Aalto University
School of Science

SeCo

Semantic Web
An Introduction

CS-E4410 Semantic Web, 13.1.2021

Eero Hyvönen
Aalto University, Semantic Computing Research Group (SeCo) *http://seco.cs.aalto.fi*
University of Helsinki, HELDIG *http://heldig.fi*

eero.hyvonen@aalto.fi

Outline

The idea of Semantic web

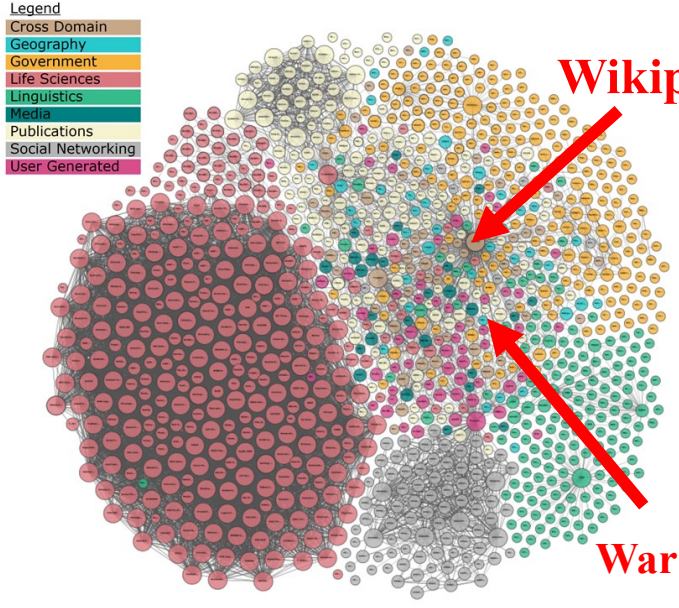
Core technological basis of Semantic web

- Metadata, ontologies, reasoning
- Review of the technological solutions and standards

Application domains

Linked Data & Semantic Web

Human Knowledge on the Web: ”Web of Data” for machines



Linked Open Data Stats 2018 (http://stats.lod2.eu/):
10 000 datasets,150 billion triples

Traditional Web ”Web of Pages” for humans

Google Index (2020):
55 billion pages (excluding hidden web)

Big Boys Have Entered the Game: Knowledge Graphs

http://schema.org

- Google Knowledge Graph
- Microsoft Satori
- IBM Watson
- eBay Products
- Facebook Open Graph
- …

Google

Microsoft

IBM

ebay

facebook

Why Linked (Open) Data?

- Enriching everybody's data collaboratively from separate silos
  - Everybody wins by collaboration!
- Creating Findable, Accessible, Interoperable, Re-usable data
  - The value of data increases!
- Creating more intelligent applications for the public, curators, and researchers
  - The machine "understands" linked data!

GO FAIR

https://www.go-fair.org/fair-principles/

Fundamental barrier for the development of the Web of Data: machine-”understandability”

The web contents are created for human readers

- HTML, PDF, JPEG, ...

Machine mediates and displays, but does not “understand” contents of the web

- E.g., a Finnish text article

A web service ≈ machine helps human

- Requires machine-”understandability” of the contents

➔ **A fundamental contradiction**

How can we build a more intelligent Web?

1. **Applications are programmed to be more intelligent**

- The contents stay as they are
- The machines operate more human-like (Artificial Intelligence)

2. **Contents are represented in a more intelligent way**

- The contents are easier to understand
- Machines stay more or less as they are

In practice, both ways are needed

- More intelligent systems process more intelligently represented contents

Approach 1: Develop more intelligent applications using AI

Automatic interpretation of natural language is difficult

- Free form of the documents
- Semantics of the content

Non-textual contents

- Pictures, sound, music, video, software, ...
- How to interpret algorithmically unstructured data?

