KDI
DERA Methodology

Fausto Giunchiglia and Mattia Fumagallli

University of Trento

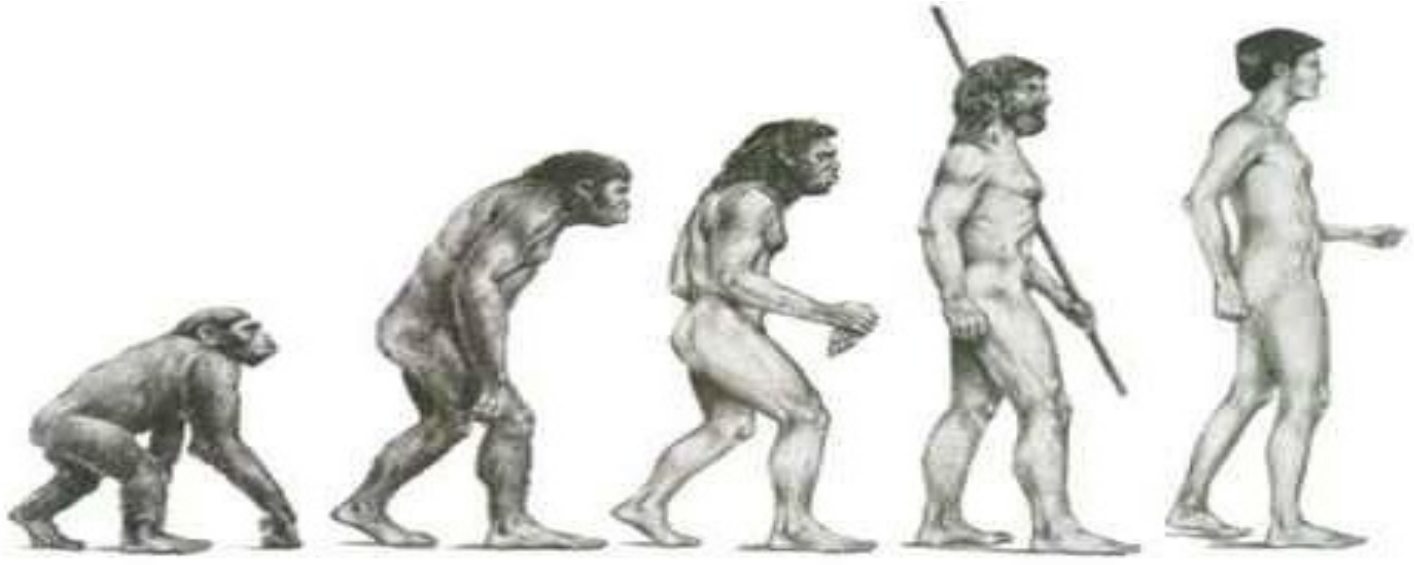
Methodologies for content generation

Methodologies for content generation

Roadmap

- Introduction
  - Motivation
  - The original faceted approach
- Primitive notions in DERA
- Steps in the methodology
- Guiding principles
- Converting DERA ontologies into DL
- Applications
- Exercises

WHY DO WE NEED A METHODOLOGY?



Humans and chimps share a surprising 98.8 percent of their DNA.

How to build ontologies which are of the highest quality possible?

Methodologies to ontology development

- Several methodologies have been developed for the construction and maintenance of ontologies (KR) or controlled vocabularies (KO)
- The faceted approach **[Ranganathan, 1967]** from library science is known to have great benefits in terms of quality and scalability
- It is based on the fundamental notions of *domain* and *facets*, which allow capturing the different aspects of a domain and allow for an incremental growth.
- Originally facets were of 5 types (PMEST): Personality, Matter, Energy, Space, Time.
- A key feature is compositionality (meccano property),
  i.e. the system allows a subject to be constructed by freely combining some basic components (facets).

