Semantic Problems of Thesaurus Mapping

Martin Doerr

Information Systems Group

Institute of Computer Science
Foundation for Research and Technology - Hellas

Lund,
April 10-12, 2002

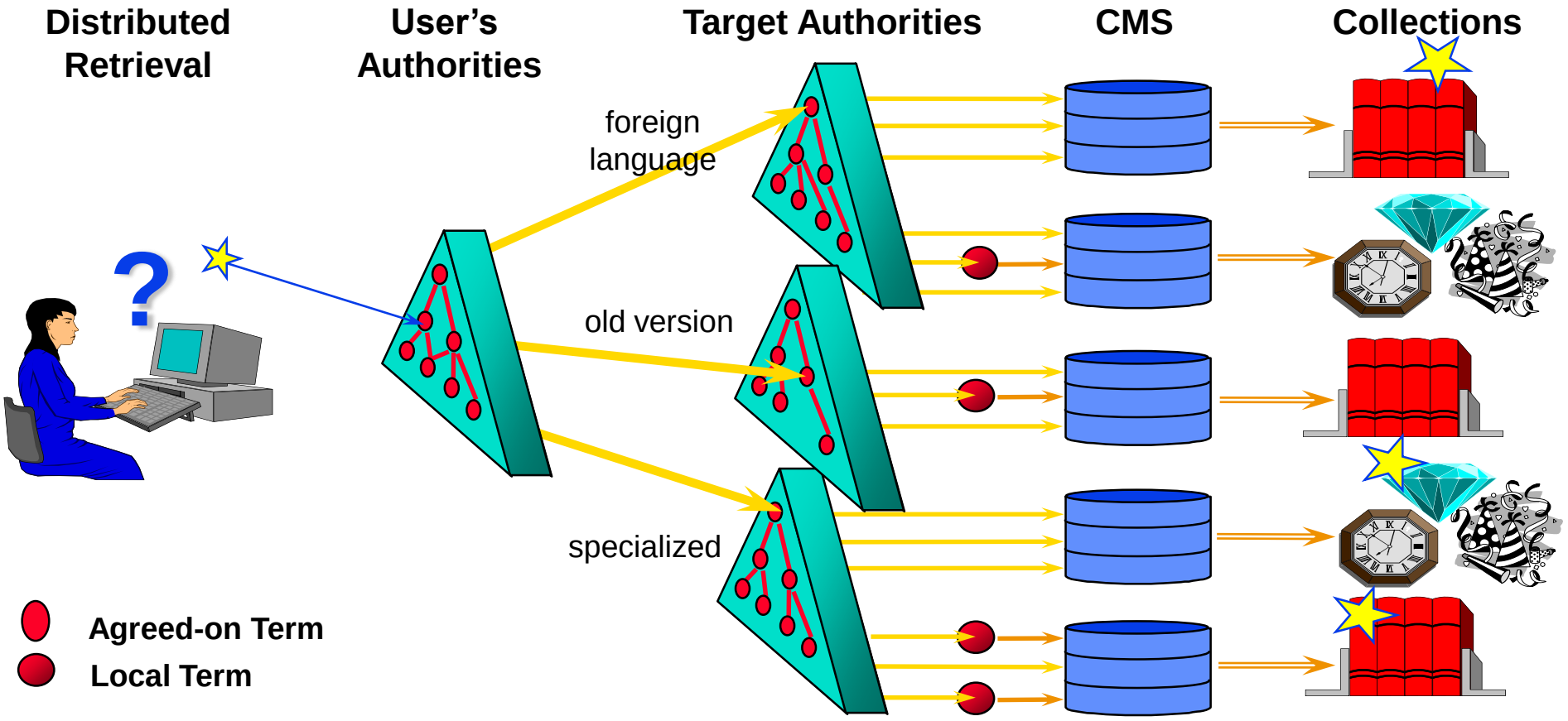
Thesaurus Mapping
Thesaurus Interoperability

- Objectives: Global access to heterogeneous information sources
- Contextual problems of information sources:
  - Different providers
  - Different objectives
  - Overlapping topics/ themes
- Where do we need thesauri ?
  - Enhancing full text retrieval, query formulation aids
  - Querying structured data & metadata with controlled vocabularies
  - Classification systems for information organization

