

# Inf 722 Information Organization (Fall 2011)

PhD Program in Information Science, CCI, SUNY Albany

September 19, 2011

## Table of Contents

- ▶ Knowledge, Intuition, and Information Sharing
- ▶

*I have often pondered over the roles of knowledge or experience, on the one hand, and imagination or intuition, on the other, in the process of discovery. I believe that there is a certain fundamental conflict between the two, and knowledge, by advocating caution, tends to inhibit the flight of imagination. Therefore, a certain naivete, unburdened by conventional wisdom, can sometimes be a positive asset.*

R. Langlands, in **Harish-Chandra**  
Biographical Memoirs of Fellows of the Royal Society  
31 (1985) 197 - 225.

The following slides are from my old notes. I have included them for completeness.

# Objectives

- ▶ Development of lexical resources such as dictionaries and thesauri
- ▶ Exploration of textual data to support analysis of its properties
- ▶ Indexing of text to retrieve information to answer queries
- ▶ Architecture of Search Engines

## Dictionaries

Dictionaries can be built based on either of the following:

- ▶ historical principles (as in case of older versions of the Oxford English dictionary: <http://www.oed.com/>)
  - ▶ usually prescriptive
  - ▶ all forms of usage may not be covered
  - ▶ may have biases based on the socio-economic conditions of the compilers
- ▶ based on the analysis of the text in a corpus (as in the case of the Collins Cobuild Dictionaries:  
<http://www.collins.co.uk/books.aspx?group=140>)
  - ▶ usually descriptive
  - ▶ biases can be reduced by choosing a corpus that is representative of the usage

Dictionaries are usually based on the written word, but they do not have to be.

## Thesauri

Thesauri contain relationships between words

- ▶ For nouns and adjectives:  
Antonyms & synonyms
- ▶ For verbs:  
Troponyms: walk – amble, shuffle, file, race, roam, meander, stroll, strut, stump, tread, trudge, foot (informal), leg (informal)

**Some good sites:** [www.dictionary.com](http://www.dictionary.com), [www.visualthesaurus.com](http://www.visualthesaurus.com)

## Thesauri

Other Lexical relations include:

### ▶ Nouns

- ▶ **hypernyms:** Y is a hypernym of X if every X is a (kind of) Y (canine is a hypernym of dog)
- ▶ **hyponyms:** Y is a hyponym of X if every Y is a (kind of) X (dog is a hyponym of canine)
- ▶ **coordinate terms:** Y is a coordinate term of X if X and Y share a hypernym (wolf is a coordinate term of dog, and dog is a coordinate term of wolf)
- ▶ **holonym:** Y is a holonym of X if X is a part of Y (building is a holonym of window)
- ▶ **meronym:** Y is a meronym of X if Y is a part of X (window is a meronym of building)

## Thesauri

Other Lexical relations include:

### ▶ Verbs

- ▶ **hypernym:** the verb Y is a hypernym of the verb X if the activity X is a (kind of) Y (to perceive is an hypernym of to listen)
- ▶ **troponym:** the verb Y is a troponym of the verb X if the activity Y is doing X in some manner (to lisp is a troponym of to talk)
- ▶ **entailment:** the verb Y is entailed by X if by doing X you must be doing Y (to sleep is entailed by to snore)
- ▶ **coordinate terms:** those verbs sharing a common hypernym (to lisp and to yell)

## Thesauri

Other Lexical relations include:

### ▶ Adjectives

- ▶ **related nouns:**
- ▶ **similar to:** Hard – firm, solid; adamantine; woody; stony, granitic, granitelike, rocklike; unyielding; stonelike, petrous; hardened, case-hardened, hardboiled; steely; calculative, calculating, conniving, scheming, shrewd;
- ▶ **participle of verb:** *rolling* stone; *smiling* crocodile; *written* apology;

### ▶ Adverbs

- ▶ **root adjectives:** *extremely* interesting; *faster* walk;

Based on Wordnet in the Wikipedia  
(<http://en.wikipedia.org/wiki/WordNet>)

*You shall know a word by the company it keeps*

*John Rupert Firth. 1957:11  
Papers in Linguistics 1934-1951 (1957)  
London: Oxford University Press.*

