Illuminating Chaos
Using Semantics to Harness the Web

Dagobert Soergel

Department of Library and Information Studies,
University at Buffalo

UDC Seminar, The Hague, October 28-29, 2009

Outline

- **Overview of issues**
  - Semantics for whom and for what
  - Representation to assist with query formulation
  - Representation for comprehension
  - Systems of representation
  - Support for finding: Indexing
  - Building KOS
  - How can it all get done
- **Zeroing in on the conceptual foundation**

Semantics, structure, meaning

- **Classification**
- **Meaningful arrangement**
- **All kinds of relationships**

Semantics for whom?

- **Semantics for computer systems**
  inference
  answers and solutions instead of lots of Web pages

- **Semantics for people**
  assist users in creating meaning and making sense
  structure for learning

Semantics for what

- Finding
- Comprehending
- To know what to look for, a user (a person or a system) must first comprehend something – a cycle
- Both finding and comprehending require navigating in an information space – need meaningful structure

Representation
to assist with query formulation

Problem clarification for search

JG	prevention approach
JG10	. **individual-level prevention**
JG10.2	. . individual- vs. family-focused prevention
JG10.2.2	. . . individual-focused prevention
JG10.2.4	. . . family-focused prevention
JG10.4	. . prevention through information and education
JG10.4.2	. . . social marketing prevention approach
JG10.4.4	. . . prevention through information dissemination
JG10.4.6	. . . prevention through education
JG10.4.8	. . . peer prevention
JG10.8	. . prevention through spirituality and religion
JG10.10	. . prevention through public commitment
JG12	. **environmental-level prevention**
JG12.4	. . social policy prevention approach
JG14	. **multi-level prevention**

Browse structure for search

- Make a table of contents for the entire Wikipedia using UDC

Facet structure to guide search

A Area of ability combines with **B Degree of ability**

A1 psychomotor ability A2 senses A2.1 . vision A2.1.1 . . night vision A2.2 . hearing A3 intelligence A4 artistic ability	B1 low degree of ability, disabled B2 average degree of ability B3 above average degree of ability B3.1 . very high degree of ability
Examples	A2.1B1 visually impaired A2.2B1 hearing impaired A3B1 mentally handicapped A3B3 intellectually gifted

Provide front-ends to assist users

- Elicit a query with a facet-based interfaces,
  then the system creates a free-text query
- Create a structure that normalizes terms assigned through social tagging and arranges them in a meaningful structure.
  The user can than browse and select concepts
  The system maps to all appropriate tags

Problem space for diseases

Used by people or computer systems
for search and arranging search output

Pathologic process

Body system affected

Cause (condition, organism, chemical substance, environmental factors)

Treatment

Representation for comprehension

A question of information representation (knowledge representation)

- For computer systems: formal representation
- For people: Text, images, graphical representation, visualization
- Transformations between representations, such as
  - from text to formal: information extraction
  - from text to a map showing the text structure
  - from a conventional thesaurus display to a concept map

Two representations

Text (for people)

High blood pressure is a serious disease often caused by being overweight. In kids 4 – 12 it can be treated highly effectively with Nystatin.

Formal representation (for computer system)

Causation (HighBloodPressure, Obesity)

Treatment (HighBloodPressure, {Human, [Age, 4-12y]}, Nystatin, [Effectiveness, 4])

Answering questions

Question

How can high blood pressure be prevented?

Answer

Loose weight?