Thesaurus Mapping
The Problem

- I ask for Cactus - you know Cholla...
  - I "chaffinch" - you "fringilla coelebs
  - I "dolls, Hopi" - you "kachina"
  - I "Champs Elysees" - you "France"
  - I "Greece, Acropolis" - you "restaurant Acropolis"
  - I "Architecture (studies)" - you : "Architecture (buildings)"
- Thesauri differ
  - in language: natural, scientific or by convention
  - in subject: coverage, completeness and detail
  - in version: state of development

Thesaurus Mapping
"Thesaurus Transition"



Thesaurus Mapping
Why do we need mapping?

- **Thesaurus mapping is central for:**
  - Thesaurus merging
  - Thesaurus correlation / interlinking
  - Thesaurus federation
- **Mapping can be concept-based:**
  - Terms are identified with the set of objects they correctly classify
  - Broader terms are regarded to classify supersets
  - Correct mapping is defined through equivalent query results
  - Depends on term use rather than comprehension of a term
  - Mapping logic should conform with query paradigm (Z39.50?)

Thesaurus Mapping
Two approaches – Three communities

Automatic mapping:

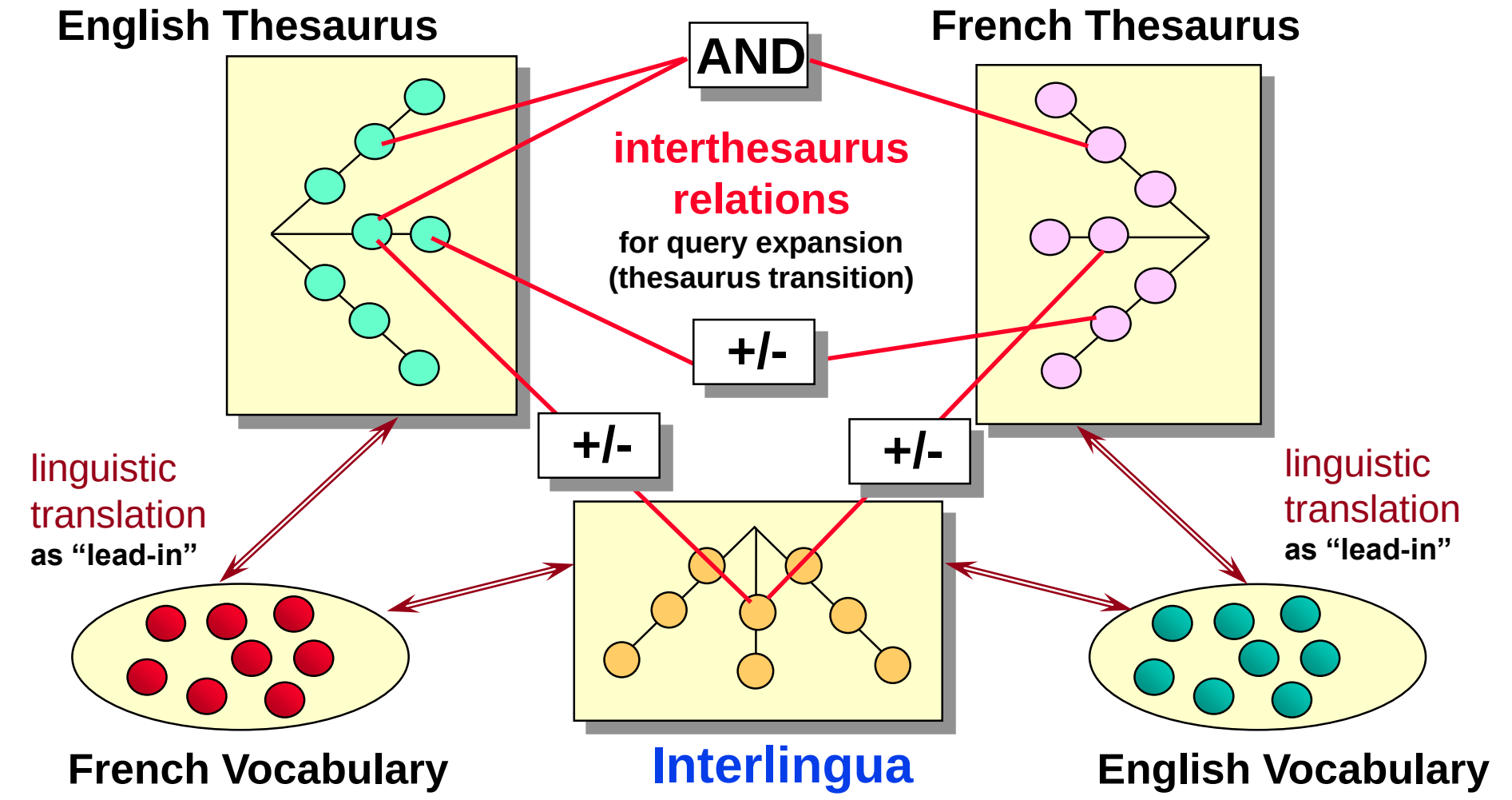
- Based on parallel indices/ similar documents
- Statistical & neural network methods
- Cheap and with optimal coverage
- Missing intellectual insight
- Cannot separate if terms express different aspects or if terms are used for different aspects. (May confuse mapping of concepts with concept co- occurrence in the document sample)
- limited precision

