Thesaurus Building

Martin Doerr, Maria Daskalaki,

Chrysoula Bekiari

Center for Cultural Informatics
Institute of Computer Science
Foundation for Research and Technology - Hellas

Berlin
December 18, 2014

Thesauri

Words, Terms, Concepts

- **Words**
  - Constituents of **natural** languages. Categorical meaning, in contrast to "proper names". Multiple senses depend on context. (Example: "order")

- **Term**
  - Constituent of **expert** language. A word with a specific (categorical) meaning, either defined in a scientific document or common to an expert group and discipline. (Example: "hepatitis A")

- **Concept**
  - A class or set of items grouped together on the basis of some implicit or explicit criterion or rule. The criterion can be unconscious or even innate ! (Example: "δημόσιος υπάλληλος" civil servant).
  - A concept is **not** a term and **not** a language element!

Thesauri

Functions of Terminology

- **Unambiguous scientific expression**
  - Use in expert discussions, expert opinions (diagnoses!) and scientific publication. Defined in disciplinary dictionaries.

- **Research**
  - Defined ad-hoc to **discriminate** items in a research project (archeology!). Conclude **from** form **on** function, form on provenance etc.

- **Data search**
  - Find all items (publications, objects etc.) possibly relevant for my research question.

➔ ***Unfortunately, each function needs a different approach!***

➔ ***We deal ONLY with data search!***

Thesauri

A "Backbone Thesaurus"

- **How to agree on a common, coherent and consistent thesaurus within the framework of DARIAH?**
  - Every discipline has different concepts, **millions**!

- **Existing vocabularies:**
  - many small thematic vocabularies, discipline or just application specific, hardly hierarchically/semantically structured in a principled way
  - Larger vocabularies: Library subject headings LCSH, SWD, Rameau are not thesauri.
  - Dewey Classification: principled, hierarchical, but arbitrary & biased
  - "Good" thesauri: Getty's AAT, English Heritage, Merimee.

Thesauri

A "Backbone Thesaurus"

- **Integration efforts**
  - Libraries: Dewey translation. Aligning subject headings on Dewey.
  - Museums: AAT by translation
  - LCSH, SWD, Rameau subject headings: Manual match by "MACS", automatic matching still impossible by 2009: http://www.few.vu.nl/~aisaac/oaei2009/results.html

- **Manual mapping**
  - **The HEREIN** Project – an attempt to merge **3** thesauri in one that led to a complete failure
  - For **10** vocabularies: 45 mappings, for **100**: 4950, for **1000**: 499.500 !!!
  - ***An impossible undertaking!***

Thesauri

A "Backbone Thesaurus"

Idea: ***It is effective to globally agree only on very few concepts.***

- **Why**
  - For **resource discovery** recall is more important than precision
  - Terms for discovery can be much coarser than for documentation!
  - The more generic the concept - the higher the recall
  - "Known Item Search" works better by keywords and factual associations
- **Therefore**
  - It is most efficient to agree on the higher terms
  - Agreeing on higher terms avoids most basic incompatibility among terms
- **We propose a backbone thesaurus**
  - Map all vocabularies to one set of top terms, may be just **30**, less than 100
  - **Not imposing** terms on experts, but providing a common order

Thesauri

A "Backbone Thesaurus"

- UMLS demonstrates feasibility

- A Common "backbone" thesaurus or "metathesaurus" as top-level
  - An ***indexing language*** and "***interlingua***".
  - From which smaller vocabularies "borrow' the upper concepts
  - "streamlines" hierarchies by providing fundamental ontological distinctions as normalizing principles
  - Map vocabularies into the common backbone (alternative upper levels)
  - Stay close to Getty's AAT
  - Preserve context-focus of small vocabularies ***NOT*** as top-term, but as ***NEW*** contextual relationship or "Term Collection".

- Only a faceted classification is enough compact and unambiguous

Thesauri

What is a Faceted Classification?

- A set of facets comprises "clearly defined, mutually exclusive(*), and collectively exhaustive aspects, properties or characteristics of a class or specific subject". (Taylor, A.G., Introduction to Cataloging and Classification, 1992).
  (*) better: mutually independent, one not implying another, “orthogonal”

- Facets are fundamental concepts which appear as characteristic syntactic constituents for term composition, such as:
  “Persian X 19th century X rugs”

- Facets can thus provide a global subdivision of concepts through the reduction of the composite terms to more primitives ones.

Thesauri

S. R. Ranganathan

- **Three cognitive "planes":**
  - Idea plane - Verbal plane - Notational plane
  - confusion hinders analysis and problem solution:
  - Missing terms for existing ideas (concepts are many, words are few) and
  - notational limitations inhibit idea plane work.
- **The invention of the "facets"**
  - Priority of the idea plane (= concept, not term)
  - Conceptual structures are multidimensional
  - Shelving of books is no argument, a taxonomy is not an index.

