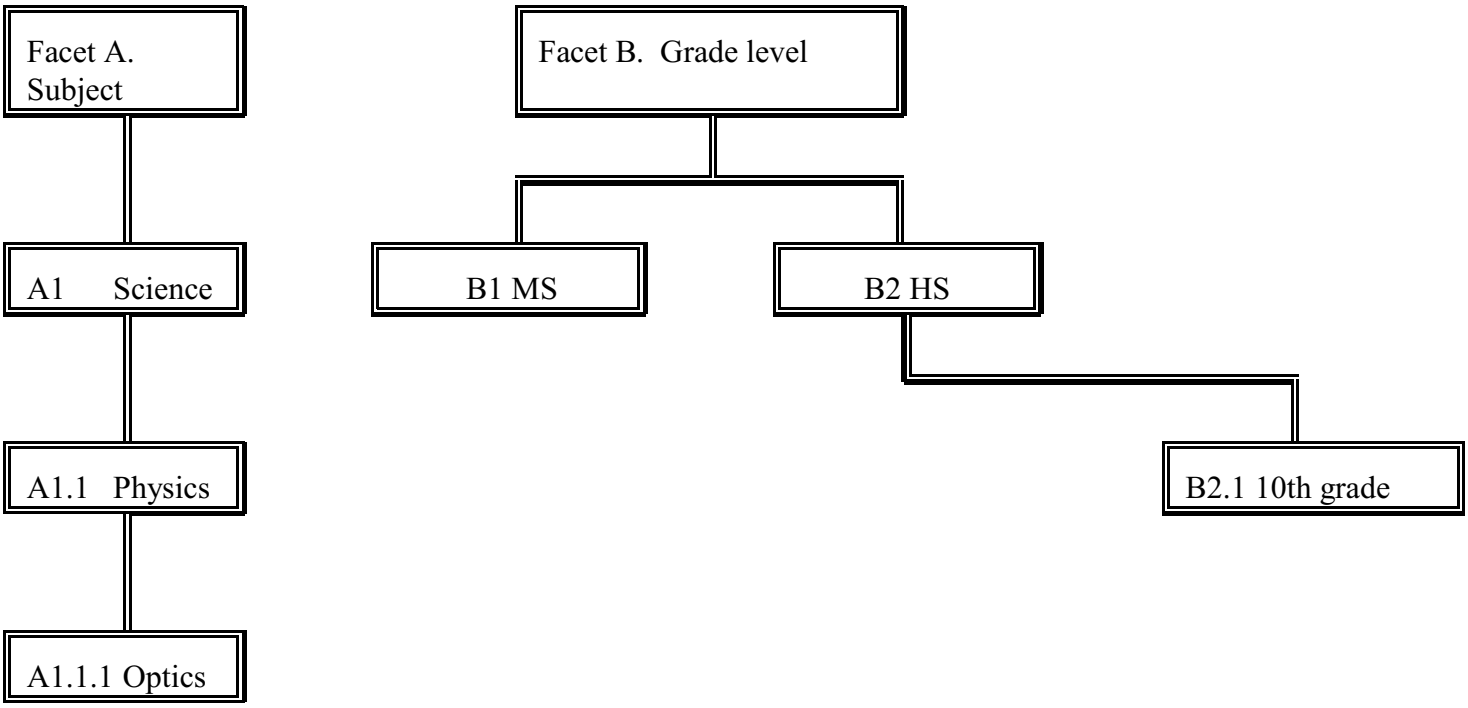
Small Groups 3. Lecture 10.1

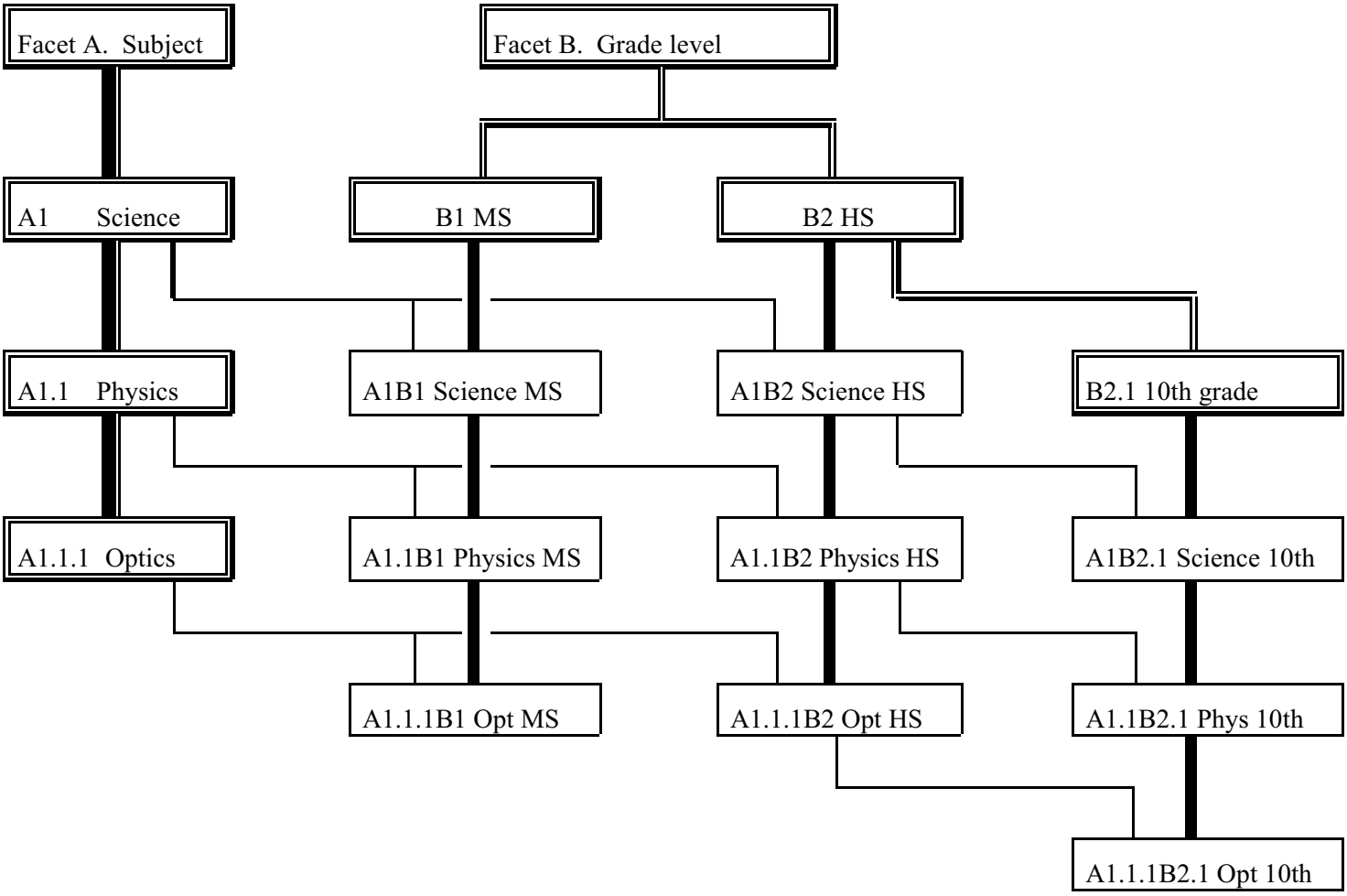
Hierarchy from Facets *March 29 - March 30*

Information system of instructional materials. Two facets, only between-facet combinations

Objectives Inherited, plus	Understand complex hierarchies that result from combining hierarchically structured facets.
Practical significance	Basis for understanding search Basis for understanding structure of DDC, LCC, LCSH, and similar systems

Process	Step 1: Form all possible between-facet combinations (do not combine with facet heads). Step 2: Find all hierarchical relationships. (Specifying all BT one level up defines the hierarchy completely.) Step 3: Represent the hierarchy Step 3a: – as a two-dimensional graph Step 3b: – as a linear arrangement with indention plus cross-references.
Application to retrieval	In a system using only elemental descriptors In a system using precombined descriptors with multiple entry (such as LC Subject Headings) In a system using precombined descriptors with single entry (such as Library of Congress Classification)


Facet A.
Subject
A1 Science
A1.1 Physics
A1.1.1 Optics
Facet B. Grade level
B1 MS
B2 HS
B2.1 10th grade



A Facet A. Subject

. A1 Science
. . A1B1 Science MS NT A1.1B1; BT B1
. . A1B2 Science HS NT A1.1B2; BT B2
. . . A1B2.1 Science 10th grade NT A1.1B2.1; BT B2.1