Thesaurus Mapping
Two approaches – Three communities

- **Intellectual mapping:**
  - **Manual, based on expert knowledge about terms**
  - **Can be supported by Description Logics ("Ontologies")**
  - **Expensive, but with high precision**
  - **Insight in structure and long-term stability**

- *Proposition: The intellectual structures are complex. Their investigation is helpful for better intellectual and refined statistical mapping methods.*

Thesaurus Mapping
Translation and Mapping



Thesaurus Mapping
Logics of Mapping for Z39.50

Interthesaurus relations (ISO 5964):

- **partial equivalence**

  Must become: broader equivalence (is subset of)
  narrower equivalence (is superset of)

- **exact equivalence (same set as)**

- **inexact equivalence (overlaps with)**

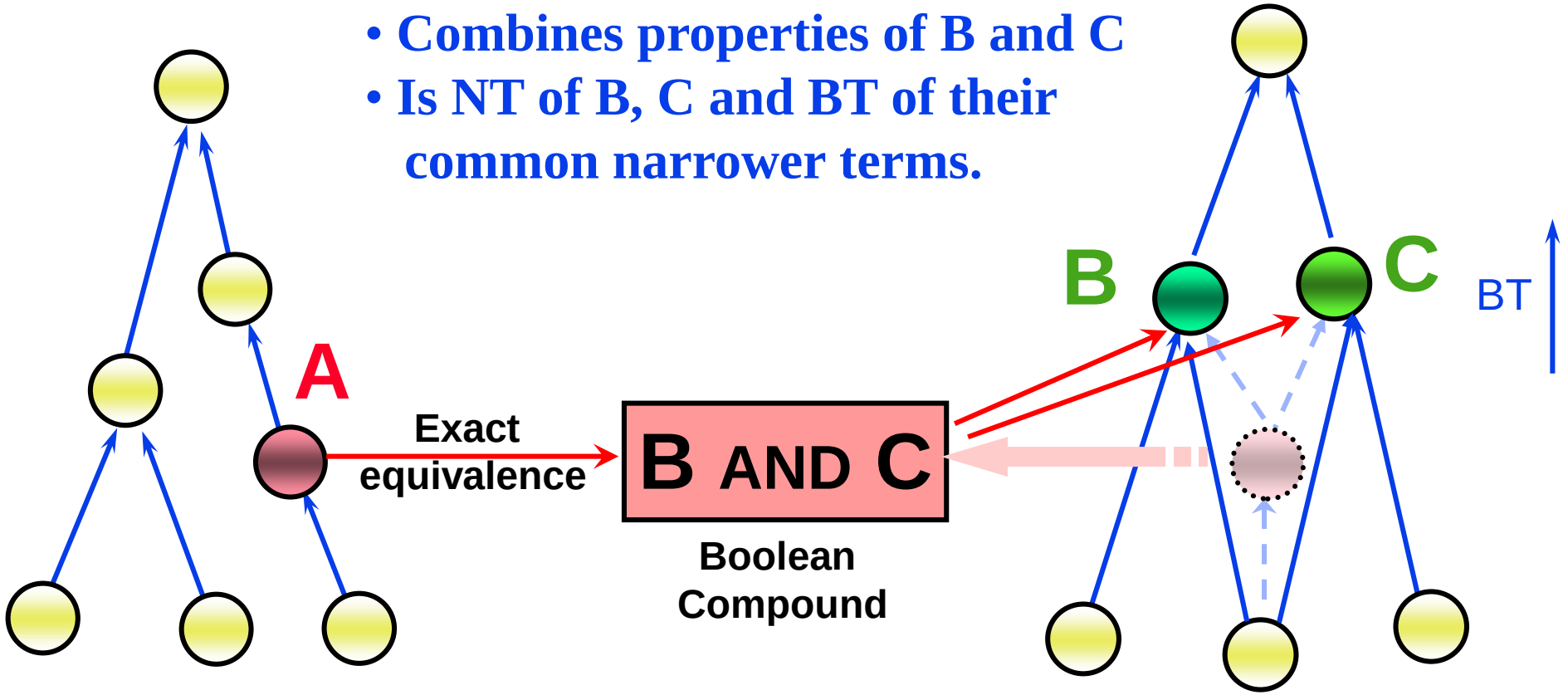
  good for FTR only

- **single to multiple equivalence**

  Must become: exact equivalence to BOOLEAN combination of target terms:
  "AND" (intersection), "OR" (union), "NOT" (complement)

Thesaurus Mapping
Boolean AND-Combinations

- Uses instances of both, B and C
- Combines properties of B and C
- Is NT of B, C and BT of their common narrower terms.



Thesaurus Mapping
Issues of Mapping Logics for Z39.50

How to use Boolean expressions inversely :

- Calculation of inferences
- Boolean combinations to a post-coordinated thesaurus:

  How to index the existence of an incoming link ?

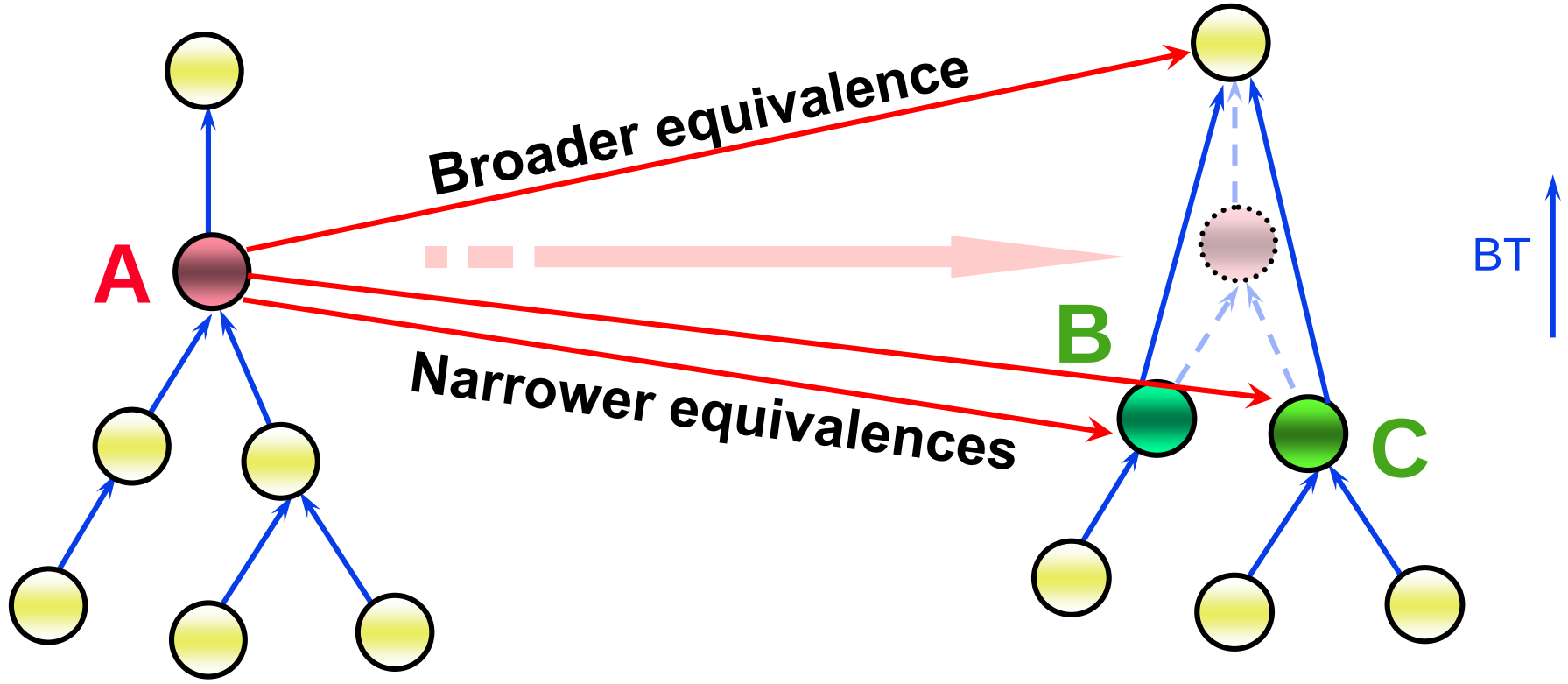
Mappings must be complete:

- Should guarantee recall over non-equivalent terms :

  preservation of precision or recall should be selectable
- Should avoid redundancies, need consistency control !
- Should avoid Combinatorial explosion:

  Need cascading Thes A => Thes B => Thes C

Thesaurus Mapping
Approximation by Inclusion



Thesaurus Mapping
Obstacles to Thesaurus Transition

- **Unclear coverage & incompatible organisation.**
  - Special vocabularies often contain general terms, contract upper levels. No global abstraction levels.
  - Missing or contradictory NT/BT relations.
  - "Loose" NT semantics (like part-whole, see-also etc.).
  - Arbitrariness of monohierarchies :

    E.g. : A hierarchy of colorants, like "red organic dye":

    organize it: by composition, production method or origin ?

    by color ?

    by physical property or function ?

Thesaurus Mapping
Obstacles to Thesaurus Transition

Term semantics.

- **Post-coordination should make use of DL:**
  - Combinations from disjoint facets: “factories + grinding”.
  - Unclear rules for allowed combinations.
  - How to attach and index synonyms in a post-coordinated hierarchy.
- **Use-induced incompatibility:**
  - E.G. Subject/object : “brigde” - “bridge construction.”
- **“Complementary polysemy” (Pustejowsky):**
  - Context-induced shifts of meaning: door, architecture etc.

  … cause context-related differences in hierarchy.