- **Colon Classification** *is a system of library classification developed by* ***S. R. Ranganathan*** *between 1925-1965. It uses five primary categories, or facets, to further specify the sorting of a publication. Collectively, they are called PMEST.*

Thesauri

Typical Facets

- **MARC: *subdivision*** by Period, Geography, Genre
  (Grammatical element of an indexing expression)

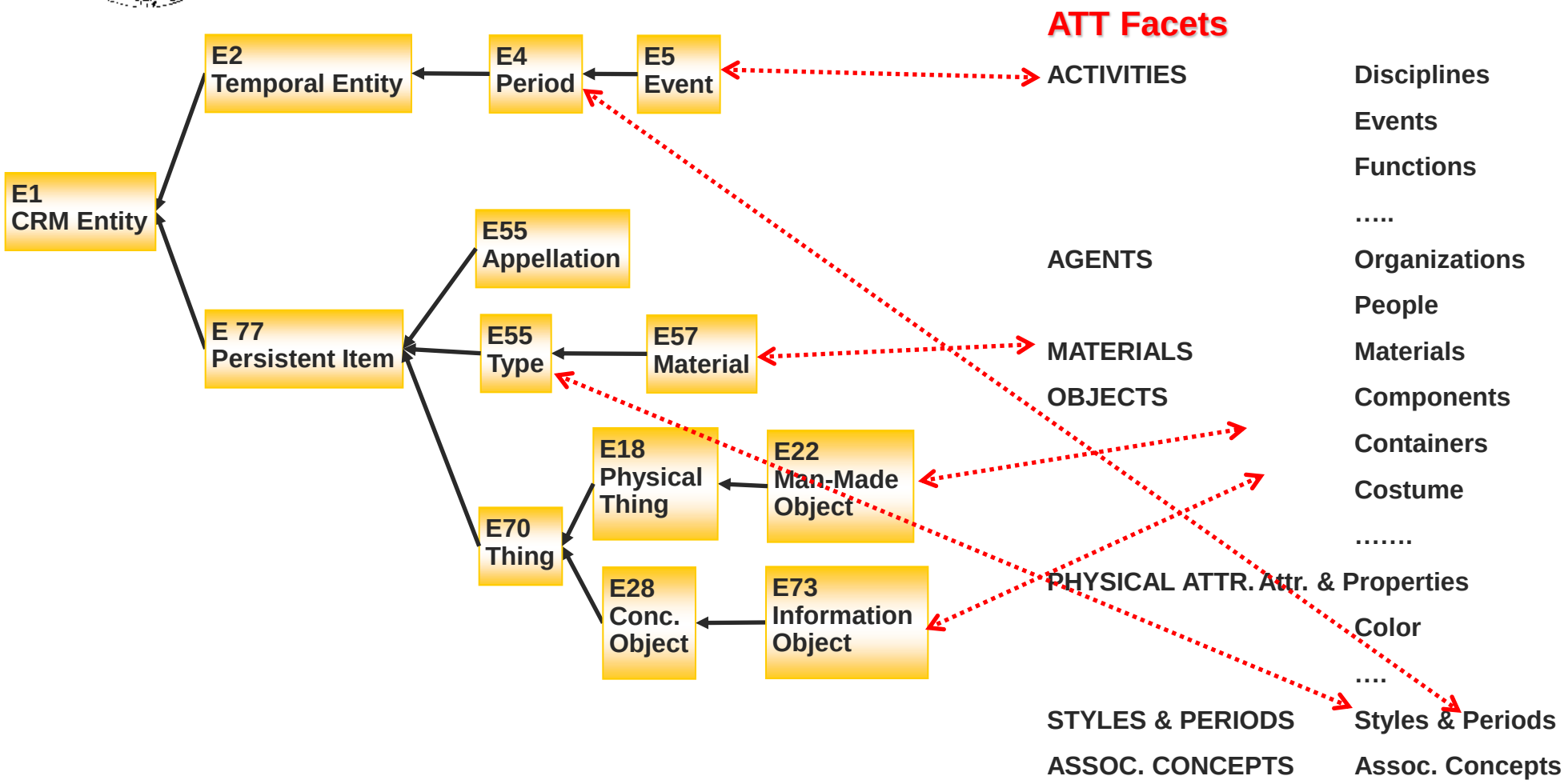
- **Ranganathan:** Personality, Matter, Energy, Space, Time

- **AAT:** Agents, Objects, Materials, Styles and Periods, Activities, Physical Attributes, Associated Concepts.

- **CIDOC CRM**: Actor, Physical Thing, Conceptual Object, Material, Type, Language, Period, Place, Time-Span, Dimension

Thesauri

CIDOC CRM / AAT mapping



Thesauri

Principles

How to design an effective faceted classification system

- Steps:

1. ***Define the functional restrictions of the indexing language:***

   What is the purpose of the classification?

   –> to facilitate a successful search of existing knowledge.Therefore, we have to find the *"bonds"* between the terms and not to demarcate/discriminate them.

   What is the domain of discourse of the indexing language?

   –>humanities (e.g.: the term "dependence" has different meaning in the context of computer science, medicine, psychology or social relations. All these specific meanings have a common basis: is a kind of (unequal) relationship, but this is not useful for indexing anything.)

Thesauri

2. Defining Concepts by Intensional Properties

2. Detection of the intensional properties of concepts (substance, nature, Erkennungsmerkmale).

- Characteristics expressing the nature/substance of a concept and providing an unambiguous recognition of an item as belonging to a category.
- Intensional properties are essential: necessary and sufficient conditions for belonging to a category, cannot be replaced without loss of meaning.

e.g. *bachelor* is defined as 'unmarried man'. Not being married is an essential property of a bachelor, because one cannot be a bachelor unless he is an unmarried man (necessary condition) and any unmarried man is a bachelor (sufficient condition).