Two representations

Text

Kids begin grazing independently from their mothers at three months

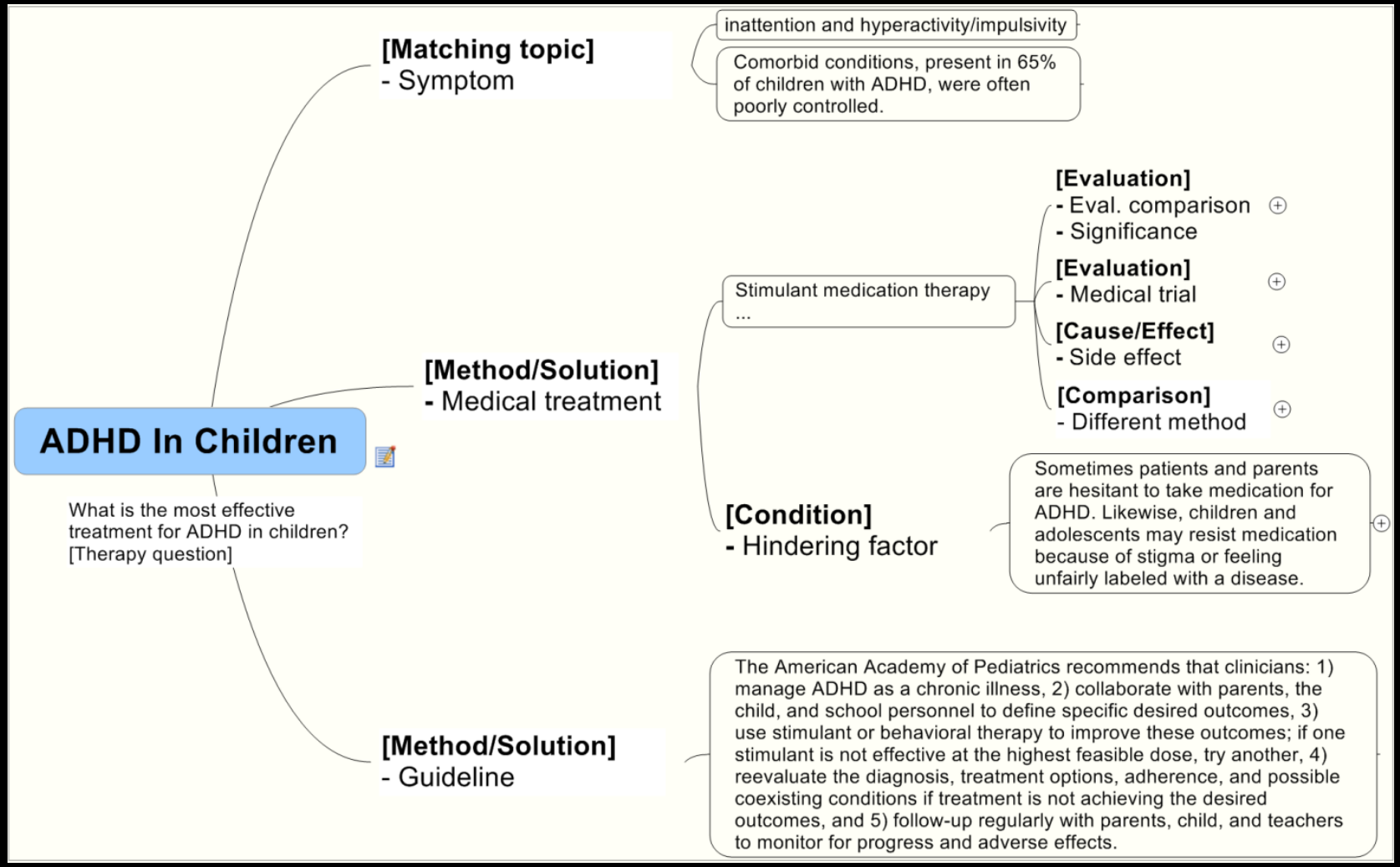
Formal representation

Separation (Mother, Child, {Goat, [Age, 3m]})

Information extraction

- Information extraction produces representations needed for the semantic Web
- Also useful for people if formal expressions are transformed into sentences that state the findings of a document as individual "bullets"
- Could arrange statements from one or more documents in UDC order as a kind of summary
- Information extraction needs rich KOS

Representation of text structure



Meaningful arrangement of terms assigned in social tagging

The Martyrdom of Saint Bartholomew

Matching topic (Direct)

- **Image content: Focal**
  - *Reference*
    - nude body
    - old man
    - Saint Bartholomew
    - executioner
    - knife
  - *Elaboration (Adj.)*
    - Bearded
    - physical anguish
    - profound emotion {emotional}
    - luminous
  - *Elaboration (Adv.)*
    - expressive hands
    - gestures
    - confronts
    - flayed alive
    - torture
- **Image theme**
  - martyrdom
  - mystical experience
  - biblical
  - religious
- **Image content: Peripheral**
  - *Elaboration (Adv.)*
    - lurking

Comparison

- **By similarity: Metaphor / analogy**
  - Christ's sacrifice and crucifixion {Christ metaphor}

Context

- **Biographic info: Artist**
  - Jusepe de Ribera
- **Biographic info: Time / period**
  - 1634
  - 17th century

Cause / Effect

- **Reaction or feeling**
  - Intensity
- **Effect / Outcome**
  - Pulls the viewer into the scene

Support comprehension through links to KOS

- Map text term to concept in KOS,
  show definition,
  show place in hierarchical structure

Comprehension "in the large"

- Learning and sense making require comprehension across multiple sources
- Requires structure – can be supplied by KOS
- Require tools for the manipulation of external structures the learner / sensemaker builds

Representation systems

Representations need rules

- Formal representations need logical formalisms, such as full first-order logic or subsets (for ease of processing) or extensions (to be more expressive)
- Text needs rules of syntax and broader document structure
- Graphical representations need rules of design

Representations need names for entities

- Names for (abstract) concepts – classification
- Names for many different types of other entities, such as persons, places, buildings, events, currencies, … (named entities)
- Systems of such names – Knowledge Organization Systems, authority lists of personal names
- Mappings between such systems

Representations need relationships

- Relationships are used to connect entities, thus forming statements
  obesity <*causes*> high blood pressure
- Need system of relationships
  Many such systems exist (a type of KOS)
  Problem of mapping

Rhetorical relationships

- To map text structure
- To discern how a retrieved document, paragraph, statement, or image relates to the topic of a search

Topical relevance typology

Function-based

Rhetorical structure

- Matching topic
- Evidence (Indirect)
- Context
- Comparison
- Evaluation
- Method / Solution
- Purpose/ Goal

Argument structure

- Grounds
- Warrants
- Claim

Reasoning-based

- Generic inference
- Comparison-based
- Induction / rule-based
- Causal-based
- Transitivity-based

Semantic-based

(Green & Bean, 1995)

- Taxonomy
- Partonomy
- Frame-based,
- etc.