. . A1.1 Physics
. . . A1.1B1 Physics MS NT A1.1.1B1; BT A1B1
. . . A1.1B2 Physics HS NT A1.1.1B2; BT A1B2
. . . . A1.1B2.1 Phys 10th gr NT A1.1.1B2.1; BT A1B2.1
. . . A1.1.1 Optics
. . . . A1.1.1B1 Optics MS BT A1.1B1
. . . . A1.1.1B2 Optics HS BT A1.1B2
. . . . . A1.1.1B2.1 Optics 10th grade BT A1.1B2.1

B Facet B. Grade level

. B1 MS NT A1B1
. B2 HS NT A1B2
. . B2.1 10th grade NT A1B2.1

A Facet A. Subject

. A1 Science NT B1A1, B2A1
. . A1.1 Physics NT B1A1.1, B2A1.1
. . . A1.1.1 Optics NT B1A1.1.1, B2A1.1.1

B Facet B. Grade level

. B1 MS
. . B1 A1 MS Science BT A1
. . . B1 A1.1 MS Physics BT A1.1
. . . . B1 A1.1.1 MS Optics BT A1.1.1
. B2 HS
. . B2 A1 HS Science NT B2.1A1; BT A1
. . . B2 A1.1 HS Physics NT B2.1A1.1; BT A1.1
. . . . B2 A1.1.1 HS Optics NT B2.1A1.1.1; BT A1.1.1
. . B2.1 10th grade
. . . B2.1 A1 10th grade Science BT B2A1
. . . . B2.1 A1.1 10th grade Physics BT B2A1.1
. . . . . B2.1 A1.1.1 10th grade Optics BT B2A1.1.1

Small Groups 3. Lecture 10.2 (25 min.) ***March 29 - March 30***

Brief discussion of Assignments 13.1-4 on the examination of classification schemes / thesauri

Analysis of Knowledge Organization Systems (classification schemes, thesauri, etc.) based on their hierarchical structure, facet structure, and citation order. See the calendar or the assignment page for schedule of assignment activities and due dates.