More than the document itself is needed for interpretation

- Context + common sense needed
- Fundamental problems of Artificial Intelligence, easy for humans!
- Great scientific and technological challenges

Approach 2: Contents represented in a more intelligent way

The foundation of Semantic web

- The information is stored in a way that a machine understands it!
- Human helps the machine
  - *Machine can also help in this (user-friendly tools for semantic content creation)*

The development was boosted in the beginning of the 2000s

- W3C Semantic Web Activity 2001
- W3C Web Services Activity 2002

Web Generations

1G WWW: early 1990's

- WWW pages for human interpretation
- HTML language

2G WWW: late 1990's

- Structured web documents for human/machine interpretation
- XML-based languages

3G WWW: Semantic Web, Web of Data, 2000's

- Explicit meaning of documents for human/machine use
- RDF-based languages

⇒ **Semantics = new foundation for intelligent web services**

- Semantic = "understandable" to machines

Limitations of non-semantic web: case MuseumFinland

```
<artifact>
  <id>NBA:H26069:467</id>
  <target>cup and plate</target>
  <material>porcelain</material>
  <creationLocation>Germany</creationLocation>
  <creator>Meissen</creator>
</artifact>
```

- This metadata cannot answer the following questions:
  - Find all vessels?
  - Find all ceramic products?
  - Find artifacts manufactured in Europe?
  - Does the city of Meissen manufacture ceramics?

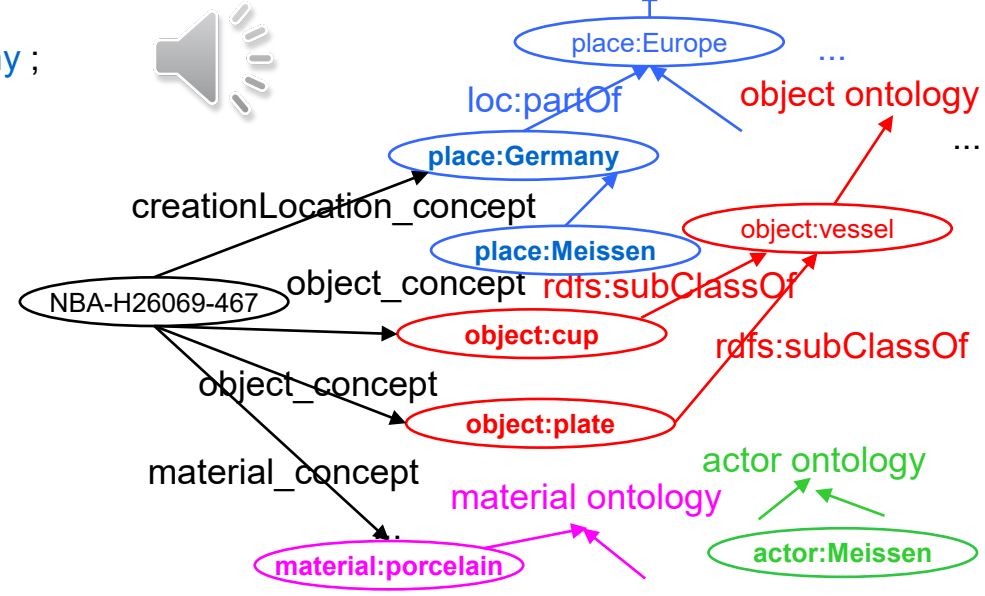
Semantic Web solution: ontologies

NBA-H26069-467
:object ”cup and plate” ;
:object_concept object:cup ;
:object_concept object:plate ;

:material ”porcelain” ;
:material_concept object:porcelain ;

:creationPlace ”Germany” ;
:creationPlace_concept place:Germany ;

:creator ”Meissen”
:creator_concept actor:Meissen .



Find all vessels?
Find all ceramic products?
Find artifacts manufactured in Europe?
Does the city of Meissen manufacture ceramics?