[D] Medicine

[E] Body Part
. Digestive System
. . Stomach

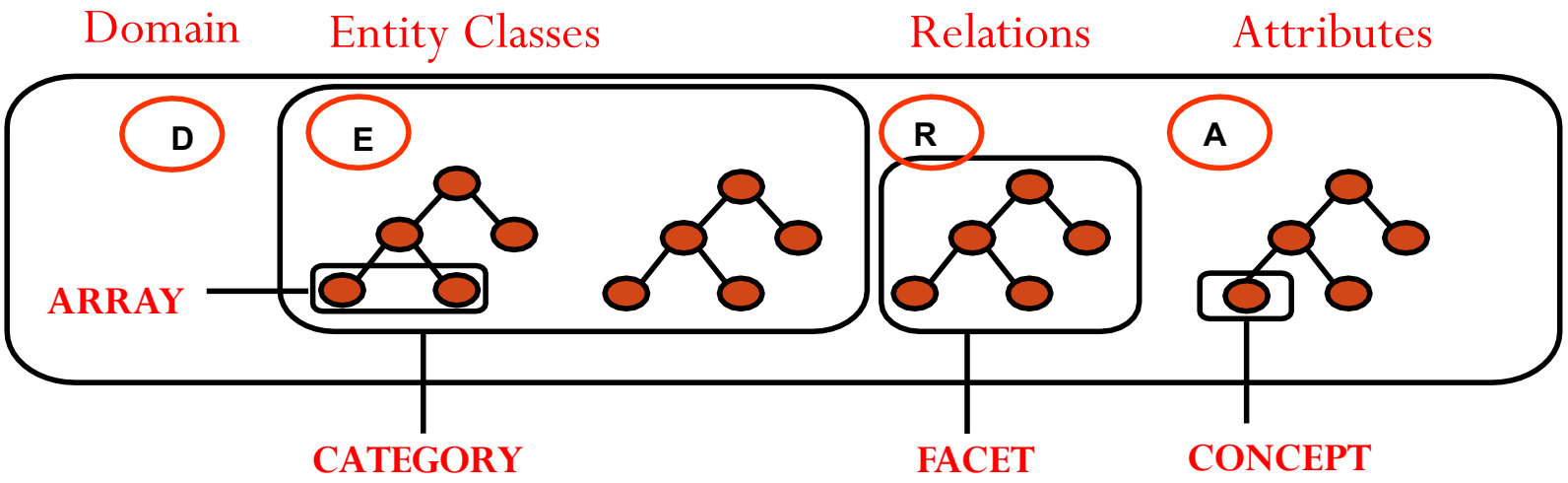
[P] Disease
. Cancer
. . Carcinoma
. . . Adenocarcinoma

[A] Action
. Treatment

[M] Kind (to be applied to [A] Action)
. Chemotherapy

The DERA framework

- To capture terminology relevant to a specific domain
- DERA is **faceted** as it is inspired to the faceted approach
- DERA is a **KR approach** as it models entities of a domain (D) by their
  entity classes (E), relations (R) and attributes (A)
- Terminology can be directly codified into Description Logic



Domains

- Any area of knowledge or field of study that we are interested in or that we are communicating about that deals with specific kinds of entities:

- Domains are the main means by which the ***diversity of the world*** is captured, in terms of language, knowledge and personal experience.

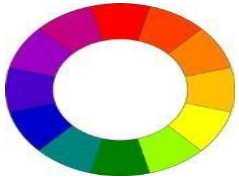
Primitive notions

- **Entity**: a (digital) description of any real world physical or abstract object so important to be denoted with a proper name. A single person, a place or an organization are all examples of entities.

- **Entity Class:** any set of objects with common characteristics.

- **Relation**: any object property used to connect two entities. Typical examples of relations include part-of, friend-of and affiliated-to.

- **Attribute**: any data property of an entity. Each attribute has a name and one or more values taken from a range of possible values.