RST+ Functional Role

Matching topic (Direct)
- Manifestation
- Image content
- Image theme

Evidence (Indirect)

Context
- Scope
- Framework
- Environmental setting
- Social background
- Time & sequence
- Assumption / expectation
- Biographic information

Condition
- Helping or hindering factor
- Unconditional
- Exceptional condition

Purpose / Motivation

Cause / Effect
- Cause
- Effect / Outcome
- Explanation (causal)
- Prediction

Comparison
- By similarity (analogy) / By difference (contrast)
- By factor that is different

Method / Solution
- Method / Approach
- Instrument
- Technique / Style

Evaluation
- Significance
- Limitation
- Criterion / Standard
- Comparative evaluation

Functional role: Comparison

Comparison

. By similarity vs. By difference (Contrast)
. . By similarity
. . . Analogy & metaphor
. . By difference (Contrast)
. By factor that is different
. . Different external factor
. . . Different time
. . . Different place
. . Different participant
. . . Different actor
. . . Different subject acted upon
. . Different act or experience
. . . Different act
. . . Different experience

Support for finding: Indexing

- **Finding based on text:**
  Knowledge-based expansion of query
  Front-end as discussed earlier

- **Finding based on indexing:**
  Semantically enriched documents

A semantically enriched document

Reis et al. (2008)
Impact of Environment and Social Gradient on Leptospira infection in Urban Slums (doi:10.1371/journal.pntd.0000228).

Infectious disease studied:	Leptospirosis
Pathogen (causative agent of disease):	Leptospira spirochete
Vector of disease pathogen:	Rat (Rattus norvegicus)
Pathogen host subjected to study:	Human (Homo sapiens)
Number of subject individuals in study:	3,171
. . .	
Purpose of study:	Quantify risk factors for leptospirosis . . .
Principal finding 1:	Prevalence of Leptospira antibodies . . .
Principal finding 2:	Disease risk . . .open sewers . . .

(http://dx.doi.org/10.1371/journal.pntd.0000228.x002)

A semantically enriched document

Tag Trees of Individual Semantic Classes of Highlighted Terms	ID = Infectious Disease Ontology GO = Gene Ontology term used in ID
disease infectious diseases diarrheal disease childhood diarrhea dengue **leptospirosis** human leptospirosis meningococcal disease pulmonary hemorrhage syndrome visceral leishmaniasis Weil's disease occupational disease zoonotic disease	ID:0000012 immunity ID:0000017 mortality ID:0000023 zoonotic ID:0000025 pathogenicity ID:0000034 endemic ID:0000038 parasite ID:0000056 host ID:0000057 carrier ID:0000063 vector ID:0000064 pathogen ID:0000066 infectious agent ID:0000069 primary pathogen ID:0000104 infection

ID = Infectious Disease Ontology GO = Gene Ontology

IDO:0000000 ! process
- IDO:0000083 transmission
  - IDO:0000231 horizontal transmission (GO:0000031)
- IDO:0000104 infection
- IDO:0000084 pathogenesis
- IDO:0000221 ! infectious disease progression
  - IDO:0000100 ! pathogen evasion of host immune response
    - IDO:0000111 antigenic variation
      - IDO:0000115 genetic diversification
- IDO:0000226 pathogen life cycle (GO:0000026)

IDO:0000001 ! role
- IDO:0000036 ! colonizer
  - IDO:0000038 parasite
- IDO:0000048 symptom
- IDO:0000056 host
  - IDO:0000057 carrier
- IDO:0000059 reservoir
- IDO:0000063 vector
- IDO:0000064 pathogen
  - IDO:0000066 infectious agent
    - IDO:0000069 primary pathogen
- IDO:0000200 mode of transmission (GO:0000000)

IDO:0000002 ! quality
- IDO:0000215 ! quality of host population
- IDO:0000098 infectious disease

Semantically enriched documents

- **Semantic enrichment supports semantic retrieval**
  - Broad area of its own
- **Many different forms**
  - Explicit document structure
  - Concept and named entity tagging and identification
  - Assigning additional concepts or named entities
  - Assigning extracted propositions
- **Closely linked with information extraction**
  - IE produces elements of semantic enrichment

Need KOS

Needed for all this

- Large Knowledge Organization Systems
- Large knowledge bases with mappings
- Methods and procedures for developing KOS

How to get all this work done?

The forces that created the problem
also support the solution

- Use automation
  - Automated information extraction gets better every day and also provides input to building KOS
  - Automated classification could be used for the UDC Wikipedia project
- Use Web-enabled collaborative work ("crowdsourcing")
- Use computer systems to assist people
- Use Web-based systems to collect and integrate results
- Bootstrap: The more knowledge is in formal systems, the more information extraction and structuring tasks can be automated

Example: Guided tagging

- Use facet structure to get taggers think a bit more out of the box
  For example, could ask
  What does this image remind you of
- Could assign some terms automatically, for example, extracting terms from text assigned to an image

Salvador Dali, 1904-1989
Galatea of the Spheres, 1952
oil on canvas
Teatre Museu Dali

Comment

Comment on this image

View All

View all comments

Instructions

Start tagging by entering terms into the General box. Press enter after each entr
tags will be added automatically to the categories below. If you feel a tag belong
different category, simply drag and drop it into the correct box. You may also ap
entering them directly into the category boxes. See the Help section for more in
on applying tags.