Thesaurus Mapping

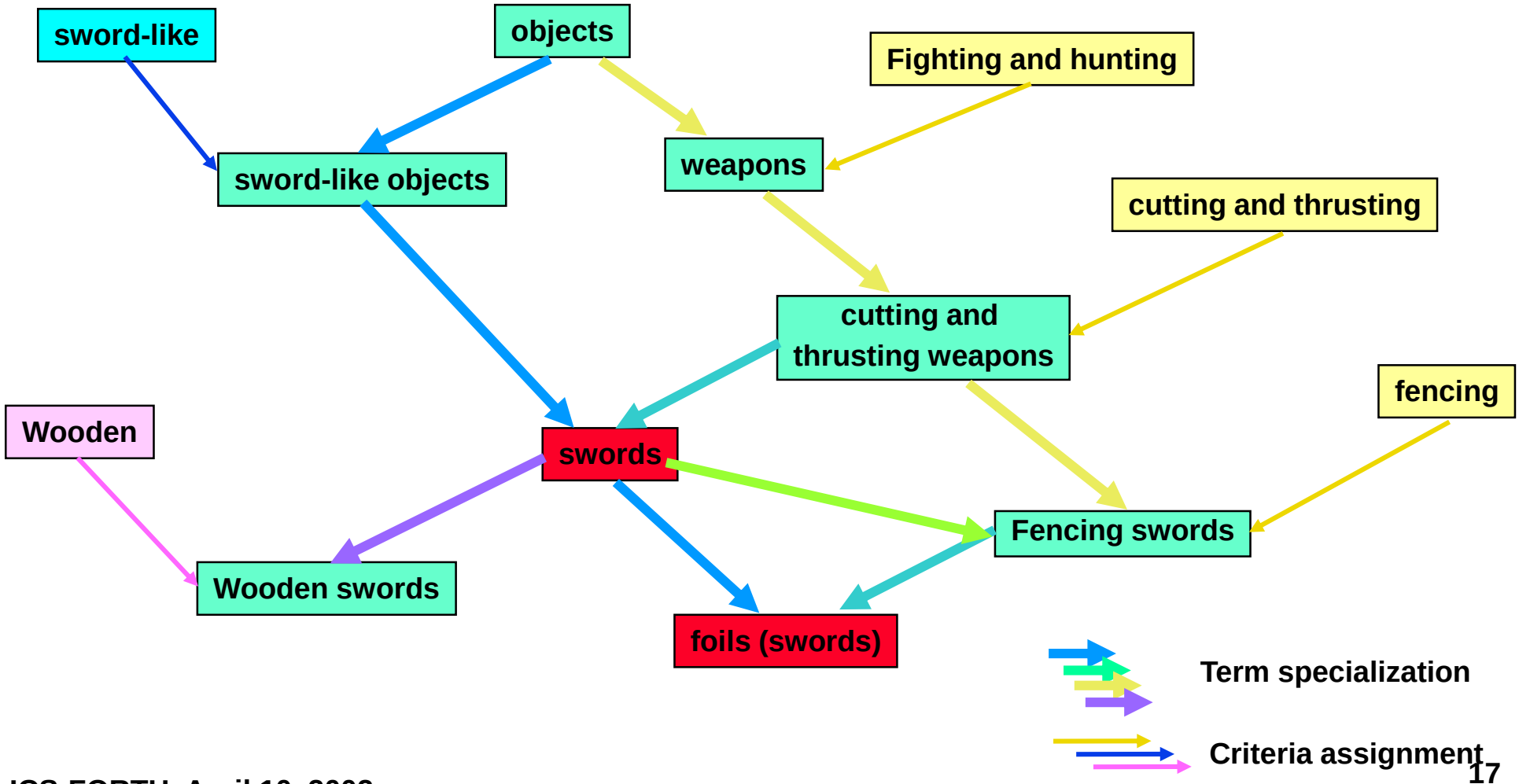
Complementary Polysemy and Minor Facets

- "Minor facets" provide explicit context criteria:
  - E.G. MDA archeological thesaurus:
    armour by construction : scale armour
    armour by form : cuirass
    armour by function : parade armour
  - Are these criteria idiosyncratic?
  - How do they relate to each other ?
  - How do they relate to compound term formation?