Case Rijksmuseum Amsterdam: CHIP Demonstrator

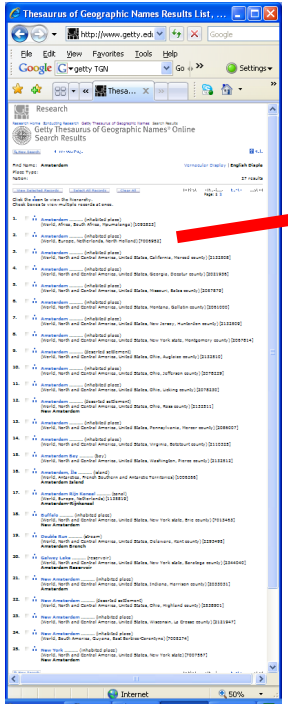
Example in Turtle notation

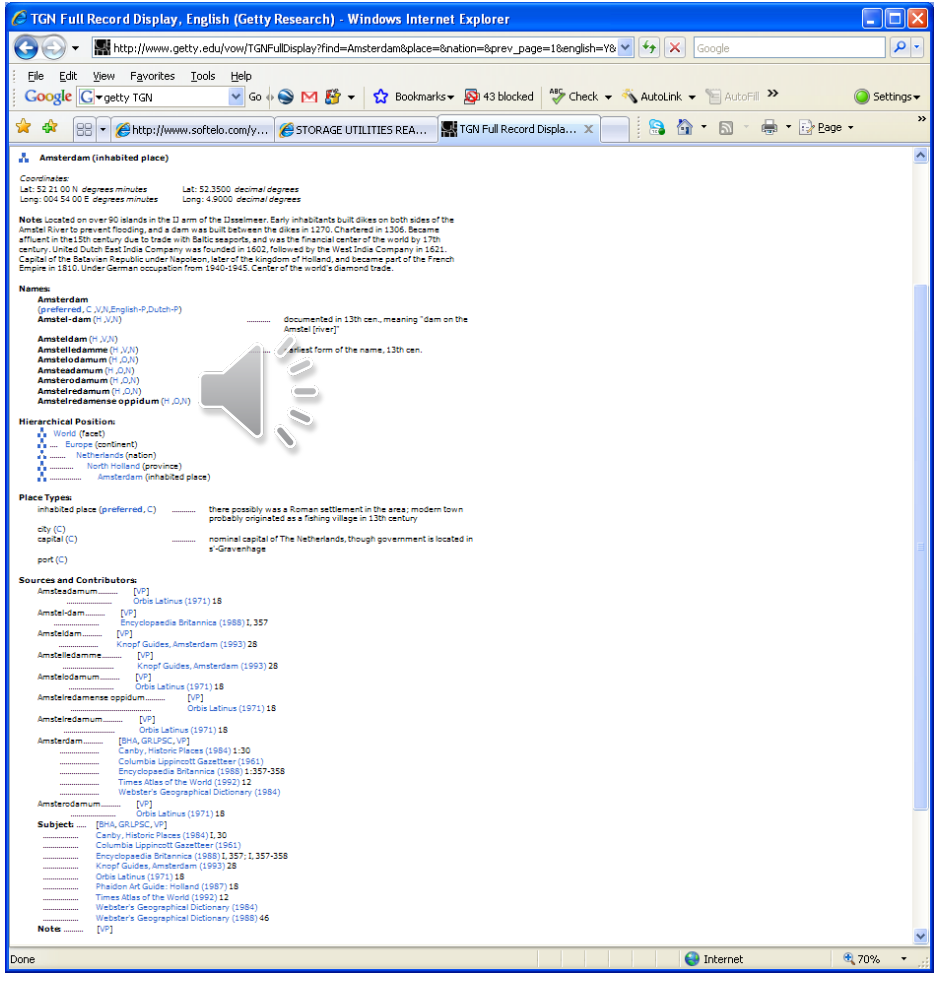
- VRA metadata schema (extension of Dublin Core)
- (Aroyo et al., 2007)

```
rijks:artefactSK-C-K
  vra:type vra:Work ;
  vra:title "The Night Watch" ;
  vra:date "1642" ;
  vra:creator: 500011051 ;          # Rembrandt
  vra:subject iconclass:45F31 ;     # Call to arms
  vra:culture tgn:7006952 ;         # Amsterdam
  vra:material aat:30015050 .       # Oil paint
```