General

surealism
peaceful
blue

Image Content

woman
blue

Style

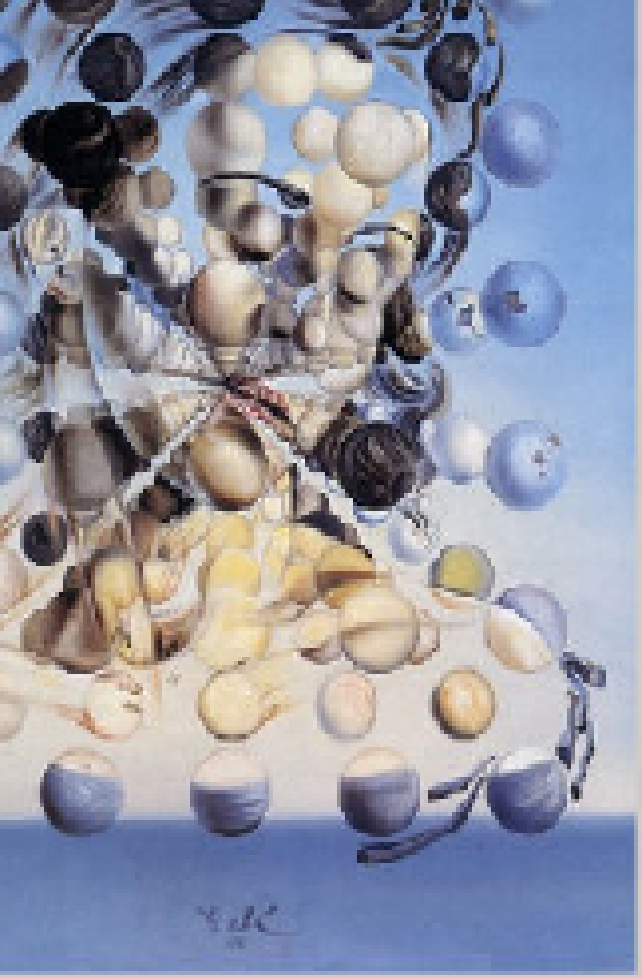
surrealism

surrealism

Facet

Interpretation

peaceful



Dali, 1904-1989
f the Spheres, 1952
vas
seu Dali

t

tags will be added automatically to the categories below. If you feel a ta
different category, simply drag and drop it into the correct box. You may
entering them directly into the category boxes. See the Help section for
on applying tags.

General

surealism
peaceful
blue

Disambiguation and Spelling Pop-up

Select 1. (n) **blue**, blueness (blue color or pigment; resembli
color of the clear sky in the daytime) "he had eyes of bl
blue"

Select 2. (adj) **blue** (used to signify the Union forces in the Am
Civil War (who wore blue uniforms)) "a ragged blue line

Select 3. (adj) gloomy, grim, **blue**, depressed, dispirited, dow
downcast, downhearted, down in the mouth, low, low-s
(filled with melancholy and despondency) "gloomy at t
thought of what he had to face"; "gloomy predictions";