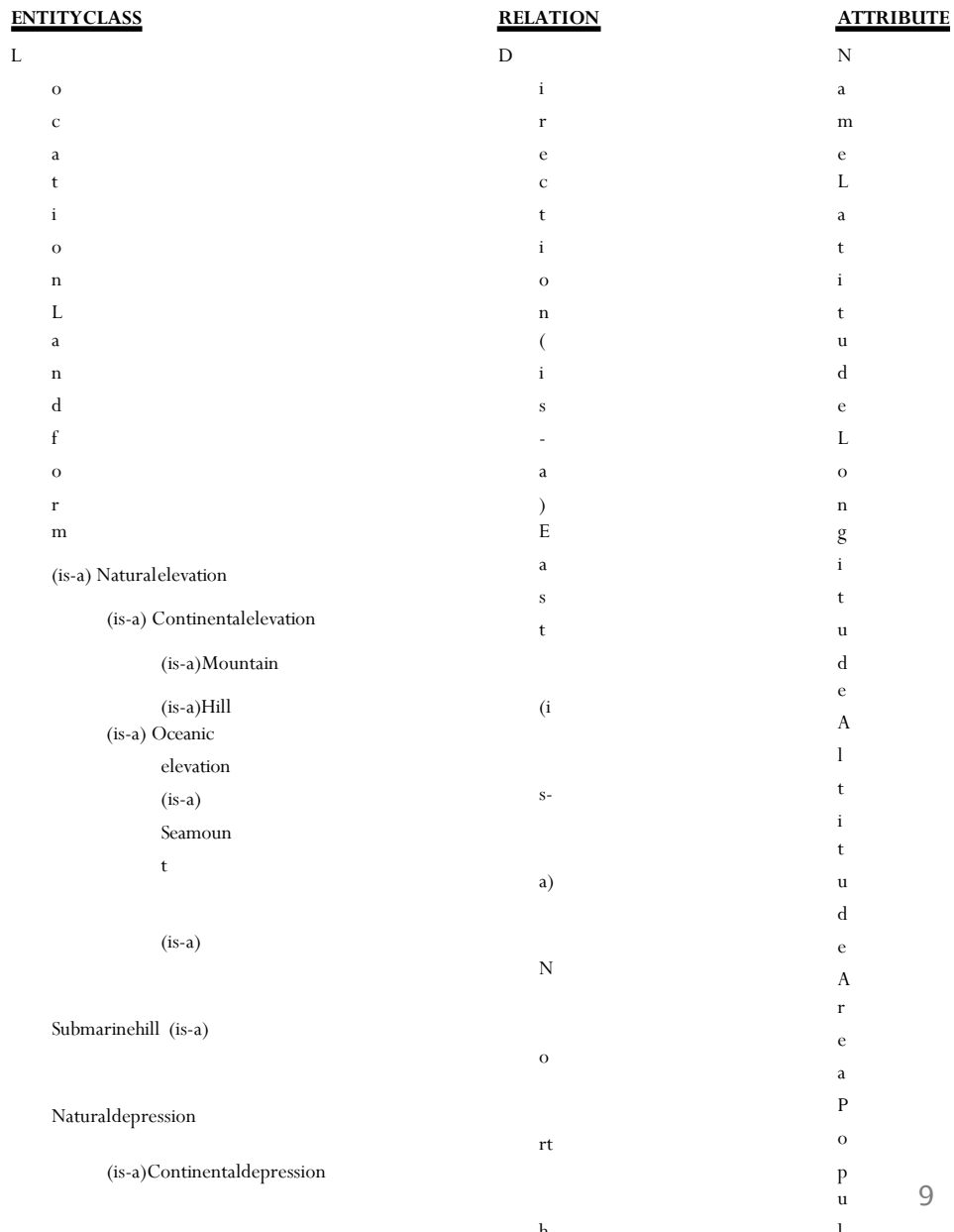
Elements of DERA

A DERA domain is a triple D = <E, R, A> where:

- **E** (for Entity) is a set of facets grouping terms denoting entity classes, whose instances (the entities) have either perceptual or conceptual existence. Terms in these hierarchies are explicitly connected by is-a or part-of relation.
- **R** (for Relation) is a set of facets grouping terms denoting relations between entities. Terms in these hierarchies are connected by is-a relation.
- **A** (for Attribute) is a set of facets grouping terms denoting qualitative/quantitative or descriptive attributes of the entities. We differentiate between attribute names and attribute values such that each attribute name is associated corresponding values. Attribute names are connected by is-a relation, while attribute values are connected to corresponding attribute names by value-of relations.

DERA facets

- DERA provides the language required to describe entities of a certain entity type in a given domain (D)
- Language comprises entity classes (E), relations (R) and attributes (A), names and values.
- Concepts and semantic relations between them form hierarchies of homogeneous nature called facets, each of them codifying a different aspect of the domain.
- Each facet is a descriptive ontology [Giunchiglia et al., 2014]



Analysis of the term "school"

Term:School			
Source	**Definition**	**Genus**	**Differentia**
WordNet	an educational institution	institution	educational
Oxford dictionary	an institution for educating children	institution	for educating children
Merriam-Webster	an institution for the teaching of children	institution	for the teaching of children
Wikipedia	an institution designed for the teaching of students (or "pupils") under the direction of teachers	institution	for the teaching of students

The term school is in general highly polysemous. Among others, school may denote a building. In the context of educational organizations, as from above, it seems there is quite an agreement about the fact that it indicates a kind of educational institution, but in some cases (such as fore WordNet) the meaning is left very generic. We coined the following definition: *"an educational institution designed for the teaching of students under the direction of teachers"*.