Objectives	1 Develop a more complete understanding of the general concepts of classification structure by applying them to concrete schemes. 2 Gain practical experience with a semi-faceted classification used on the Web (Assignment 13.4 Yahoo), create familiarity with specific schemes that are widely used in libraries in the US (Assignments 13.1 DDC, 13.3 LCSH and 13.4 LCC), and become acquainted with a wide range of schemes used for a wide variety of purposes (Assignment 13.2 ERIC and Lecture 13.1, Exploration of classification schemes and thesauri). You can grasp the structure of these schemes better by applying a general conceptual framework to their analysis. Assignments 13.1 - 13.4 help you understand specific schemes by application to specific problems in cataloging (indexing) and query formulation for searching. Where available, these assignments introduce the electronic form of a scheme as well.
Practical significance	• A good working knowledge of faceted classification principles is important for the conceptual analysis of queries as a basis for developing good query formulations in any system. • Knowledge of specific schemes is important for searching catalogs and indexes based on those schemes (including catalogs of Web documents). • Knowledge of the variety of schemes that exist for different purposes is important for being able to work in many different applications and for recognizing where classification could be useful.

Over

Turn to Assignment 13, Assignments p. 137 (or thereabouts)

Assignment 13. Subject cataloging and searching practice

- Assignment 13.1, Dewey Decimal Classification (DDC) (6 hours)
- Assignment 13.2, ERIC Thesaurus (3 hours)
- Assignment 13.3 Library of Congress/Sears Subject Headings (LCSH) (5 hours)
- Assignment 13.4 Yahoo, Yahoo classification (a semi-faceted classification) (6 hours)
  OR LCC, Library of Congress Classification (LCC) (6 hours)
  OR Media. Media Streams iconic classification (6 hours)
  OR Own choice

March 29 - March 30, 2011 **Free-write 10**

Lecture 10.1. Hierarchy from facets

Lecture 10.2 Brief discussion of Assignment 13.1-4 on the examination of classification schemes / thesauri

- **Reflect** – what you learned, what was most important, what was most interesting, what was extraneous;
- **Ask questions** – ask for more explanation, how is a concept connected to other concepts, why is a concept important, how can it be applied, why is a reading important;
- Offer **critique and suggestions**;
- Say anything else you want to.

Lecture 11.1 ***April 6***

Index language structure 2: database organization (Organizing Info., Chapter 15)