# Exploration of Textual data

Why?:

- ▶ Voluminous data. Medline alone has well over 15 million records
- ▶ Traditional literature mining impossible
- ▶ Manual digital curation (preservation, maintenance) and provenance management very costly
- ▶ Information retrieval methods are often inadequate for aiding knowledge discovery

# Exploration of textual data

- ▶ Construction of corpus
- ▶ Markup of documents in the corpus
- ▶ Text pre-processing
- ▶ Concordances
- ▶ Interpretation

**Source:** Developing Linguistic Corpora: a Guide to Good Practice  
John Sinclair, Tuscan Word Centre  
<http://ahds.ac.uk/guides/linguistic-corpora/index.htm>

**Definition:** A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.

## Principles:

- ▶ The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise.
- ▶ Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen.

# Corpus Construction

- ▶ Only those components of corpora which have been designed to be independently contrastive should be contrasted.  
(Normative, historical, monitor, varietal corpora, specialised corpora)
- ▶ Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination.
- ▶ Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications.
- ▶ Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible. This means that samples will differ substantially in size.

- ▶ The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken.
- ▶ The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components.
- ▶ Any control of subject matter in a corpus should be imposed by the use of external, and not internal, criteria.

# Markup of Documents in the corpus

- ▶ Part of speech tagging
- ▶ Sense tagging
- ▶ Semantic tagging
- ▶

Supervised and unsupervised algorithms are used for word-sense disambiguations for sense tagging. For details, see: *Sense Tagging: Semantic Tagging with a Lexicon* by Yorick Wilks and Mark Stevenson,  
<http://www.aclweb.org/anthology-new/W/W97/W97-0208.pdf>

# Text Preprocessing

- ▶ Tokenization
- ▶ Expansion of abbreviations and acronyms to their canonical orthographic representations
- ▶ Sentence boundary detection
- ▶ Morphological analysis/ Lemmatization

# Concordances

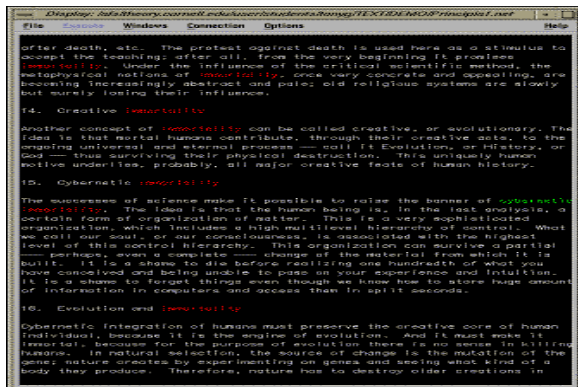
*An alphabetical arrangement of the principal words contained in a book, with citations of the passages in which they occur*  
Oxford English Dictionary

Headword	No.	Context...	Word	...Context	Reference
HEAR	15	That my own	heart	drifts and cries, having no...	Deep Analysis
HEARD	9	By the shout of the	heart	continually at work	And the wave
HEARING	7	Nothing to adapt the skill of the	heart	to, skill	And the wave
HEARS	3	The tread, the beat of it, it is my own	heart	,	Träumerei
HEARSE	1	Because I follow it to my own	heart		Many famous
HEART	25	My	heart	is ticking like the sun:	I am washed t
HEART'S	2	The vague	heart	sharpened to a candid co...	The March Pa:
HEART-SHAPED	1	Contract my	heart	by looking out of date.	Lines on a Yo
HEARTH	1	Having no	heart	to put aside the theft	Home is so Se
HEARTS	7	And the boy puking his	heart	out in the Gents	Essential Bea
HEARTY	1	A harbour for the	heart	against distress.	Bridge for the
HEAT	6	These I would choose my	heart	to lead	After-Dinner F
HEAT-HAZE	1	Time in his little cinema of the	heart		Time and Spax
HEATH	1	This petrified	heart	has taken,	A Stone Churr
HEATS	1	How should they sweep the girl clean...	heart	,	I see a girl dra
HEAVE	1	Hands that the	heart	can govern	Heaviest of fic
HEAVEN	4	For the	heart	to be loveless, and as col...	Dawn
HEAVEN-HOLDING	1	With the unguessed-at	heart	riding	One man walk
HEAVIER-THAN...	1	If hands could free you,	heart	,	If hands could
HEAVIEST	2	That overflows the	heart		Pour away the