Synthesis of educational organizations

Educational Institution

<by level of complexity>

Preschool School

Primary school Secondary school Post-secondary school

<by programme orientation>

Training school

Vocational school

Technical school

Graduate school

College

University

Synthesis of educational organizations

Educational Institution (*an institution dedicated to education*)

Preschool (*an educational institution for children too young for primary school*)

School (*an educational institution designed for the teaching of students under the direction of teachers*)

Primary school (*a school for children where they receive the first stage of basic education*)

Secondary school (*a school for students intermediate between primary school and tertiary school*)

Tertiary school (*a school where programmes are largely theory based and designed to provide sufficient qualification for entry to advanced research programmes or professions with high skill requirements and leading to a degree*)

Training school (*a tertiary school providing theoretical and practical training on a specific topic or leading to certain degree*)

Vocational school (*a tertiary school where students are given education and training which prepares for direct entry, without further training, into specific occupation*)

Technical school (*a tertiary school where students learn about technical skills required for a certain job*)

Graduate school (*a tertiary school in a university or independent offering study leading to degrees beyond the bachelor's degree*)

College (*an educational institution or a constituent part of a university or independent institution, providing higher education or specialized professional training*)

University (*an educational institution of higher education and research which grants academic degrees in a variety of subjects and provides both undergraduate education and postgraduate education*)

Guiding principles

Principle	Example
Relevance	breed is more realistic to classify the universe of cows instead of bygrade
Ascertainability	flowing body ofwater
Permanence	spring as a natural flow of groundwater
Exhaustiveness	to classify the universe of people, we need both male and female
Exclusiveness	age and date of birth, both produce the samedivisions
Context	bank, a bank of a river, OR, a building of a financial institution
Currency	metro station vs. subwaystation
Reticence	minority author, blackman
Ordering	stream preferred to watercourse

Guidelines for the formal language

- **Concepts:** facets in UKC are descriptive ontologies where each concept denotes a set of real world entities (classes) or a property of real world entities (relations and attributes).
- **Look for essential concepts:** a property of an entity (that we codify as a concept) is essential (as opposite of accidental) to that entity if it must hold for it. As special form of essence, a property is rigid if it is essential to all its instances [Guarino and Welty, 2002].
- **Avoid complex concepts:** e.g. "red car".
- **Avoid redundancies:** e.g. "nursery school" and "kindergarten" are synonyms
- **Avoid individuals:** e.g. "United States military academy"
- **Pay attention to meronymy relations:** while part-of is assumed to be transitive in general, substance-of and member-of are not. Therefore, the latter two cannot be considered as hierarchical. In fact, [Varzi, 2006] describes some of the paradoxes that would be generated in assuming otherwise.

Guidelines for the natural language (I)

- **Terms and synsets:** terms are grouped into synsets. In UKC multiple languages are accounted for by developing multiple dictionaries, i.e. by assigning either a synset or a GAP to every concept.
- **Lemmas:** for the selection of terms we focus on lemmas.
- We do not accept in UKC:
  - **articles** (e.g. the) and **plural forms**;
  - **capitalization**, except for cases such as acronyms and abbreviations;
  - **punctuation characters and parenthesis**;
- The following are instead accepted, but not recommended:
  - **loan terms**, i.e. terms borrowed from other languages, if widely used. For instance, the term kindergarten in English is typically well accepted.
  - **transliterations**, i.e. when a terms is a transcript from one alphabet to another one.

Guidelines for the natural language (II)

- ***Parts of speech:*** noun, adjective, adverb and verb. A lemma can be a single word (e.g. bank), a multi-word (e.g. traffic light) or a prepositional phrase (e.g. place of warship).
- ***Homographs:*** terms which are spelled the same, but have different meaning. The same term can be associated to multiple concepts.
- ***Glosses:*** in line with principle of reticence, a gloss should not convey any cultural, temporal or regional bias.

Primary school: a school for young children; usually the first 6 or 8 grades **Infant school:** British school for children aged 5-7 **Junior school:** British school for children aged 7-11	NO
Primary school: a school for children where they receive the first stage of basic education **Infant school:** a primary school for very young children where they learn basic reading and writing skills **Junior school:** a primary school for young children where they learn basic notions of core subjects such as math, history and other social sciences	YES