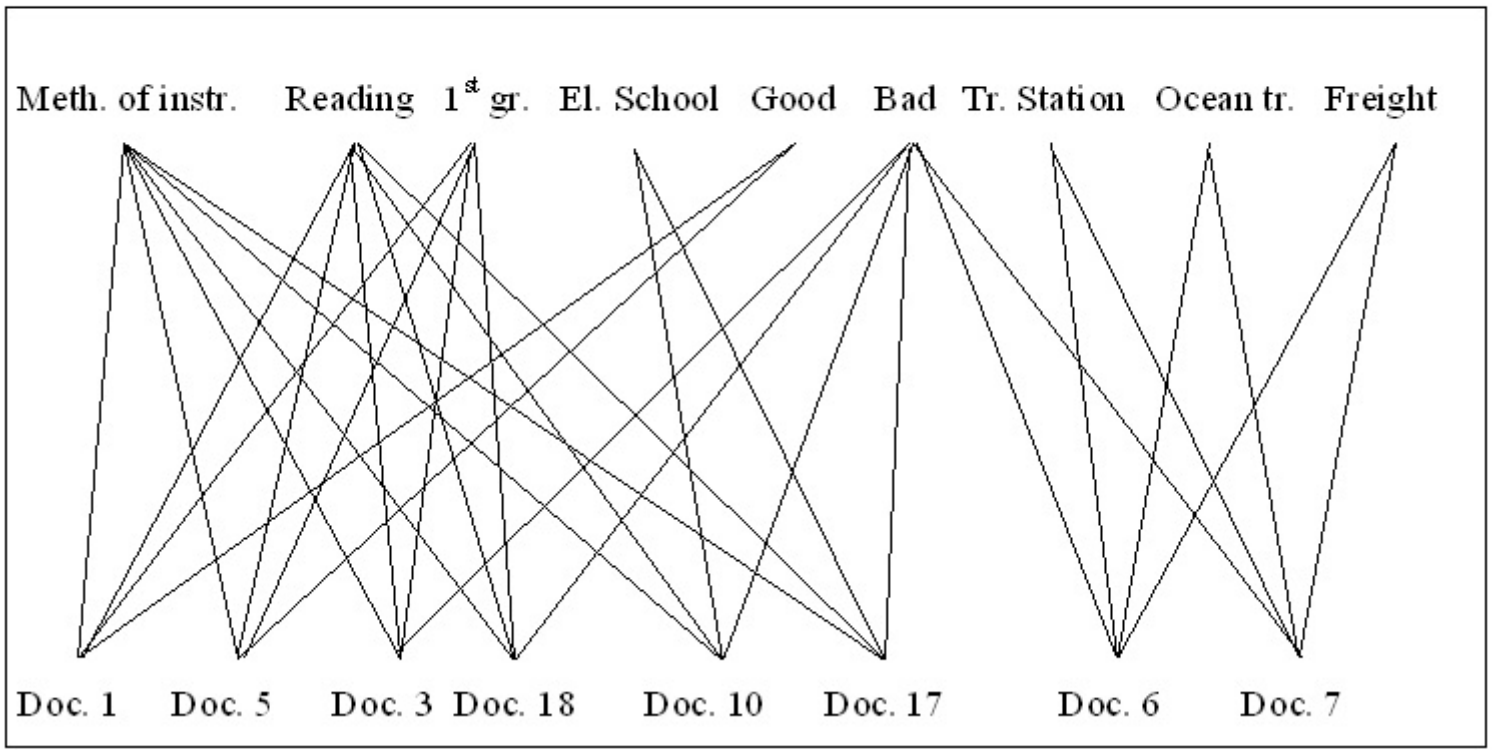
Objectives Inherited from Lect. 9.1-11.1 plus these	1 Understand **postcombination** and **precombination** - more generally, the degree of precombination — and how they relate to the retrieval mechanism used. 2 Be able to match the index language structure to the database organization and search mechanism available. 3 Understand the effect of precombination on index language structure and searching and be able to apply this understanding to the analysis of classification schemes such as DDC and LCC and improved searching with such schemes. 4 Understand the access mechanisms that help a user find the proper descriptors in a large classification scheme with many precombined descriptors, in particular **cross-references** and a **descriptor-find index**. 5 Understand principles of meaningful arrangement of search results.
Practical significance Inherited from Lect. 9.1-11.1 plus these	In conjunction with Chapter 14, this chapter establishes the foundation for understanding • the structure of systems used in libraries — and increasingly for the arrangement and display of electronic information — such as the Dewey Decimal Classification (DDC), the Library of Congress Classification (LCC), the Yahoo classification, and the Library of Congress Subject Headings (LCSH); • the structure of Web directories designed for browsing; • ad-hoc arrangement of retrieval results based on the analysis of noun phrases as compound concepts, as in the next-generation Web search engines.
Discussion question	Consider the design of an interface to a public-access online catalog in an academic library that would assist users in finding the appropriate LC class number and the appropriate LC subject headings.

Discussion of Organizing Information, Chapter 15

Section 15.1 and 15.2	Further examination and explication of postcombination vs. precombination of the concepts chosen as descriptors and their relationship to database organization and search mechanism. Interpretation of postcombination and precombination in terms of the entity-relationship approach and semantic networks (see figures on the following two pages). Examples of applying these concepts to a better understanding of index languages such as the Library of Congress Classification and the Library of Congress Subject Headings.
15.3 and 15.4	Emphasis on looking at precombination as a matter of degree. Introducing precombined descriptors as an example of restructuring semantic networks using hierarchical inheritance.
15.5	Methods for organizing an index language for access. Emphasis on understanding the idea of a descriptor-find index. Section 15.5.2 on how then arrangement of precombined concepts is important for understanding classification schemes and for arranging search output or any type of information, in print or online.
15.6	A look into the future: the idea of a conceptually unified index language for different search mechanisms.

In-class exercises concluding Lectures 8.1-11.1
Semantic networks and precombined descriptors
Vocabulary control and hierarchical structure
Conceptual analysis and synthesis
Discussion question on OPAC (Online Public Access Catalog) interface