Image C

woman

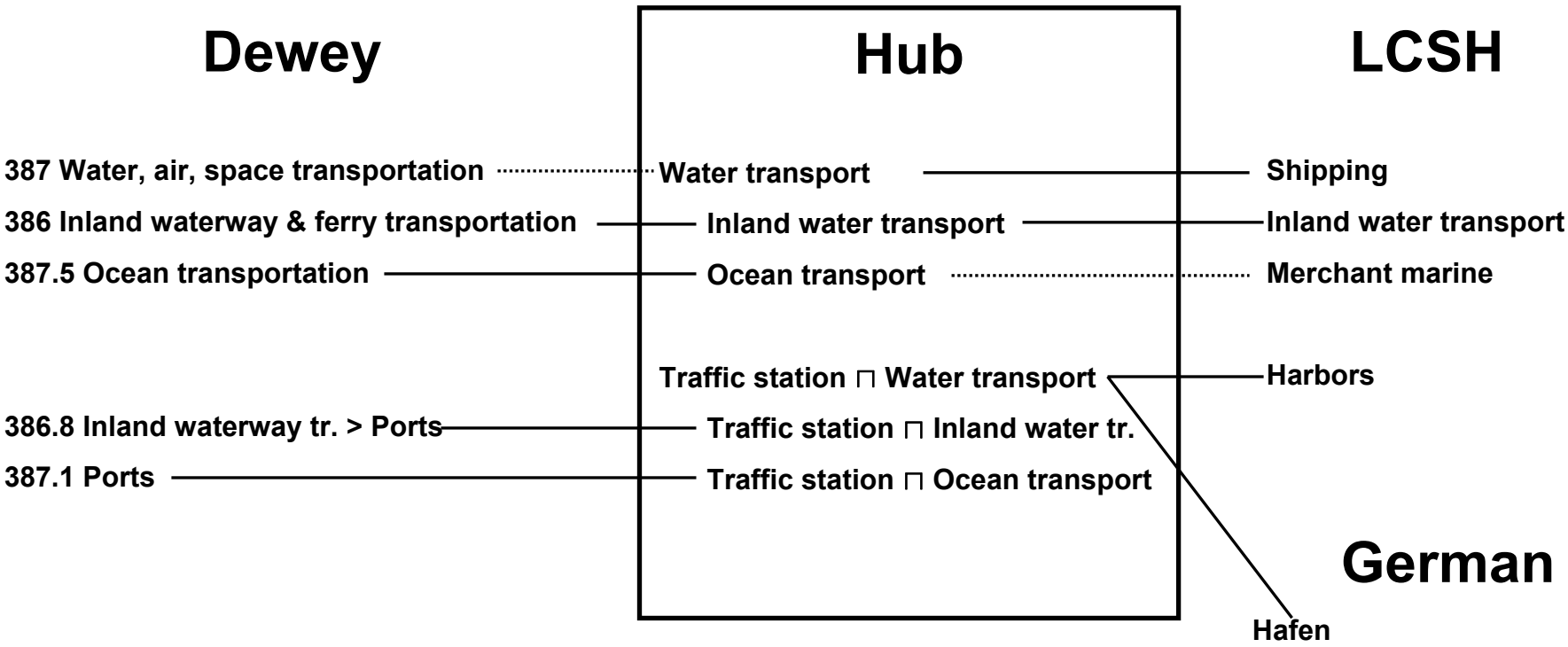
Style

surrealis

Facet

Semantic analysis as the basis for everything

Mapping through a Hub



Outline

- Objective: Interoperability Plus
- KOS concept hub
- Method: Knowledge-based, computer-assisted creation of canonical representations of concepts
- Resulting knowledge base and applications

Objective

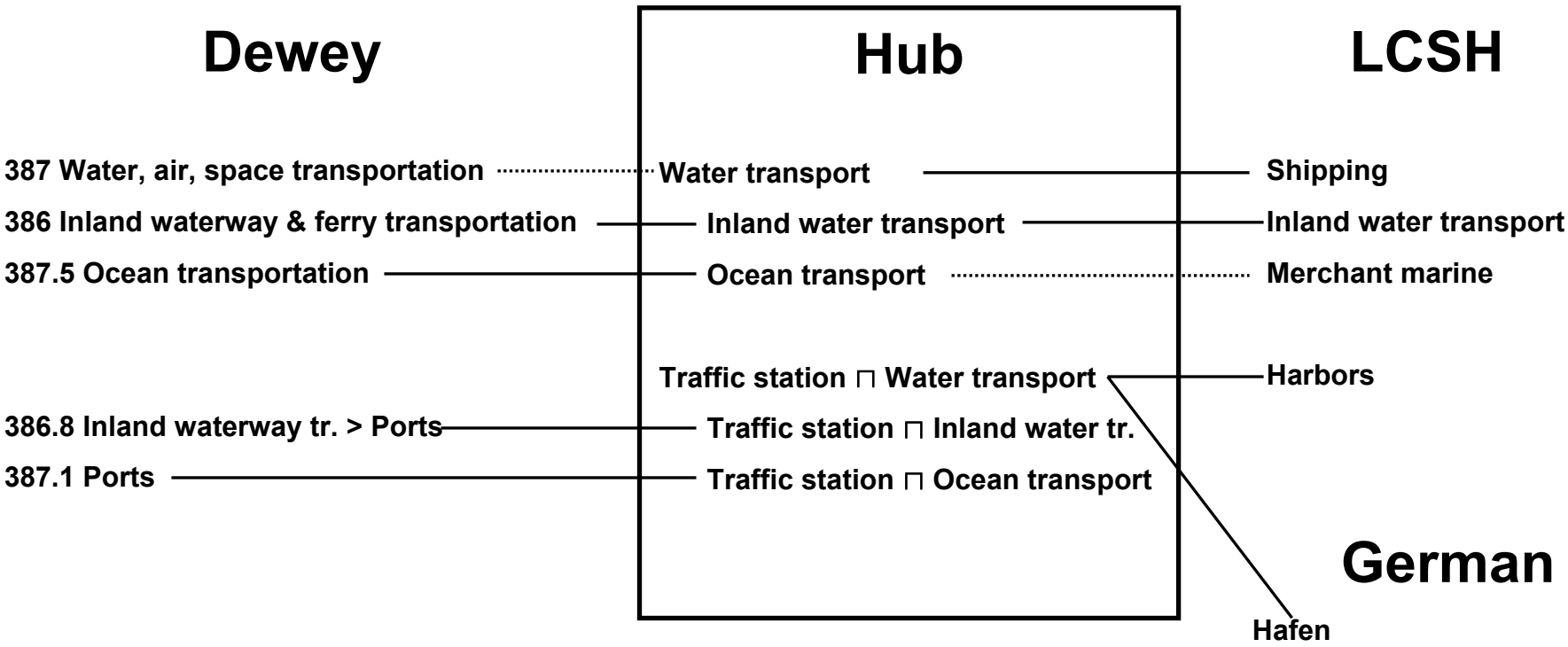
Improve semantic-based search
across multiple collections in multiple languages.