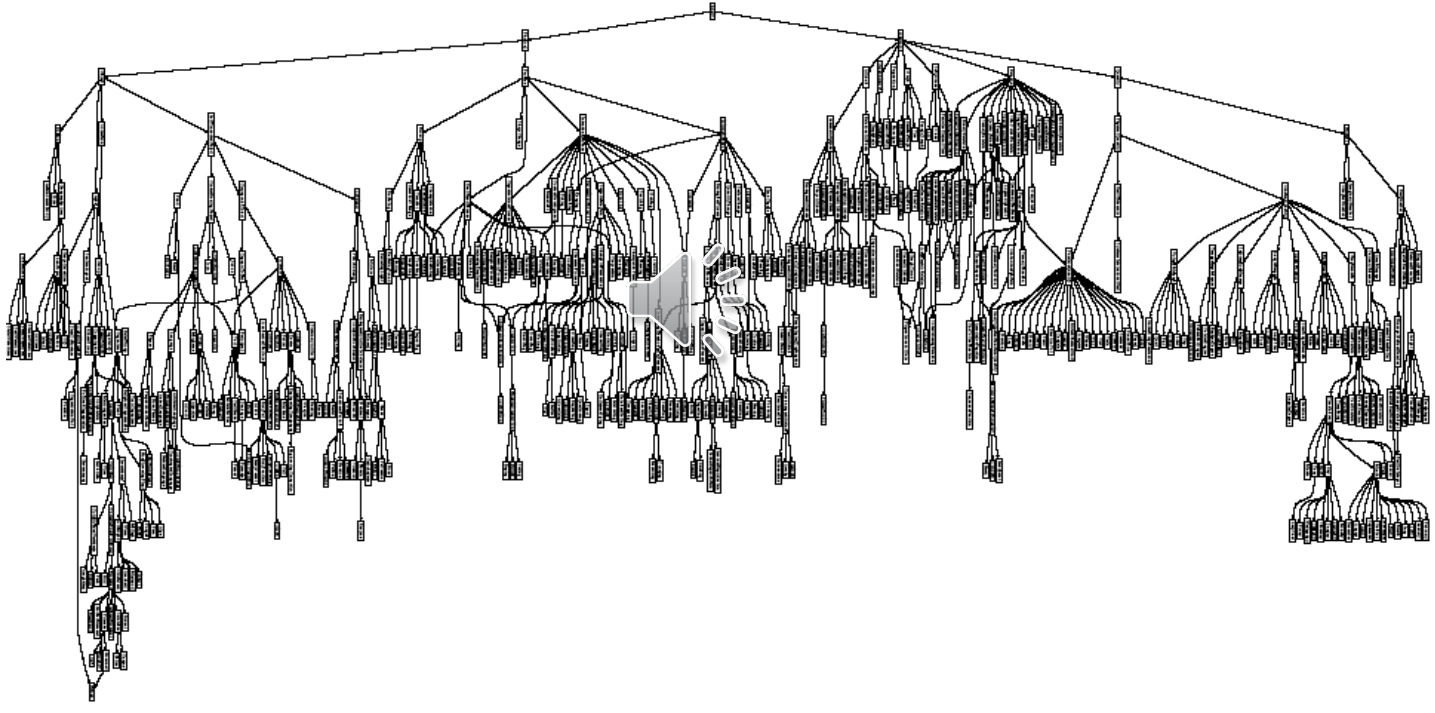
A resource in the TGN ontology / vocabulary