Words: 7318 Tokens: 37070 At word: 2990 Deleted lines: 1 [24] Word sort: Asc alpha (string) Context sort: Asc occurrence order

Source: <http://www.concordancesoftware.co.uk/>

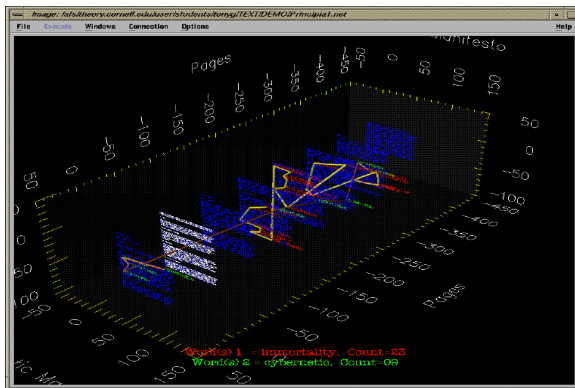
# Text Visualisation



Source:

<http://www.nbb.cornell.edu/neurobio/land/OldStudentProjects/cs490-95to96/tonyg/Language.Viz1.html>

# Text Visualisation



## Source:

<http://www.nbb.cornell.edu/neurobio/land/OldStudentProjects/cs490-95to96/tonyg/Language.Viz1.html>

Continuation

*Note:* This slide is based on the Wikipedia article on the subject

- ▶ The Semantic Web is a "man-made woven web of data" that facilitates machines to understand the semantics, or meaning, of information on the World Wide Web (From *Wikipedia*)
- ▶ Based on the concept of the Semantic Network Model was coined in the early sixties by the cognitive scientist Allan M. Collins, linguist M. Ross Quillian and psychologist Elizabeth F. Loftus to represent semantically structured knowledge.
- ▶ Machine-readable metadata about pages and how they are related to each other are inserted in the documents thus enabling automated agents to access the Web more intelligently and perform tasks on behalf of users
- ▶ Objective: to find, share, and combine information easily

# Semantic Web Stack 2000

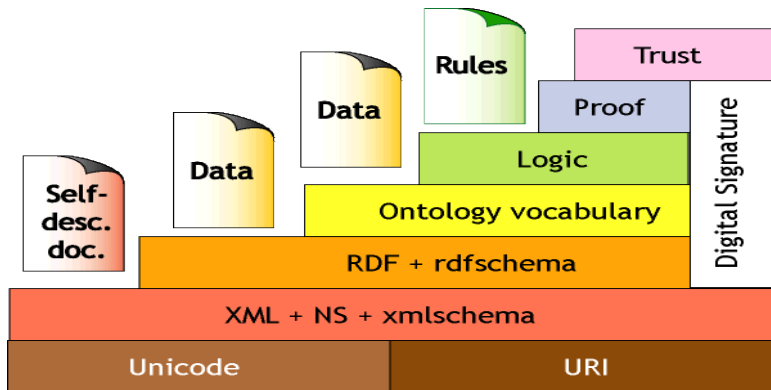


Figure: Semantic Web Stack

## Source:

<http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html>

# Semantic Web Stack 2005

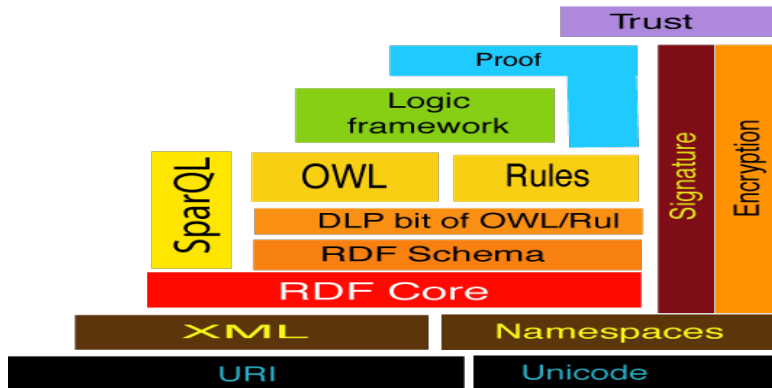


Figure: Semantic Web Stack

## Source:

<http://www.w3.org/DesignIssues/diagrams/sw-stack-2005.png>

# Resource Description Framework

- ▶ XML and RDF: Motivation
- ▶ RDF Vocabulary
- ▶ Tips on RDF Modeling & RDF Syntax
- ▶ Topics in RDF
- ▶ RDF Schema
- ▶ Drawbacks of RDF