- Recognition must be based on **accessible** information
- When sufficient intensional properties are implicit or not commonly accessible, the term is defined through confining or referring to commonly known phenomena:

e. g. human being, necessary: DNA, genetical. accessible: confining morphological characteristics.

Thesauri

2. Defining the intensional properties

- Intensional properties justify/reveal hierarchical IsA relationships

e.g.: any bachelor must be a man.

- potential properties are consequences of the nature of a thing. They may be confined to a category or not. They may appear at some instances at some time.

e.g.: potential properties of the bachelor: no children, is male or female (not a child), live alone etc. ***Not confined to bachelor!***

e.g.: potential properties a person: can drive a car. ***confined to person!***

- Intensional properties allow for deduction of the potential properties of items belonging to this category.

Thesauri

3. What is a Good Broader Concept?

3. searching for broader categories, which enable an "open world".

- Generalizing a concept into a broader category means that it can be ensured that this concept possesses these more general intensional (and potential) properties, **possibly** together with other concepts under that category.
- All items (terms, classes) which ***are not*** included in a broader category are not characterized by the intensional properties of that category. It must be possible to identify things not belonging to this category.

e.g.: What is not a "Research Object" ?

A **good broader category** "confines" many potential properties ("behavior") that can only apply to this intension. Each refinement of intension may confine or guarantee another set of potential properties.

e.g.: "Material Object" can have weight, elasticity. A "living individual" consumes energy
e.g.: "Activity" is better than "shoemaking"

Thesauri

Good and Bad Terms

"*The function of a term is to conclude potential properties from intensional ones*"

- Concepts defined by potential relationships are bad for indexing:

e.g.: "Research Object" can be anything. No further independent property can be derived. The fact is generally not accessible. It is **incidental**. It can be derived from context.

e.g.: Defining "human" by "can driver a car". Then, an handicaped is not a man, but a robot is….

- In particular context of use, context of interest, spatiotemporal contexts must not be defining criteria

- Never define by negation (antonymity, complements): In an open world "having not a property" does not imply anything. Complete decomposition is a kind of negation!

e.g.: female human = not male human. ***What are transsexuals?***

Thesauri

Building Hierarchies

Forming broader categories from common substance/nature which enables/confines/guarantees potential properties ("behavior") does not divide the world into disjoint classes. A particular thing may fit to multiple intensions.

The higher categories can't be justified in a logically exhaustive and strict way but can be reached intuitively, by common sense and by reducing the more complex terms and concepts to their primitive ones.

Building such categories **bottom-up,** we eventually reach the context-independent levels, and finally elementary concepts we posses and through which we perceive and conceptualize our reality. Those we call **facets.**

Thesauri

The Golden Rules of Hierarchy Building

Only IsA relation and the effect of intension on behavior counts:

- *Levels of hierarchies are never absolute. Even Facets may have generalizations.*
- *Levels of hierarchy are never complete.*
- *Generalizations are never unique.*
- *Sets of sibling concepts are never complete. Anything that does not fit goes into the next level category, until a better specialization is found.*
- *Don't complete levels by "other objects" or "elephants and none-elephants"*
- *Particulars (gazetteers, person lists) are NOT terminologies (but other KOS)*

Based on this, we avoid most arbitrariness and context dependency

Collaborative development of an upper level becomes feasible.

Thesauri

Benefits

benefits of the faceted classification:

- Reveals the complexity of a term and reduces it to its fundamental components.
- It is not an artificial classification of the terms or a “top to bottom” classification, but is generated from the analysis/decomposition of one term in its elementary characteristics.
- A term can be classified in multiple hierarchies (e.g. doll toys/visual works).
- Is independent of the context, within which each a term appears although the context is crucial for the classification of a term in facets.
- Is based only in a restricted number of fundamental concepts.
- Can be expanded without disrupting or disorganizing existing facets and hierarchies and enables thus compatibility between different classification systems from different domains without imposing terms on the experts.
- It does not presuppose knowledge regarding the exact context of the terms.