Amsterdam in TGN





An Ontology Concept Hierarchy: Standard Upper Merged Ontology SUMO



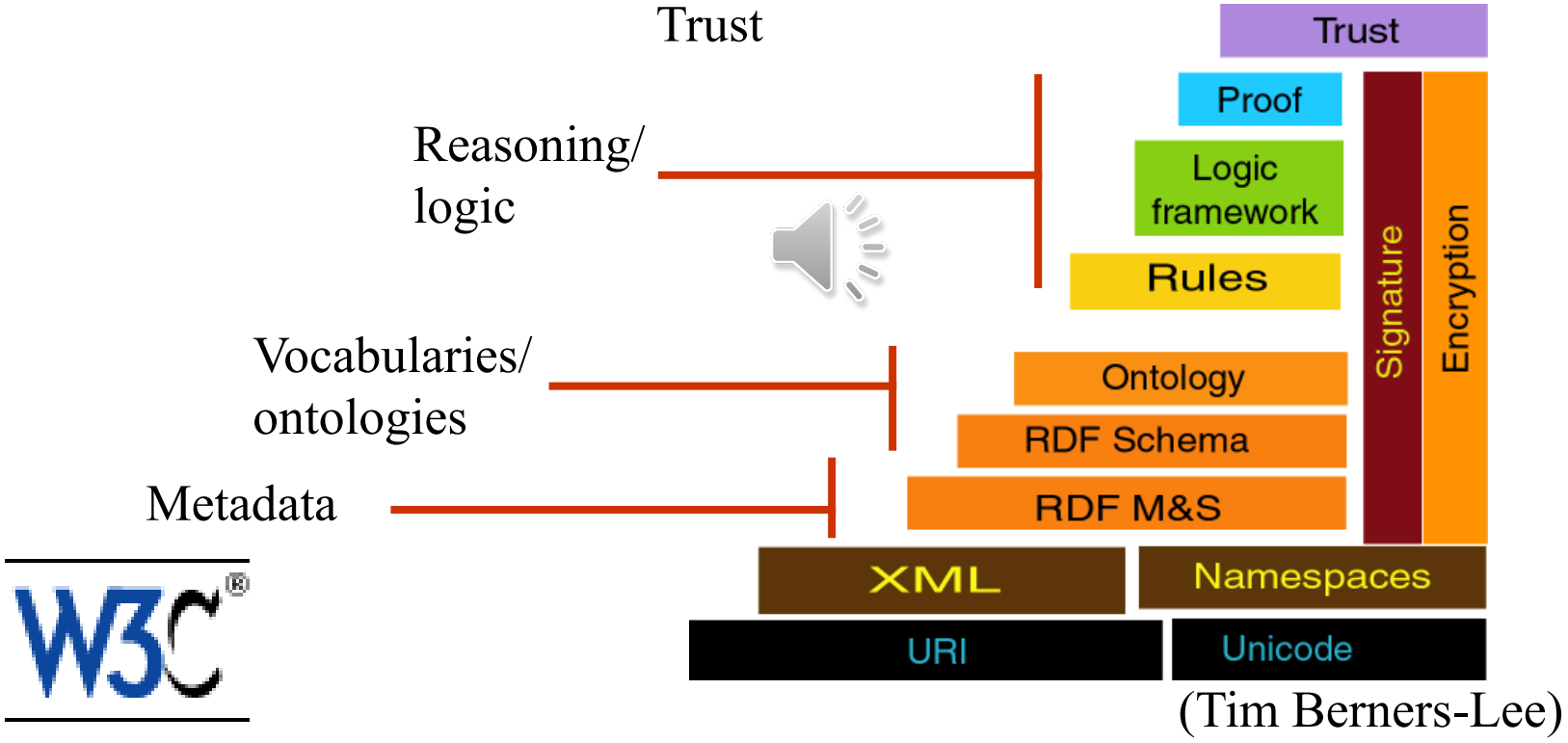
Technological basis of Semantic Web

Aalto University
School of Science

Department of
Computer Science

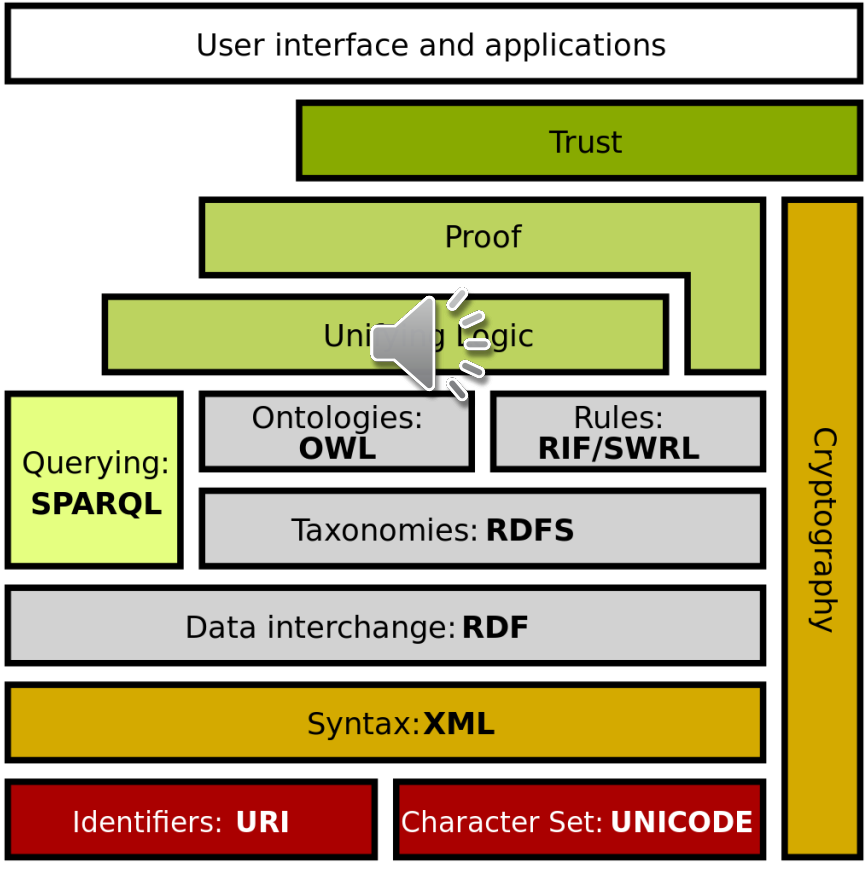
SeCo

The Original ”Layer Cake Model” of Semantic Web



(Tim Berners-Lee)

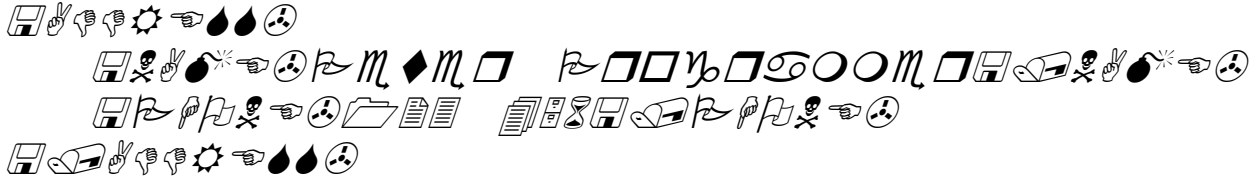
Newer Version of Layer Cake Model



Metadata Level

Why isn't XML alone sufficient for the basis of Semantic web?

- Interpretation of XML languages has to be defined in a domain-specific way
- Combining different XML languages is often difficult
- We need a markup language, whose interpretation is:
  - *Machine-"understandable"*
  - *Shared across different application domains*
  - *Commonly agreed*
- The semantics of XML is only in human brain

  - *<ADDRESS>*
    <NAME>Peter Programmer</NAME>
    <PHONE>123 456</PHONE>
    </ADDRESS>
  - 

The Semantic Web solution for Metadata: RDF Resource Description Framework

- General metadata description model and language for web resources
- Relational model, *not* a syntax (as opposed to XML)
  - *RDF description = directed (knowledge) graph*
- Semantics is defined based on logic
- RDF has syntaxes/serializations, too
  - *XML-based RDF/XML, especially for machines*
  - *Simpler notations (Turtle, N-triples, N3) for humans*
- Standardized and commonly used
  - *W3C draft 1999*
  - *W3C recommendation RDF 1.0, 10.2.2004*
  - *W3C recommendation RDF 1.1, 25.2.2014*