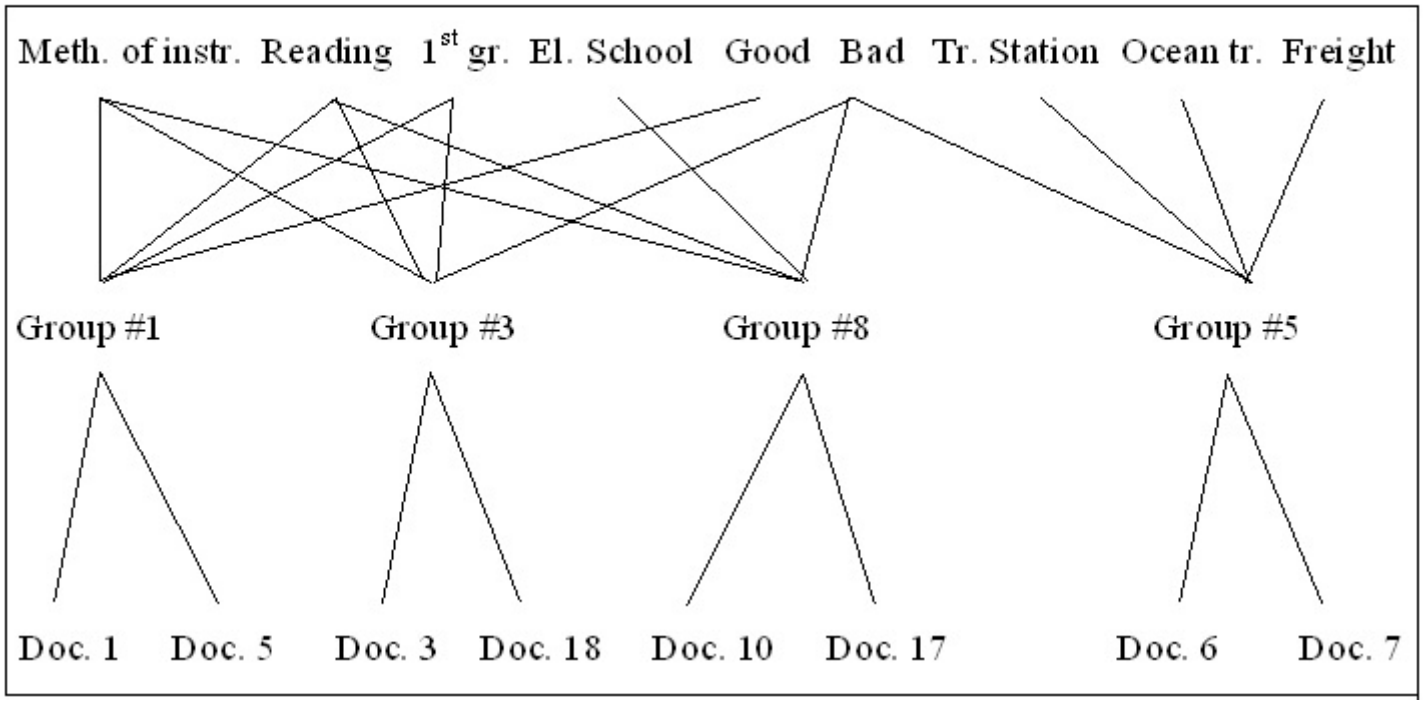
Semantic network representation of Database 15.1 (Organizing Info., p. 296)



In-class exercise: Semantic networks and precombined descriptors

Semantic network representation of Database 15.2

Semantic network representation of Database 15.2 (Organizing Info., p. 296)



Meth. of instr. Reading 1st gr El. School Good Bad Tr. station Ocean tr. Freight

Doc. 1 Doc. 5 Doc. 3 Doc. 18 Doc. 10 Doc 17 Doc. 6 Doc. 7

In-class exercise: Vocabulary control and hierarchical structure

The following is a list of terms that have occurred in query statements and in document titles. Organize it for purposes of information retrieval.

Book Campaign Candidate Department of State Elections Foreign Office Issue Journal Movement Periodicals Roll-call vote Running for Governor Running for Office State Department Vote	

This task calls on you to apply your knowledge from Soergel, Chapters 12-15. Therefore, no further guidelines are provided. (You may have to do this on your own on a much larger scale, in real life.) Since the list of terms is so small, facet analysis and synthesis is not required in this task.

In-class exercise: Conceptual analysis and synthesis

Organize the following list of terms for purposes of information retrieval.

Terms to work on	**Additional terms** (just to think about)
U.S. Congress	Foreign Office
State Court	British Parliament
County administration	United Nations
State legislature	Prime minister
Federal court	House of Commons
U.S. Senate	House of Lords
U.S. House of Representatives	UN Secretary-General
State administration	UN Security Council
State senate	UN General Assembly
State assembly	World Court

Procedure: Facet analysis and synthesis

Step 1:	Factor concepts into semantic components, resulting in elemental concepts
Step 2:	Organize the resulting elemental concepts in facets
Step 3:	Combine the facets (form all combinations)

The resulting hierarchical structure is to be shown graphically as well as in a linear sequence with cross-reference.

Note: The combinations produced in step 3 show gaps in the original list of terms.

****

Discussion question

Consider the design of an interface to a public-access online catalog in an academic library that would assist users in finding the appropriate LC class number and the appropriate LC subject headings.

Lecture 11.2 ***April 6***

Indexing and system performance (Organizing Information, Chapter 16)

Objectives	1 Understand the concepts of exhaustivity and specificity of indexing and their effect on searching. 2 Understand the concept of weights in indexing. 3 Be able to ascertain the exhaustivity and specificity of indexing in a given system and apply this knowledge to appropriate query formulation. 4 Be able to apply indexing weights in query formulation (including analogous techniques in free-text searching). 5 Be able to determine the proper levels of exhaustivity and specificity of indexing for a new IR system based on user requirements.
Practical significance	An examination of indexing parameters, especially **exhaustivity** and **specificity** and **term weighting**, their measurements, their effect on retrieval performance (which is often oversimplified), their dependence on various factors in the indexing process, and their costs (Chapter 16). A correct understanding of these relationships is important for optimal query formulation in online systems, including the more sophisticated Web search engines, as well as for system design.

Discussion questions	How could one gauge the exhaustivity of indexing in a database if indexers' instructions are not available? How could one tell if within one and the same database exhaustivity varies from subject to subject? Give examples of exhaustivity and specificity of indexing for other types of entities and relationships.