Thesauri

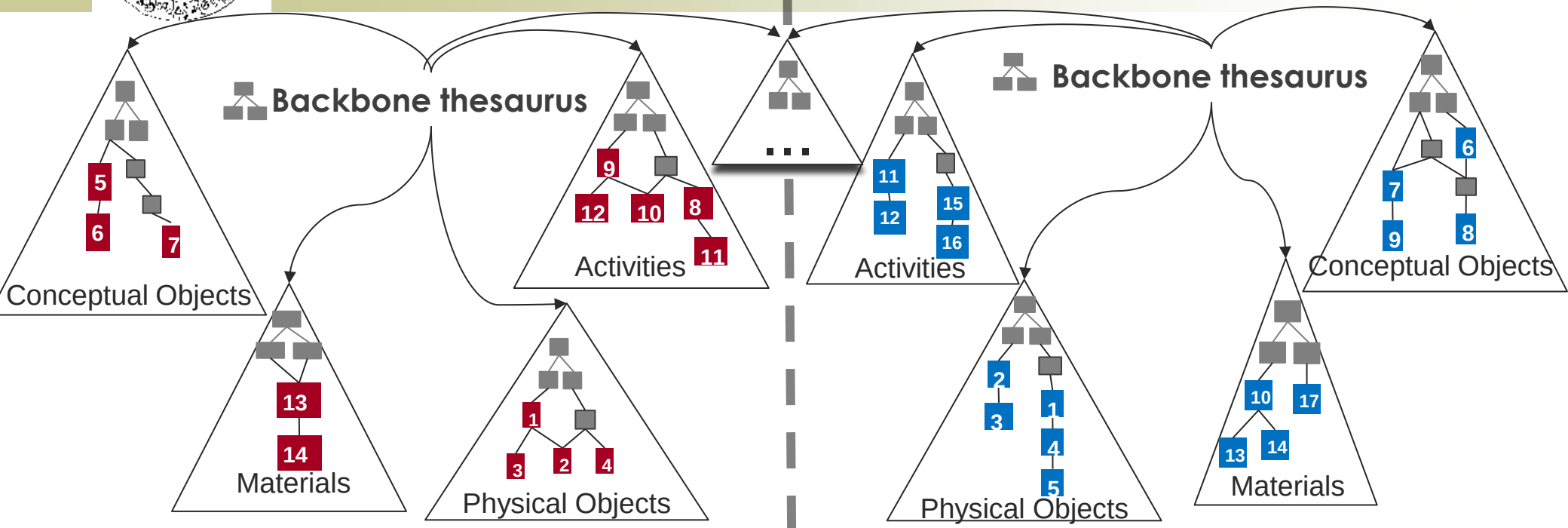
More benefits

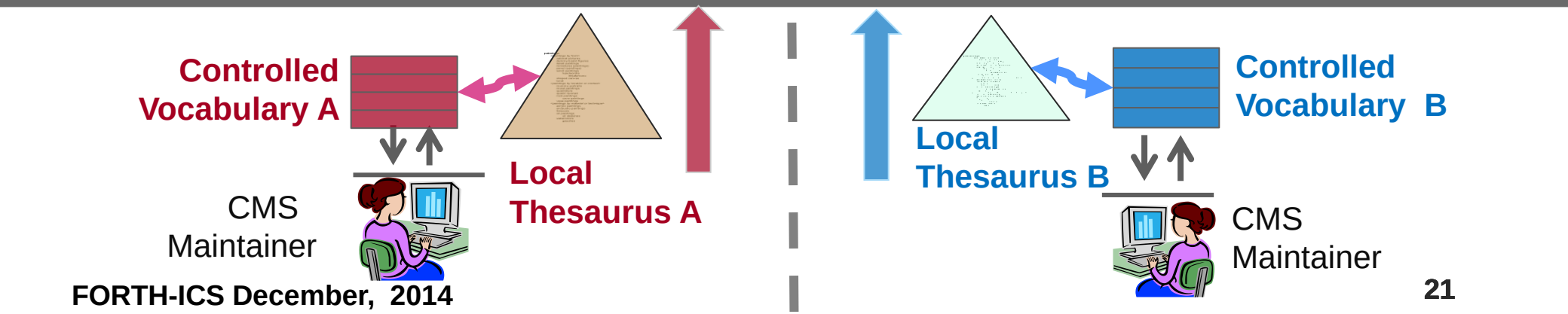
Faceted classification

- Helps us to discover concepts that are needed in searching or that enhance the logic of the concept hierarchy (e.g. train/bus station, harbor, airport=>traffic station)
- Does not divide the world in closed spheres of meanings according to specific characteristics, but brings to light hidden connections between the terms and establishes concept relationships.