- Interoperability between any two participating KOS (Knowledge Organization Systems)
- Support for search, esp. facet-based search
  - for any collection indexed by a participating KOS
  - for search based on free-text or free-form social tagging
- Assistance in cataloging (metadata creation) by catalogers or users (social tagging)
- Long-range goal: Web service where a KOS can be uploaded and mappings to specified target KOS are returned

KOS Concept Hub

- Interoperability is achieved by expressing concepts from all participating KOS as a canonical representation, such as a description logic formula using atomic concepts and relationships
- The backbone of the proposed system is a faceted core classification of atomic concepts together with a set of relationships
- Mapping from KOS to KOS is achieved by reasoning over these canonical representations

Mapping through a Hub



Method: How to get DL formulas

Key: Efficient creation of canonical representations (DL formulas)

- Apply existing knowledge:
  Large knowledge base ➡ less effort for processing a new KOS
- Use knowledge of KOS structure for hierarchical inheritance
- Use linguistic analysis of terms and captions
- Eliminate redundant atomic concepts
- Check or produce mapping results from assignment of concepts to the same records
- Get human editors' input and verification where needed through a user-friendly interface
- KOS "owners" may verify and edit data pertaining to their KOS

Knowledge base

Requires an ever larger classification and lexical knowledge base containing many kinds of data:

1. A faceted classification of atomic concepts

Seeded from sources with well-developed facets such as

- UDC
- the Alcohol and Other Drug (AOD) Thesaurus
- the Harvard Business Thesaurus
- the Art and Architecture Thesaurus
- various systems called ontologies

Knowledge base 2

Requires an ever larger classification and lexical knowledge base containing many kinds of data:

2. **Linguistic knowledge bases** such as WordNet and mono-,bi-, and multi-lingual dictionaries and thesauri

3. **Many KOS (Knowledge Organization Systems),** such as LCC, UDC, DDC, DMOZ directory, LCSH, Gene Ontology, Schlagwortnormdatei

4. These will over time be **fused into one large multilingual knowledge base** with many terminological and translation relationships and relationships linking terms to concepts,
   with an increasing number of concepts semantically represented by a DL formula.

Examples of deriving DL formulas

L00 Transportation and traffic

- L10 Traffic system components
  - L13 Traffic facilities
    - L15 Traffic stations
  - L17 Vehicles
- L30 Modes of transportation
  - L33 Air transport
  - L37 Water transport

P00 Buildings, construction

- P23 Buildings
- P27 Architecture
- P43 Construction

R00 Engineering

- R30 Acoustics
- R37 Soundproofing

T70 Military vs. civilian

- T73 Military
- T77 Civilian

HE Transportation HE550-560 Ports, harbors, docks, wharves, etc.	L00 Transportation and traffic ⊓ T77 Civilian *Inherited:* L00 Transportation and traffic ⊓ T77 Civilian *Added by editor*: L15 Traffic stations ⊓ L37 Water transport *Resolved to*: **L15 Traffic stations ⊓ L37 Water transport⊓ T77 Civilian**

NA6300-6307 Airport buildings	*From database already established*: Airport = L15 Traffic stations ⊓ L33 Air transport Buildings = P23 Buildings *Added by editor* T77 Civilian *Resolved to* **L15 Traffic stations ⊓ L33 Air transport ⊓** **P23 Buildings ⊓ T77 Civilian**

TL681.S6 Airplanes. Soundproofing	*From database already established*: Airplane = L17 Vehicles ⊓ L33 Air transport Soundproofing = R37 Soundproofing *Added by editor*: Nothing *Resolved to* **L17 Vehicles ⊓ L33 Air transport ⊓ R37 Soundproofing**

Aeroplanes-Soundproofing	*From database already established*: Aeroplanes = Airplane [Spelling variant] *Therefore* Term is recognized as same as Airplanes. Soundproofing *Resolved to* **L17 Vehicles ⊓ L33 Air transport ⊓ R37 Soundproofing**

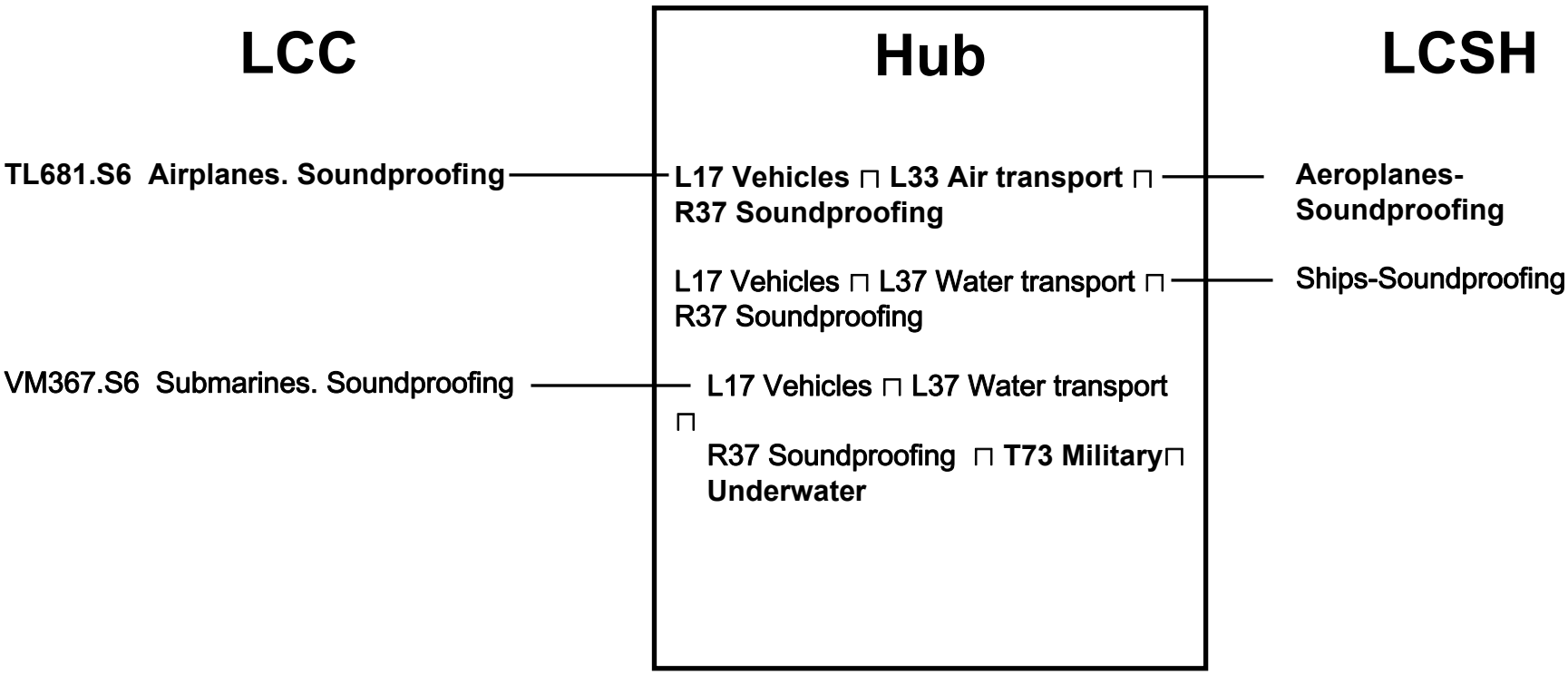
Any class formed by geographical subdivision Such as NA6300-6307 Airport buildings **NA6305.E3 Egypt**	Recognized using a dictionary of geographical names Inherits from subject class above it; simply add the country L15 Traffic stations ⊓ L33 Air transport ⊓ P23 Buildings ⊓ T77 Civilian ⊓ **Egypt** No editor checking needed