XXX Combine Ch. 16 and logical evidence paper into new Ch. 16

Over

Discussion of Text Organizing Information, Chapter 16 and the reading

Section 16.2.1	**Definitions of exhaustivity and specificity. Indexing weights.** Put in the context of the conceptual data schema of a system. Indexing specificity has to do with the entity values for the entity type subject (or of other entity types, for example **Date**, to which the concept of specificity can be applied). The rules for exhaustivity in indexing are a special case of rules for establishing relationships, such as relationships between a document and subjects. Analogous rules can be defined for many types of relationships. Indexing with weights requires three-place relationships, such as Document *deals with or is relevant for* (Subject, Weight)
16.3.1	**Effects of indexing exhaustivity on retrieval performance** Important conclusion: The query formulation must be adapted to the exhaustivity of indexing for best retrieval results.
Other questions	**Questions on the remainder of the chapter and the reading.**

Discussion questions (repeated)

How could one gauge the exhaustivity of indexing in a database if indexers' instructions are not available? How could one tell if within one and the same database exhaustivity varies from subject to subject?

Give examples of exhaustivity and specificity of indexing for other types of entities and relationships.

April 6, 2011

Free-write 11

Lecture 11.1. Index language structure 2. Database organization

Lecture 11.2. Indexing and system performance

- **Reflect** – what you learned, what was most important, what was most interesting, what was extraneous;
- **Ask questions** – ask for more explanation, how is a concept connected to other concepts, why is a concept important, how can it be applied, why is a reading important;
- Offer **critique and suggestions**;
- Say anything else you want to.

Over

Lecture 12.1 ***April 13***

Discussion and in-class exercise: DDC

Discussion of Assignment 13.1, Dewey Decimal Classification

In-class exercise: Advanced topics in DDC (as marked in the DDC worksheet)

Lecture 12.2 ***April 13***

Short Media Streams demo (to give a taste of Assignment 13.4 Media)

April 13, 2011 Name (optional)

Free-write 12

Lectures 12.1-12.2. DDC and MediaStreams

- **Reflect** – what you learned, what was most important, what was most interesting, what was extraneous;
- **Ask questions** – ask for more explanation, how is a concept connected to other concepts, why is a concept important, how can it be applied, why is a reading important;
- Offer **critique and suggestions**;
- Say anything else you want to.

Over

Lectures 13.1-13.2 ***April 20***

Lecture 13.1A

Questions on Assignment 13.2, ERIC and 13.3, LCSH

Lecture 13.1B

Introductory discussion and in-class exercise on Assignment 13.4 Yahoo

We will start going through the worksheet, index a document, and formulate a query.

We will be online to http://dir.yahoo.com/

Lecture 13.2

Introductory discussion and in-class exercise on Assignment 13.4: LCC

We will start going through the worksheet, index a document, and formulate a query.

April 20, 2011 Name (optional)

Free-write 13

Lectures 13.1-13.2. Yahoo and LCC

- **Reflect** – what you learned, what was most important, what was most interesting, what was extraneous;
- **Ask questions** – ask for more explanation, how is a concept connected to other concepts, why is a concept important, how can it be applied, why is a reading important;
- Offer **critique and suggestions**;
- Say anything else you want to.

Over

Lecture 14.1 ***April 27***

Exploration of classification schemes and thesauri

Discussion of Reading
Be prepared

Lecture 14.2 ***April 27***

Concluding discussion and comparison of classification schemes and thesauri

Objectives Inherit from Lecture 11.2b In addition	1 Get an overview of different types of KOS (Knowledge Organization Systems) and the wide variety of systems in which KOS are used and the wide variety of purposes for which they are used. 2 Solidify general principles underlying all KOS through applying them to the analysis of many quite different examples.
Practical significance Inherit from Lecture 11.2b In addition	• You will be able to "sell your skills" to a wider variety of organizations, increasing opportunities for work • Knowing the general principles that underlie all KOS will enable you to evaluate KOS, to improve existing KOS, and to build new KOS (after taking LIS 514 Indexing and Surrogation).

April 27, 2011 Name (optional)

Free-write 14

Lectures 14.1-14.2. Classification schemes and thesauri

- **Reflect** – what you learned, what was most important, what was most interesting, what was extraneous;
- **Ask questions** – ask for more explanation, how is a concept connected to other concepts, why is a concept important, how can it be applied, why is a reading important;
- Offer **critique and suggestions**;
- Say anything else you want to.

Over

Conclusion

May 4

Lectures 15.1-15.2

Final review

Numbers at left margin indicate number of minutes = number of points.

15 1. There are a wide variety of "documents" on the World Wide Web ("Web pages" and "Web sites"). In a catalog of Web documents, it might be useful to include an indication of the type of document in the catalog record. **Develop a typology of Web documents for this purpose**. (A typology is a list or classification of types).