Thesauri

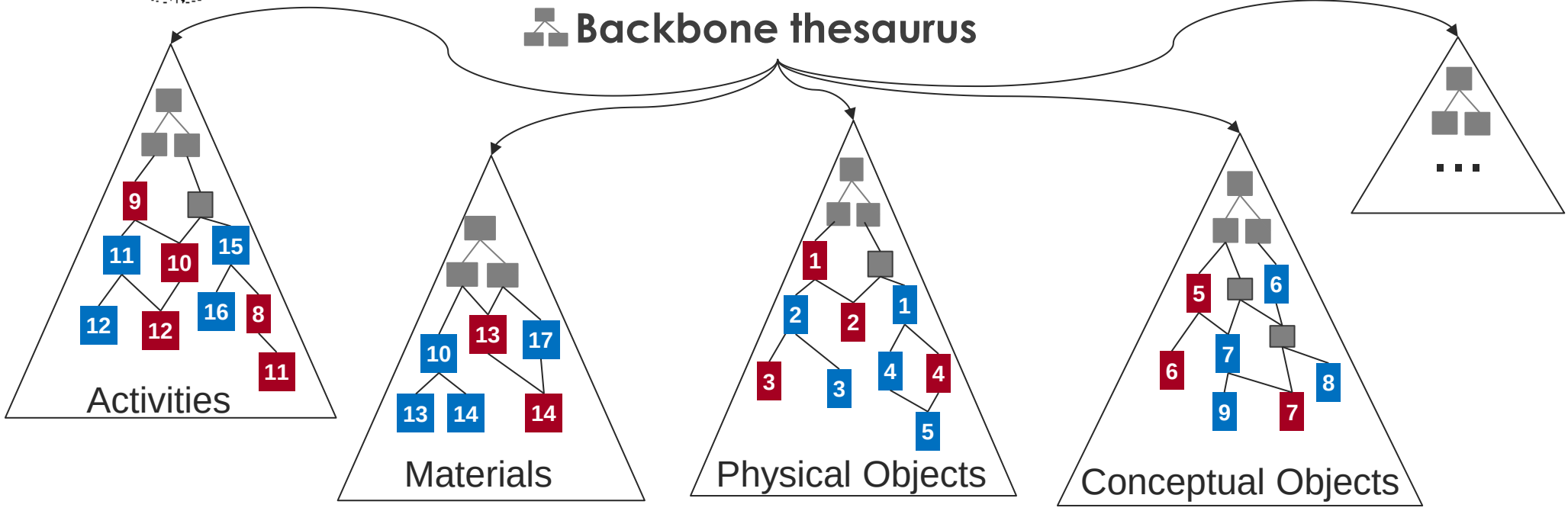
Mapping to the Backbone

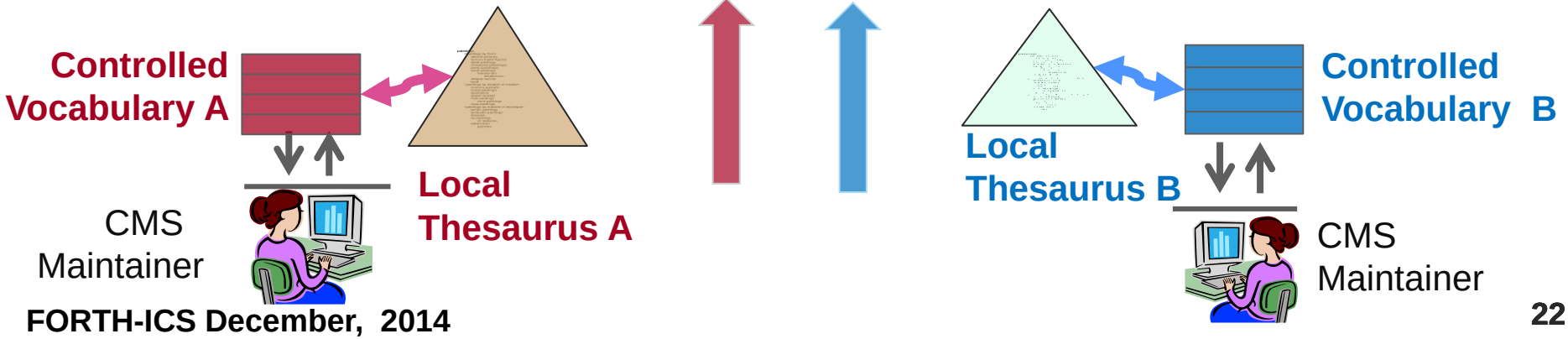




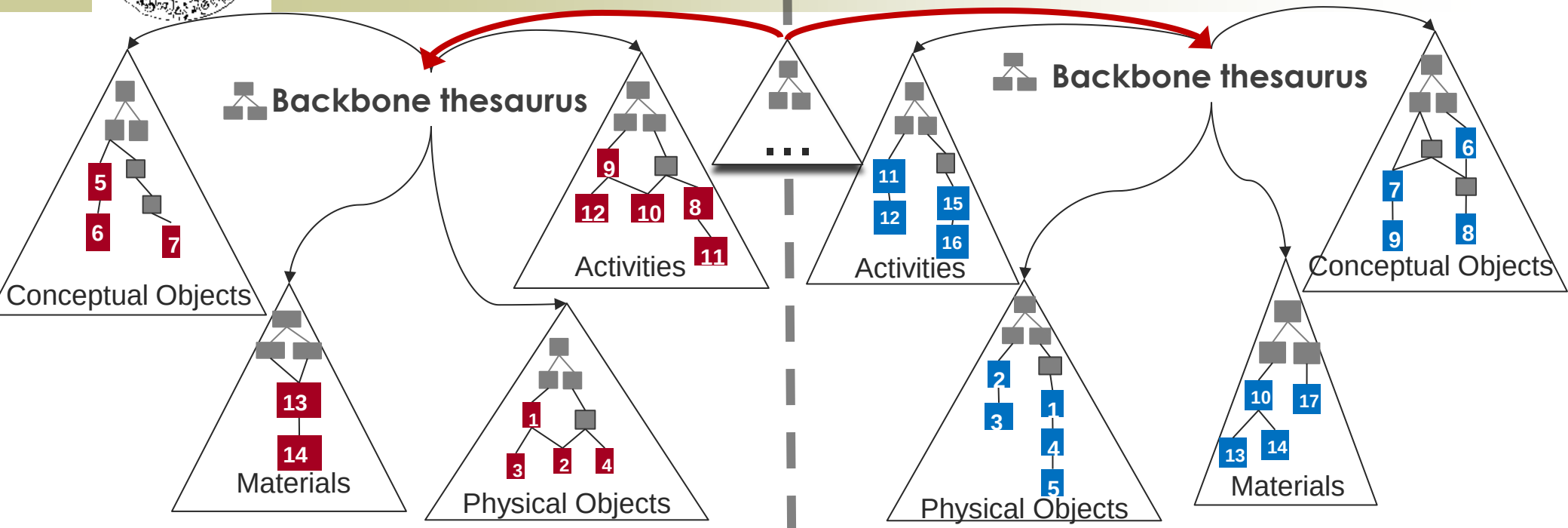
Thesauri

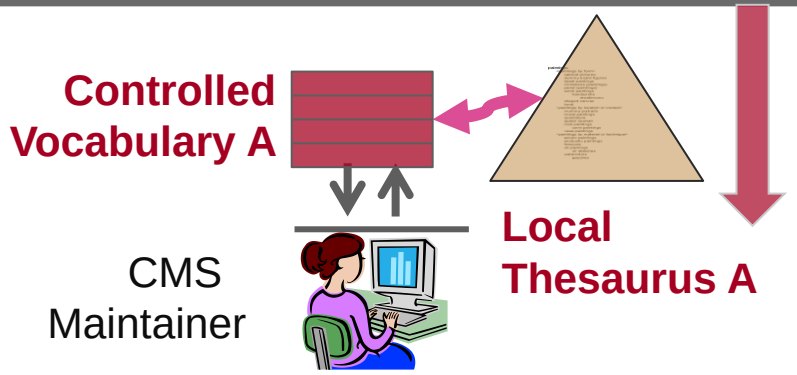
Merging under the Backbone

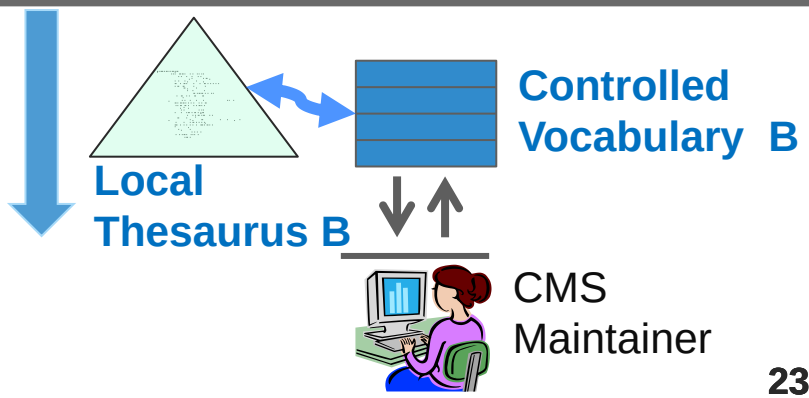




Thesauri *Projecting out of the Backbone*







Thesauri

Practical Steps

1. *Agree on the Method/ Principles*
2. *Partners propose individual facets and high-level concepts (“hierarchies”) and justify by principles and application. Reuse/improve existing ones.*
3. *Show how a facet integrates parts of vocabularies and provides better indexing power using SKOS.*
4. *Map into CIDOC CRM, AAT and ???*
5. *Agree on Facets step-by-step. Better gaps than ad-hoc generalizations that cause conflicts.*
6. *Continue maintenance and update mappings for ever.*

Thesauri

For illustration: our experimental facets

1. Facet: Materials

2. Facet: Material objects
   Hierarchies: i) monuments
   ii)artifacts/objects
3. Facet: Conceptual objects
   Hierarchies: i) symbolic objects=>information object
   ii) propositional objects=>information object
   =>methods=> processes, techniques

4. Facet: Natural Processes (“CRM Temporal Entity”)
   Hierarchies: i) natural disasters
   ii) natural geneses
5. Facet: Epochs (“CRM Temporal Entity”)

Thesauri

For illustration: our experimental facets

6. Facet: Activities (“CRM Temporal Entity”)

Hierarchies:

i) Disciplines => α) production of material objects and installations
b) conception and comprehension of phenomena
c) provision of knowledge and expertise (know-how)
d) production of aesthetic phenomena

ii) Events=> a) social events
b) conflicts
c) political, social and financial phenomena
d) administration

iii) Functions

iv) ….Other Activities : this is not a hierarchy!

Thesauri

For illustration: Scope notes

- **Facet: Activities**