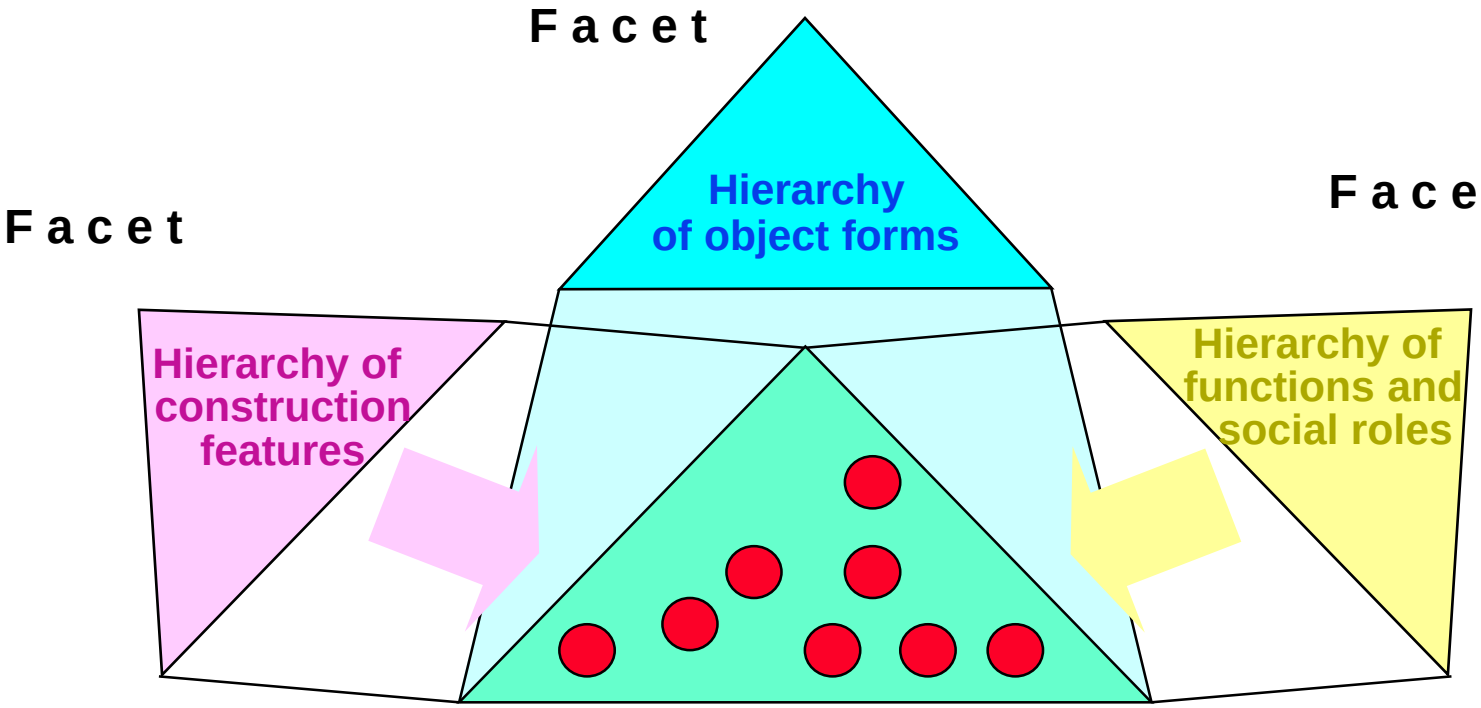
Thesaurus Mapping
Minor Facets in the AAT

- **The "object" facet (1998 edition) contains:**
  - About 1640 facet indicators,
  - About 600 with explicit criteria ("by form etc..")
  - Using 150 ! criteria
- **Preliminary frequency analysis of criteria:**
  - Form: 35%, function: 30%, placement: 15%, construction: 15%, social context: 5%…
- **Hypothesis:**
  - Minor facets criteria can be systematically generalized
  - Minor facet criteria are different kinds of NT relations

Thesaurus Mapping
Narrower Terms for three Facets



Thesaurus Mapping
Explicit facet criteria for objects



Thesaurus Mapping
Summary of Semantic Problems

We could identify four semantic problems

(statistical methods are not sensitive to semantic problems)

- **Logics of query term expansion between compatible hierarchies**
- **Theory of concept formation by compound terms, linguistic and semantic. "KR " should collaborate with experienced thesaurus editors.**
- **Understanding of context –dependency of term hierarchies: understanding of the role of complementary polysemy differences between subject and object classification.**
- **Meaning of terms versus meaning of term used for a document**

Thesaurus Mapping
What To Do

Research: Deeper understanding.

- **Investigation of polyhierarchies, polysemy and BT/NT semantics.**
- **Theory of concept formation by compound terms, from linguistics and logic.**
- **Use of ontologies as "top-level thesauri", to provide.**
  - highest levels (like physical objects, actors, events).
  - "roles" for concept formation (e.g. "using", "made for", "made in").
  - transition between single terms and terms in multiple fields (e.g. type: "sword", material: "wood" versus "wooden sword").

Thesaurus Mapping
What To Do

Protocols: enabling dynamic thesaurus transition

- **Metadata for description of the logic of a thesaurus**
  - BT/NT semantics, organization principles, lead-ins
- **Recall/precision control in thesaurus transition**
- **DL-based post-coordination rules. Explicit use of "Roles".**

Practice: Analysis of semantic heterogeneity

- **Comparing thesauri wrt logic of construction and intended use.**
- **Understanding semantics of automatic mappings, integration of intellectual and automatic methods.**