15 2. **Reorganize thesaurus information to take less reading and less storage space.**

The ERIC Thesaurus has the following entries:

Autoinstructional aids
- RT Audiovisual aids
- RT Computer assisted instruction
- RT Courseware
- RT Individualized instruction
- RT Learner controlled instruction

Programmed instructional materials
- RT Audiovisual aids
- RT Computer assisted instruction
- RT Courseware
- RT Learner controlled instruction
- RT Workbooks

Teaching machines
- RT Computer assisted instruction
- RT Courseware
- RT Learner controlled instruction
- RT Pacing

Another example that could be used for this question is on the next page.

Other example:

free participation
RT health care delivery and administration
RT health care economics

payment-based participation
RT health care delivery and administration
RT health care economics

subsidized payment
RT health care delivery and administration
RT health care economics

full cost-recovery payment
RT health care delivery and administration
RT health care economics

20 3. You have to **design a controlled-vocabulary IR system** (with human indexing) that gives the searcher the option of emphasizing either discrimination (one factor determining precision) or recall. List the features that are important for achieving this flexibility.

4. **Query formulation for free text.** A user needs information on the following topic:

Validity of the evaluation of instructors through undergraduate students in social science courses.

A free-text search for this topic is to be made in a bibliographic database

(1) in database 1, searching is by terms occurring in the **title** of the document,

(2) in database 2, searching is based on terms that occur in the **title and/or the abstract of the document.**

20 a. For each database give the **conceptual query formulation** that you would use (do not worry about terminology at this point). Give your rationale.

10 . b. **Give the free-text query formulation for database 2.** Assume that the search from the previous question is to be made in the system searching on **titles and abstracts** (system 2). Any word or phrase (multi-word term) occurring in the title or abstract can be used as descriptor for searching. Briefly describe how you would go about developing the query formulations in terms of descriptors (3 min.) Start doing it (7 min.)

40 5. You are charged with the design and development of an online information retrieval system for courses at the University of Maryland. The system should serve
(1) students in course selection and
(2) curriculum committees who want to know what courses exist in a given area (such as *statistics* or *communication in organizations* before approving a new one.
Discuss your approach (describe the workings of the system you propose to the extent feasible in 40 minutes; bulleted lists for some pieces are fine)

20 6. This question deals with **retrieval in archives**; sufficient background is provided so that you can answer it even if you are not familiar with archives. Archives are a collection of documents (letters, memoranda, reports, etc.) produced by an organization, its various units, and the persons working in the units. (For this question assume an organization of the complexity of the Federal Government with many organizational units interrelated hierarchically and otherwise.) The organization of archives usually allows for easy retrieval of all documents produced by an organizational unit or by a person; a document is linked to its producer at its creation so that the archivist need not do additional indexing to provide this type of access. Date when created, receiving organizational unit or person, and often related documents are also known for each document. It is usually not feasible to assign subject descriptors to individual documents (too expensive), yet subject searches occur frequently. The archivist doing a subject search uses her - more or less - complete knowledge of organizational units and persons and the subjects they have been dealing with at certain times to find relevant documents; she looks under appropriate organizational units and persons.

Sketch a conceptual data schema for a computerized retrieval system for archives that implements in a formal way the approach described.

40 7. You are appointed as head of a medium-sized IR-system (about 200,000 documents) that uses three different systems for subject access:

(1) an alphabetical subject catalog of books;
(2) shelving books by subject;
(3) an independent classification scheme for filing newspaper clippings

Your analysis shows that the subject heading list and the shelving classification are both far from satisfactory. The subject headings have grown without control and no listing is available. But a cost-benefit analysis rules out major changes or revision, like introducing new schemes, especially in view of the large costs for re-indexing the old collection. On the other hand, the cost-benefit analysis also shows that some costs would be justified to improve the usability of the IR-system. What do you suggest should be done? How would you implement your suggestions?

30 8. Assume you are involved in the **design of a large lexical and classification database** that has the ambitious objective of serving as a tool for natural language processing and as a tool for indexing and retrieval. What information should be included for each term or concept?

40 9. You are given the task to design an IR system. One problem is to determine **how much money should be spent for indexing**. Discuss the data you need/the considerations on which you would base your decision.

40 10. You are given the task of **developing an index language and thesaurus** for
(1) a newly set up information center in a company, or
(2) a public information center in the inner city (choose **one)**.
What are the main points you have to take into consideration in performing this task?

20 11. **Assist users in coping with large Web search results**. A search in a Web directory, such as Yahoo or the Open Directory Project (http://dmoz.org/about.html), or a search engine, such as Google, AltaVista, or Lycos, often returns hundreds of documents. What could the system do to help the user to cope with these large numbers?

15 12. Discuss **exhaustivity** in the context **of hypertext links** made in a system.

15 13. A large subject index is to be put on microfiche. The system has two parts:

(1) The actual index on microfiche. This is an ordinary index: Under each descriptor the entries for the documents (or other retrieval objects) indexed by that descriptor are listed.

(2) To help the user find the appropriate microfiche, there is a hard copy "index to the index." This is simply a list of all descriptors, giving for each the microfiche number and the frame number on the microfiche.

Question: Should the subject index on microfiche be arranged in classified or in alphabetical order? How should the hard-copy "index to the index" be arranged?

Assume a microfiche reader where the user must manually insert the fiche and find the frame..