Scope note: The “Activities” facet comprises types of intentional actions that result in the preservation, creation, production, modification or destruction of an entity (living beings, conceptual/material objects, groups, social, intellectual, physical etc phenomena).

Hierarchies

I) Disciplines: This hierarchy comprises types of branches of professional or potentially professional occupations socially and/or legally acceptable under the criteria of sector self-subsistence, practice efficiency, adoption of common methods and transferability of knowledge and expertise. Each sector includes types of unified activities that express some sort of professional or potentially professional specialization

II) Events: This hierarchy comprises types of intentional activities carried out by at least one actor causing or changing phenomena or states of affairs on the social, political, financial, cultural and intellectual level. .

Thesauri

For illustration: Scope notes

iii) **Functions:** This hierarchy comprises types of activities that are structural parts of a relatively stable complex system of permanent and self-contained procedures that repeat within this system and thus contribute to its preservation. Although functions are part of a wider system, each function is completely distinct from the rest. As structural parts of a complex system, functions are types of actions that play a certain role within a system and aim at a specific goal, which they must accomplish.

In this respect it is not possible that the purpose which a certain function has to achieve be different from that for which the function is performed. In other words, the purpose of a function is one of its identity criteria.

Consequently, the notion of the function univocally relates the actions performed and the target achieved by these actions in such a way that, if some other target is achieved due to external factors, we speak of a different function or activity.