# XML and RDF: Motivation

- ▶ Current web is document-centric
- ▶ Documents are hierarchical in nature and best described by tree data structures
- ▶ Tree data structures are well understood, and their manipulation on computers can be very efficient
- ▶ However, tree data structures are not ideally suited to representing semantics. For example, if data is distributed across many documents, it is difficult to fuse trees in different documents into one tree that represents the collective meaning of such document fragments

# XML and RDF: Motivation

- ▶ Description of the meaning of data on the web need not be hierarchical. In fact, much of the data on the web is best described by directed graphs
- ▶ It is easier to fuse such directed graphs distributed across the web into one directed graph that coherently describes the collective meaning of such distributed fragments. This is important since often we need to borrow vocabularies from other domains to construct our own things
- ▶ It is important to choose a data structure that makes the task of manipulation easier and efficient

# XML and RDF: Motivation

- ▶ RDF adopts the directed graph data structure to represent meaning.
- ▶ Directed graph data structure permits a graph to be composed of unconnected components. A tree structure, on the other hand, requires that the tree be connected
- ▶ "Graphs in RDF are . . . better suited for the composition of distributed information sources" Hitzler, Krötzsch, & Rudolph (2010)

- ▶ Literals: Data values
- ▶ URI: *scheme:[//authority] path [?query] [#fragment]*
- ▶ Directed Graph consisting of
  - ▶ Ovals: URIs
  - ▶ Arrows: URIs
  - ▶ Rectangles: Literals

# Tips on RDF Modeling and RDF Syntax

- ▶ No direct statements about literals. So, no directed edges (arrows) originating from literals (rectangles)
- ▶ Literals can not be used as labels for directed edges (arrows)
- ▶ Easier to first model the graph and then represent it in RDF syntax
- ▶ Representation of graphs as triples
  - ▶ **Subject**
  - ▶ **Predicate**
  - ▶ **Object**
  - ▶ Only binary predicates can be used in RDF. It is always possible to reduce an n-ary predicates in to a set of binary predicates as done in data modeling for databases

- ▶ Data Types: XML Schema data types (<http://www.w3c.org/2001/XMLSchema#string>) are used for typing literals, but any externally defined data types can be used
- ▶ RDF document consists of an `rdf:RDF` element, the content of which is a number of descriptions *Example:*

```
< rdf : Descriptionrdf : about = "949318" >
```

```
< uni : name > DavidBillington < /uni : name >
```

```
< uni : agerdf : datatype = "&xsd; integer" > 27 < uni :  
age >
```

```
< /rdf : Description >
```

- ▶ In RDF external namespaces are expected to be RDF documents defining resources used to import RDF documents
- ▶ It is possible to nest descriptions
- ▶ It is possible to type descriptions using `rdf:type`
- ▶ Use of container elements such as `bag` (unordered container with multiple occurrences), `seq` (ordered container), `Alt` (set of alternatives).

- ▶ RDF makes no assumptions about domains. Therefore there is a need to explicitly define classes, hierarchies, and inheritance
- ▶ RDF Schema provides the following core classes: Resource, Class, Literal, Property, Statement. It also provides core properties for defining
  - relationships (type, subclassOf, subPropertyOf),
  - restricting properties (domain, range),
  - for re-ification (subject, predicate, object),
  - container classes (Bag, Seq, Alt), and
  - utility properties (seeAlso, isDefinedBy, comment, label)

# Drawbacks of RDF

- ▶ Limited expressivity because predicates have to be binary and only limited to class and property hierarchies and domain/range definitions
- ▶ Can not express restrictions on properties (eg., cows eat only grass, ... )
- ▶ Can not express disjointedness of classes (eg., male and female are disjoint classes)
- ▶ Not possible to build new classes by combining existing classes using union, intersection, complement
- ▶ Cardinality constraints can not be specified