12 14. Compare a system using shelf arrangement based on an index language like LCC or DDC with a system based on postcombination (such as a computerized IR system) with respect to the exhaustivity and specificity of indexing that can be achieved. What can you say about retrieval performance in both cases?

Final review. Natural language processing (NLP)

Purposes of natural language processing (NLP)

- Preparing a description of the document
  - Descriptive cataloging (e.g. from optically scanned title page)
  - Subject indexing
    - Multiple index terms
    - Assigning a class (from Dewey or LookSmart or Chemical Abstract category), also called document categorization
      - Categorizing a document by reading level(more generally: by the audiences for which the document is appropriate)
  - Abstracting / summarizing
    - Multi-document summaries
- Determining the attitudes, beliefs, or emotions underlying the document (content analysis in sociology and political science or in psychoanalytical methods)
- Determining authorship or other characteristics of the origin of the document
- Preparing a hypertext version of a document, incorporation into a larger hypertext
- Extracting data from a document. Represent the relationships expressed in a document in a more explicit and more easily manipulated way
- Assistance with query formulation
- Natural language interaction with software and systems
- Matching
  - Enhanced proximity searching
- Assistance with document creation
  - Editing assistance
    - Spell check
    - Grammar check
  - Machine translation, for example on-the-fly translation of Web documents
  - Natural language answers from databases, text generation

Natural language processing techniques

- Statistical

  Word frequency, phrase frequency, concept frequency. Frequency of words that connote an attitudinal/emotional dimension (content analysis in psychology/sociology/political science). Differential frequency. Looking for the unexpected (such as weighting rare words highly in ranking retrieval results). Association of words with classes / document categories

- Based on text macrostructure - positional approach

  For example: Introduction and conclusions useful source for abstract. Section headings and figure captions useful source for index terms. First and last paragraphs of sections, first and last sentences of paragraphs

- Cue words, phrases, and sentences

  "method", "important result", "new"

- Stemming and other morphological normalization

- Syntactic and semantic analysis

  Parsing of sentences or partial parsing to detect noun phrases

  Parsing with semantic interpretation

  Homonym disambiguation (Subject area of document or Disambiguation rules based on semantic rules (such as *laugh* takes only animate subjects)

  Inter-sentence parsing, resolution of anaphoric references

- Slot filling in frames using parsing or cues

A technique used equally by human readers and by machine systems.

Converting natural language statements into entity-relationship expressions. Applying verb case frames. Using cue words to discover type of relationship between two entities, such as *because* or *therefore* indicating causation (See Crombie example in Lecture 6.1).

Knowledge required by NLP

Final review. Precombination vs postcombination

1 Precombination vs postcombination in searching

1.1 Basic problem: Most searches are for topics or themes expressed as compound concepts, such as

the effect of alcohol on the liver or

how to improve test scores of minority children.

In a retrieval system that allows for combining descriptors (as an online system allowing for Boolean query formulations) an index language consisting only of elemental descriptors is sufficient: the user can combine the elemental descriptors that make up her search topic.

But in a retrieval system that allows only searches for single descriptors, as in a card catalog, printed index, shelve arrangement, or a Web subject directory (Yahoo, LookSmart, etc.) the system must provide precombined descriptors for the topics users want to search. The user who wants to find materials on a topic for which there is no precombined descriptor will have difficulty.

Additional reasons for introducing precombined descriptors even in a system that is mainly based on postcombination

1.2 With postcombination, the components of the query formulation may not have the right relationship in the document

Example: A search for

Air transport AND Vehicles

finds a document on

Vehicles used on the metro line to the airport

Need descriptor *Aircraft*

1.3 With postcombination, a combination of elemental descriptors might be ambiguous

Examples:

School	AND	Library	Need descriptors *School library* and *Library School*
Personnel	AND	Administration	Need descriptors *Personnel administration* and *Administrative personnel*

1.4 Requiring the user to combine elemental descriptors may be unnatural

2 Precombination vs postcombination in database organization

Documents are about topics/themes and can be usefully grouped by topics/themes

Examples: Shelf arrangement, Web subject directory, organizing Web search results (often hundreds or thousands of items) into meaningful groups. Need precombined descriptors that define groups (classes).

Problem of arranging precombined descriptors / classes in a meaningful order

Other aspect of same problem: If document can be assigned only one descriptor, that descriptor should express as many of the document concepts as possible; it needs to be precombined.

Relationship top semantic networks; precombined descriptor as an abstraction of what is in common to a group of documents.

3 User problem with systems using a large number of precombined descriptors:

Finding all precombined descriptors under which to search (because of extensive polyhierarchy in a large set of precombined descriptors, one search often requires looking under many, as the DDC and LCC assignments demonstrated)

Solution: descriptor-find index

May 4, 2011 Name (optional)

Free-write 15

Lectures 15.1-15.2. Final review

- **Reflect** – what you learned, what was most important, what was most interesting, what was extraneous;
- **Ask questions** – ask for more explanation, how is a concept connected to other concepts, why is a concept important, how can it be applied, why is a reading important;
- Offer **critique and suggestions**;
- Say anything else you want to.

Over