Thesauri

Conclusion

- ***Our goals are***
  - to characterize NOT to analyse the existing knowledge.
  - to design a consistent, stable and highly expressive set of fundamental concepts, that will enable humanities experts to find adequate generalizations.
  - to ensure interoperability between the thesauri already developed in specific scientific fields of the humanities within the Dariah project.
  - to facilitate users with their research inquiries.
  - To avoid the methodological errors that will lead to inconsistencies and incompatibilities between the terms.
  - To achieve the greatest economy in the process of organizing terms.

Our proposal is to construct a backbone thesaurus based on faceted classification!

Thesauri

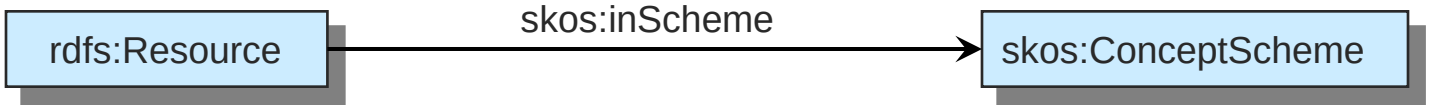
DISCUSSION

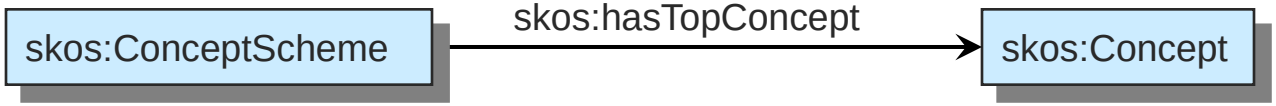
Classes

- skos:Concept
- skos:ConceptScheme
- skos:collection
- skos:OrderedCollection (sub-class of skos:collection)