Back to entities

Entity Class

Attributes

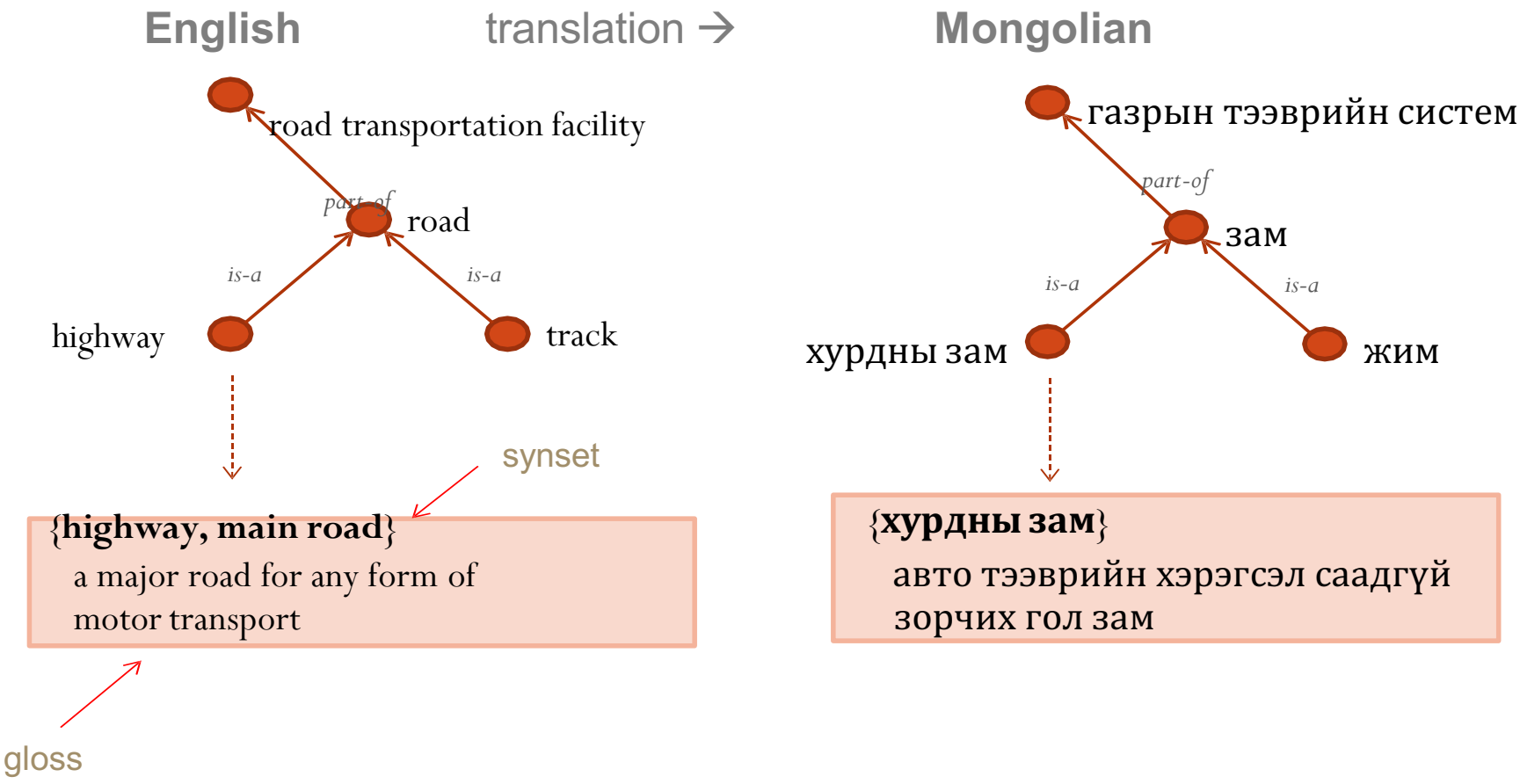
Relations

Class:	River
Name:	Thames
Latitude:	51.50
Longitude:	0.61
Length:	346 km (long)
Intersect:	UK

Thames

Each of the terms above comes from a DERA ontology in KB

Localization



Formalizing DERA into DL (I)

With the formalization, DL concepts denote either sets of entities or sets of attribute values. DL roles denote either relations or attributes.