... ontology is a *data model* that represents a set of *concepts* within a *domain* and the *relationships* between those concepts. It is used to *reason* about the *objects* within that domain. – Wikipedia

- ▶ Representation of shared conceptualisations
- ▶ Representation of knowledge regarding a domain, or a part of the world

- ▶ Quines ontological commitment: To be is to be the value of a variable
- ▶ Ontological reduction: The most economical ontology for a purpose
- ▶ Criteria of identity: "No entity without identity"
- ▶ "On the Ontological Remarks on the Propositional Calculus", "A Logical Approach to the Ontological Problem", and "On What There Is" by Quine

"We may be said to countenance such and such an entity if and only if we regard the range of our variables as including such an entity. To be is to be a value of a variable."

W.v Orman Quine

"What entities there are, from the point of view of a given language, depends on what opositions are accessible to variables in that language. There is one important sense, however, in which the ontological question transcends linguistic convention: How economical an ontology can we achieve and still have a language adequate to all purposes of science? In this form the question of ontological presuppositions of science survives."

W.v Orman Quine

# Why Ontology”

- ▶ To share common understanding of the structure of information among people or software agents
- ▶ To enable reuse of domain knowledge
- ▶ To make domain assumptions explicit
- ▶ To separate domain knowledge from the operational knowledge
- ▶ To analyze domain knowledge

- ▶ Knowledge Representation:
  - ▶ Semantic networks (directed graphs with concepts as nodes and relationships as arrows)
  - ▶ Frames (A frame is a collection of attributes or slots and associated values that describe some real world entity)
- ▶ Predicate/Propositional Logic

# Requirements for Ontology Languages

- ▶ Sufficient expressive power
- ▶ Well-defined syntax
- ▶ Formal semantics
- ▶ Efficient reasoning support
- ▶ Convenience of expression

# Requirements for Ontology Languages

- ▶ **Sufficient Expressive Power**
  - Should be sufficiently rich so that all information regarding the ontology domain can be expressed in the language
- ▶ **Well-defined Syntax**
  - Necessary condition for machine processing
- ▶ **Formal Semantics**
  - The meaning of what is expressed in the ontology must be clear
  - Enables reasoning about properties such as
    - ▶ Class Membership
    - ▶ Equivalence of Classes
    - ▶ Consistency
    - ▶ Classification
- ▶ **Efficient Reasoning Support**
  - ▶ Checking consistency
  - ▶ Checking for unintended relationships between classes
  - ▶ Automatic classification of items in classes

# Owl Sublanguages

- ▶ OWL Full:
  - Uses all OWL language primitives
  - Fully upward compatible with RDF
  - Undecidable, and so no hope of efficient reasoning support
- ▶ OWL DL:
  - Sublanguage of OWL
  - Restricts how OWL constructors can be used (OWL provides constructors to define class expressions)
  - Permits efficient reasoning support
  - Every legal OWL DL document is a legal RDF document, but not the other way around
- ▶ OWL Lite:
  - Excludes enumerated classes, disjointness statements, arbitrary cardinality, and so limited expressivity
  - Easier for users to understand

- ▶ Upward compatibility between OWL Lite, OWL DL and OWL Full
  - Legality of ontologies
  - Conclusions

# OWL Language Description

- ▶ Header
  - RDF element that specifies namespaces (rdf:RDF)
  - Housekeeping assertions element (owl:Ontology)
    - ▶ Comments
    - ▶ Versions
    - ▶ Imports
    - ▶ Label
- ▶ Class Elements (owl:class)
- ▶ Property Element
  - DatatypeProperty (owl:DatatypeProperty)
  - Object Property (owl:ObjectProperty)
- ▶ Property Restrictions
  - rdfs:subClassOf
  - owl:Restriction
  - .....
- ▶ Special Properties
- ▶ Boolean Combinations (union, intersection, complement)

# Ontologies: Current Drawbacks

- ▶ No support for modules in ontologies
- ▶ No support for defaults
- ▶ Assumption of open world (most databases assume closed world)
- ▶ Unique names assumption (most databases assume that individuals with different names are different)
- ▶ No support for procedural attachment