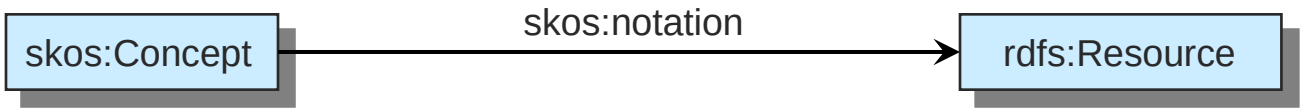
Interthesaurus relations

Concept scheme properties



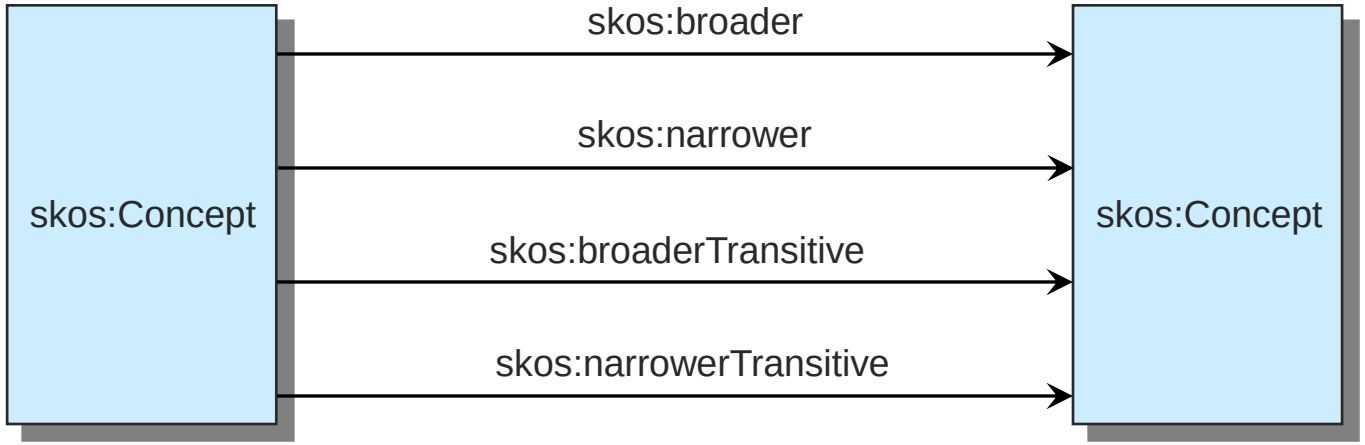


Notation property



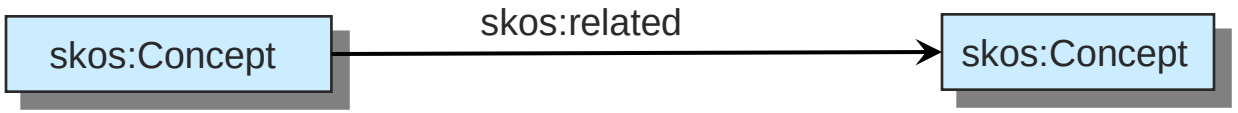
Interthesaurus relations

Hierarchical Relations

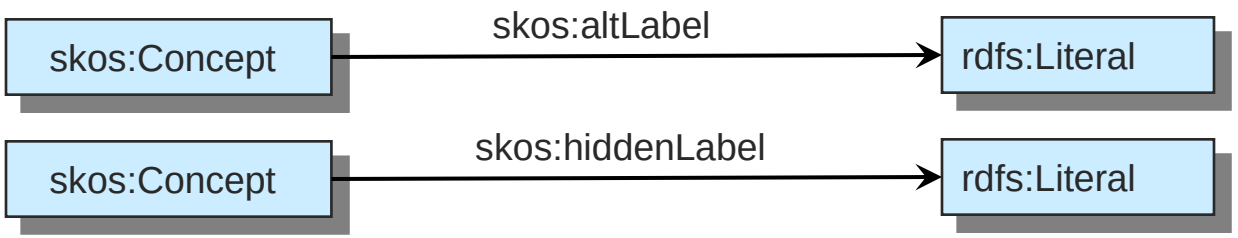


Interthesaurus relations

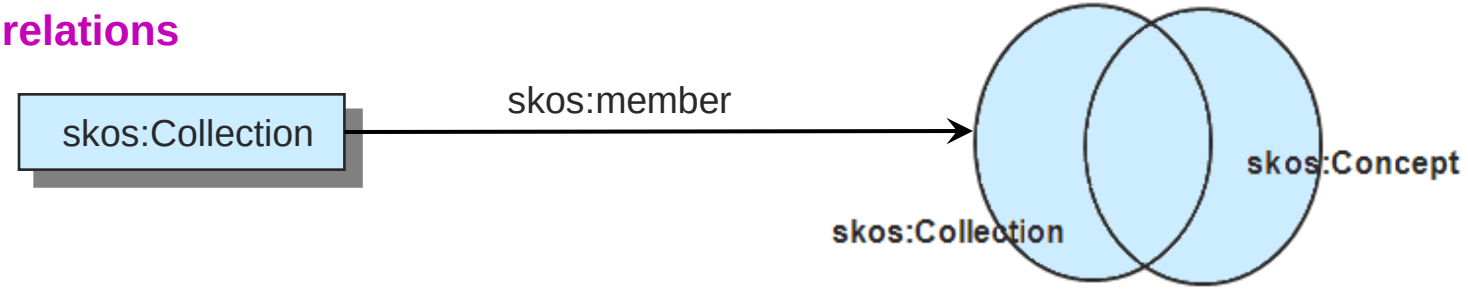
Associative Relations



Equivalence Relations

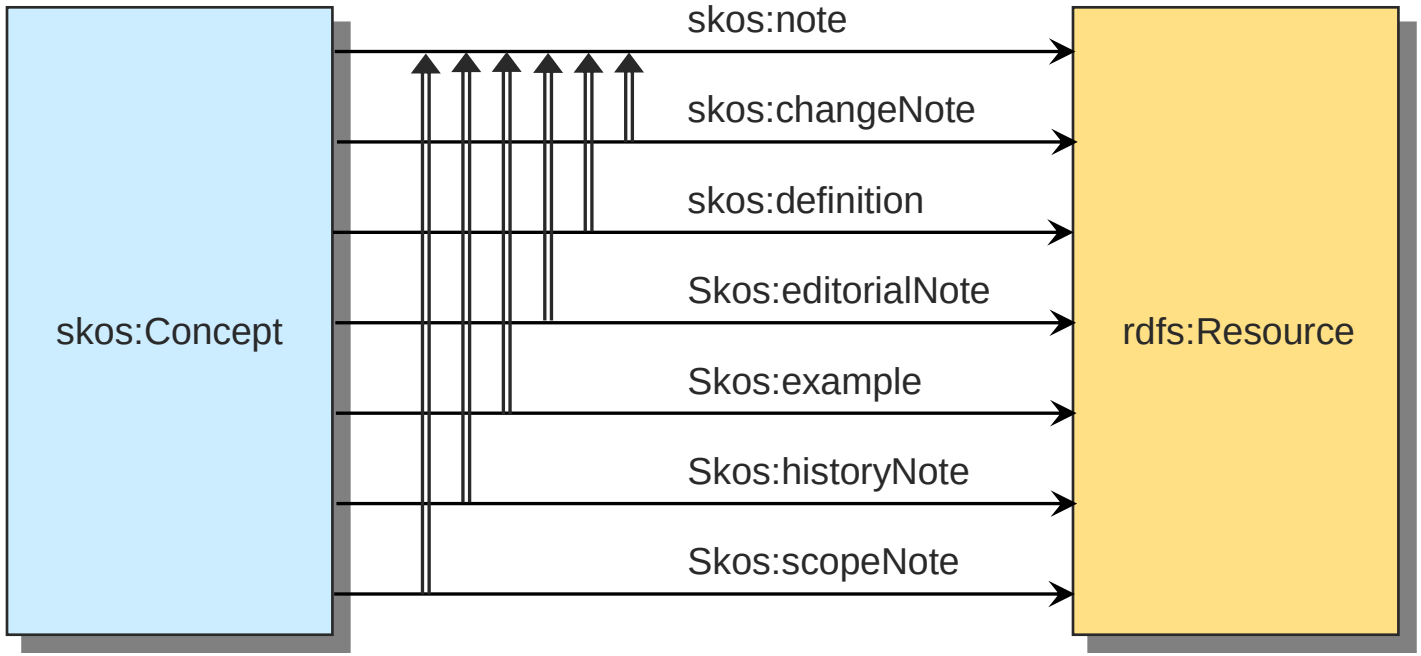


Grouping relations



Interthesaurus relations

Documentation properties



Intrathesaurus relations

Mapping properties