A DL *interpretation* $I = <\Delta, I>$ consists of the *domain of interpretation* $\Delta = F \cup G$ where:

- F is a set of individuals denoting real world *entities*
- G is a set of *attribute values*

$$E_{i_I} \subseteq F \qquad R_{j_I} \subseteq F \times F \qquad A_{k_I} \subseteq F \times G \qquad v_{r_I} \in G$$

and of an interpretation function I where:

Formalizing DERA into DL (II)

	Object	DL formalization	
$E_1, \ldots, E_p$	entity classes	Concepts	TBox
$R_1, \ldots, R_q$	relations between classes	Roles	
$A_1, \ldots, A_s$	Attributes	Roles	
value-of	hierarchical relation	role restrictions	
is-a	hierarchical relation	subsumption ($\sqsubseteq$)	
part-of	hierarchical relation	Roles	
any other relation	associative relations	Roles	
$e_1, \ldots, e_n$	entities instances	individuals in F (entities)	ABox
$v_1, \ldots, v_r$	attribute values	individuals in G (values)	

Advantages of DERA

- DERA facets have **explicit semantics** and are modeled as descriptive ontologies

- DERA facets inherits all the important properties of the faceted approach, such as robustness and scalability

- DERA allows for **automated reasoning** via the formalization into Description Logics ontologies. In particular, DERA allows for a very expressive search by any entity property

The space ontology

- Knowledge is extracted from GeoNames and the Getty Thesaurus of Geographic Names
- Terms are collected, categorized into classes, entities, relations and attributes, and synsets are generated
- Synsets are mapped to and integrated with WordNet
- Synsets are analyzed and arranged into facets
- Terms are standardized and ordered

Objects	Quantity
Entity classes (E)	845
Entities (e)	6,907,417
Relations (R)	70

Landform
- Natural depression
  - Oceanic depression
    - Oceanic valley
    - Oceanic trough
  - Continental depression
    - Trough
    - Valley
- Natural elevation
  - Oceanic elevation
    - Seamount
    - Submarine hill
  - Continental elevation
    - Hill
    - Mountain

Body of water
- Flowing body of water
  - Stream
    - River
    - Brook
- Stagnant body of water
  - Lake
  - Pond

The semantic-geo catalogue

- Knowledge is extracted from the geographical dataset of the Province of Trento
- The faceted ontology was built in English and Italian
- Usage of the ontology
  - The ontology is used in combination with S-Match within the search component of the geo-catalogue to improve search
  - The evaluation shows that at the price of a drop in precision of 0.16% we double recall

Objects	Quantity
Facets	5
Entity classes (E)	39
Entities (e)	20,162
	20,161

Body of water
- Lake
- Group of lakes
- Stream
- River
  - Rivulet
- Spring
- Waterfall
- Cascade
- Canal

Natural elevation
- Highland Hill
- Mountain
- Mountain range
  - Peak
- Chain of peaks
- Glacier

Natural depression
- Valley Mountain pass

Exercises

1. Analyse the following terms:
   - (geography) river, lake, salt lake, depth
   - (business) organization, company, business
   - (literature) newspaper, newsletter, book, archive, author, publisher, format, frequency
2. Take one domain of your choice, identify the entity types which are relevant and define corresponding terminology using DERA (concentrate on a few classes, relations and attributes).

Reference material

[Ranganathan, 1967] S. R. Ranganathan, Prolegomena to library classification, Asia Publishing House.

[Gruber, 1993] A translation approach to portable ontology specifications. Knowledge Aquisition, 5 (2), 199–220.

[Pollock, 2002] Integration's Dirty Little Secret: It's a Matter of Semantics. Whitepaper, The Interoperability Company.

[Guarino and Welty, 2002] Guarino, N., Welty, C. (2002). Evaluating ontological decisions with OntoClean. Communications of the ACM, 45(2), 61-65.

[Uschold and Gruninger, 2004] Ontologies and semantics for seamless connectivity. SIGMOD Rec., 33(4), 58–64.

[Varzi, 2006] Varzi, A. (2006). A note on the transitivity of parthood. Applied Ontology, 1 (2), 141-146.

[Giunchiglia et al., 2009] Faceted Lightweight Ontologies. In: Conceptual Modeling: Foundations and Applications, LNCS Springer.

[Giunchiglia et al., 2012a] A facet-based methodology for the construction of a large-scale geospatial ontology. Journal on Data Semantics, 1 (1), pp. 57-73.

[Giunchiglia et al., 2012b] Domains and context: first steps towards managing diversity in knowledge. Journal of Web Semantics, special issue on Reasoning with Context in the Semantic Web.

[Giunchiglia et al., 2014] From Knowledge Organization to Knowledge Representation. Knowledge Organization. 41(1), 44-56.

[Tawfik et al., 2014] A Collaborative Platform for Multilingual Ontology Development. International Conference on Knowledge Engineering and Ontology.

[Ganbold et. al., 2014] An Experiment in Managing Language Diversity Across cultures. eKNOW 2014