Examples from the resulting knowledge base

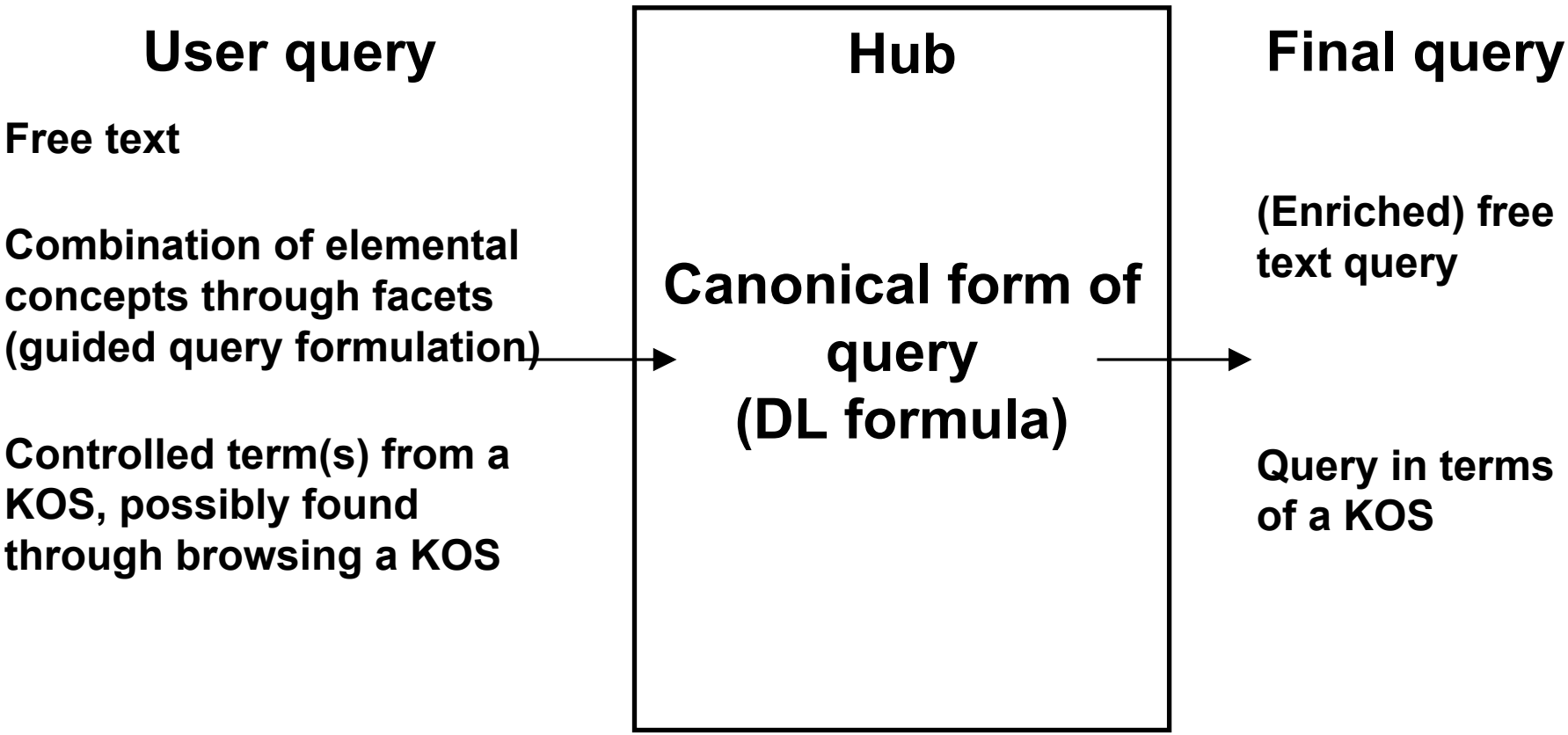
HE550-560 Ports, harbors, docks, wharves, etc.	=	L15 Traffic stations ⊓ L37 Water transport ⊓ T77 Civilian
NA2800 Architectural acoustics	=	P27 Architecture ⊓ R30 Acoustics
NA6300-6307 Airport buildings	=	L15 Traffic stations ⊓ L33 Air transport ⊓ P23 Buildings ⊓ T77 Civilian
NA6330 Dock buildings, ferry houses, etc.	=	L15 Traffic stations ⊓ L37 Water transport ⊓ P23 Buildings ⊓ T77 Civilian
TC350-374 Harbor works	=	L15 Traffic stations ⊓ L37 Water transport ⊓ R00 Engineering ⊓ T77 Civilian
TH1725 Soundproof construction	=	P23 Buildings ⊓ P43 Construction ⊓ R37 Soundproofing
TL681.S6 Airplanes. Soundproofing	=	L17 Vehicles ⊓ L33 Air transport ⊓ R37 Soundproofing
TL725-726 Airways (Routes). Airports and landing fields. Aerodromes	=	L13 Traffic facilities ⊓ L33 Air transport ⊓ Technical aspects
VA67-79 Naval ports, bases, reservations, docks	=	L15 Traffic stations ⊓ L37 Water transport ⊓ T73 Military
VM367.S6 Submarines. Soundproofing	=	L17 Vehicles ⊓ L37 Water transport ⊓ R37 Soundproofing ⊓ T73 Military ⊓ Underwater

Aeroplanes-Soundproofing	=	L17 Vehicles ⊓ L33 Air transport ⊓ R37 Soundproofing
Airports-Buildings	=	P23 Buildings ⊓ L15 Traffic stations ⊓ L33 Air transport
Buildings-Soundproofing	=	P23 Buildings ⊓ P43 Construction ⊓ R37 Soundproofing
Ships-Soundproofing	=	L17 Vehicles ⊓ L37 Water transport ⊓ R37 Soundproofing

Mapping through a Hub



Mapping user queries



Query:
L17 Vehicles AND R37 Soundproofing

TL681.S6 Airplanes. Soundproofing	[L17 Vehicles ⊓ L33 Air transport ⊓ R37 Soundproofing]
VM367.S6 Submarines. Soundproofing	[L17 Vehicles ⊓ L37 Water transport ⊓ R37 Soundproofing ⊓ Military]
Aeroplanes-Soundproofing	[L17 Vehicles ⊓ L33 Air transport ⊓ R37 Soundproofing]
Ships-Soundproofing	[L17 Vehicles ⊓ L37 Water transport ⊓ R37 Soundproofing]

Examples from NALT and LCSH

- NALT National Agricultural Library Thesaurus
- LCSH Library of Congress Subject Headings

Air pollution laws

LCSH term

Air – Pollution – Laws and regulations

[isa] Legal rule [appliedTo] {[isa] Condition [isConditionOf] Air [causedBy] Pollutant [property] Undesirable}

NALT terms

Air pollution

[isa] Condition [isConditionOf] Air [causedBy] Pollutant [property] Undesirable

Laws and regulations

[isa] Legal rule

Mapping LCSH —► NALT

Air – Pollution – Laws and regulations —► Air pollution AND Laws and regulations

Interpretation for indexing and searching in both directions

Soil moisture vs. Soil water

LCSH term

Soil moisture

[isa] Water [containedIn] Soil

NALT term

Soil water

[isa] Water [containedIn] Soil

Mapping LCSH —► NALT

Soil moisture —► Soil water

Greenhouse gardening

LCSH term

Greenhouse gardening

[isa] Gardening [inEnvironment] Greenhouse [inEnvironment] Home

NALT terms

Home gardening

[isa] Gardening [inEnvironment] Home

Greenhouse

[isa] Greenhouse

Mapping LCSH —► NALT

Greenhouse gardening —► Home gardening AND Greenhouse

Salad greens

LCSH term

Salad greens

[isa] Green leafy vegetable [usedFor] Salad

NALT term

Green leafy vegetables

[isa] Green leafy vegetable

Mapping LCSH —► NALT

Salad greens —► **BT** Green leafy vegetables

Emerging diseases

LCSH term

Emerging infectious diseases

[isa] Disease [hasProperty] Infectious [hasProperty] Emerging

NALT term

Emerging diseases

[isa] Disease [hasProperty] Infectious ??? [hasProperty] Emerging

Mapping LCSH —► NALT

Emerging infectious diseases —► Emerging diseases

Emerging infectious diseases —► **BT** Emerging diseases

Distributed implementation

- A KOS on the Web could assign DL formulas to its concepts − let's call this a
  semantically enhanced KOS or SEKOS
- Could use any of a number of faceted core classifications or even several (using a unique URI for each elemental concept)
- Core classifications could be mapped to each other
- It is now a simple matter to map from any SEKOS to any other (somewhat dependent on the core classifications used)

Take-home message

Semantics gives powerful systems

Dagobert Soergel

dsoergel @ umd.edu

www.dsoergel.com

T