RDF Example

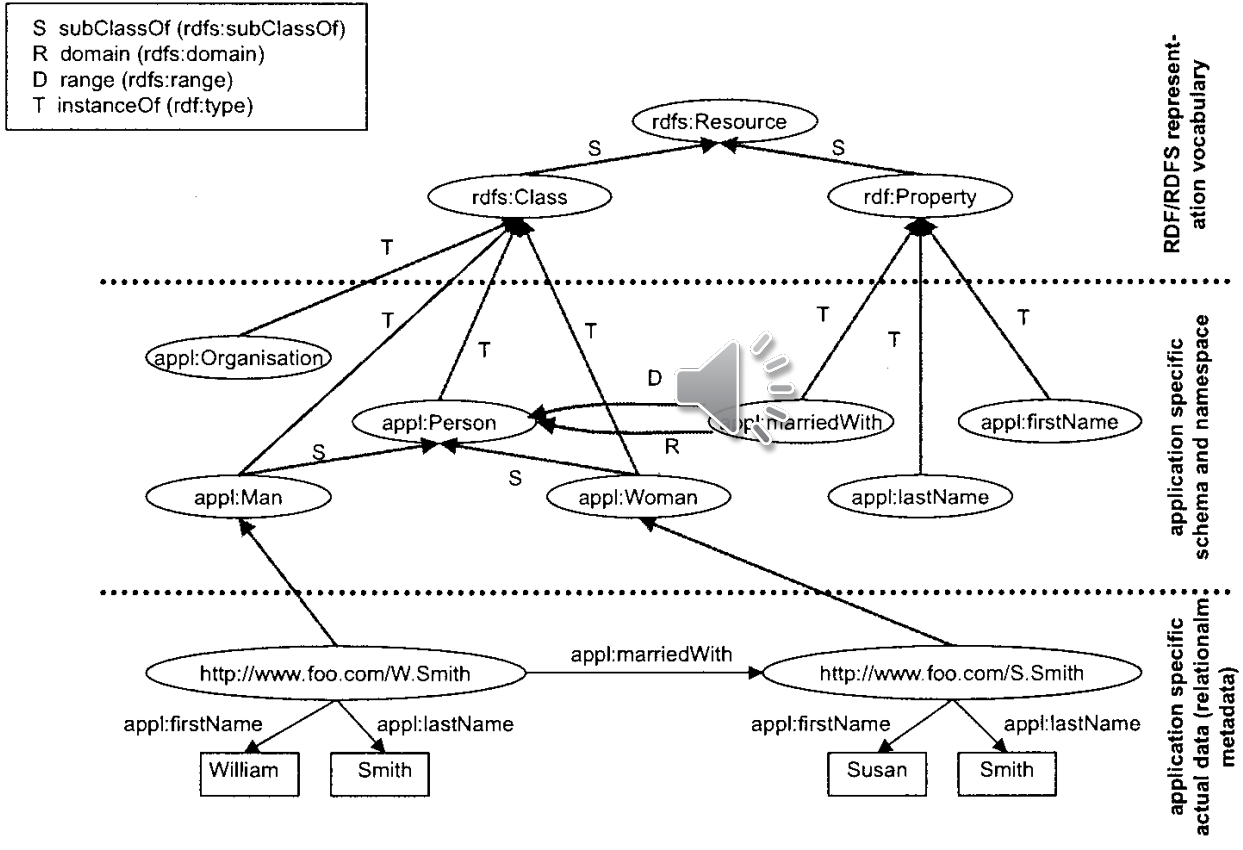


Figure 3.4. An RDF-Schema Example (Maedche, 2002)

Metadata Schemas

Standardized templates for representing (meta)data

- A set of elements (properties) describing object types
  - *E.g author, publisher, and publishing year of books*
- Value specifications for the properties describing individual objects
  - *Literal values for data (text, number, date, …)*
  - *URI identifier for related concepts/resources*

Different content types typically require different schemas (element sets)

- E.g. books, persons, paintings, places, …

Example: Dublin Core metadata schema for describing web documents

Set of 15 general properties for different content types

- Dublin Core Metadata Element Set (ISO Standard 15836)
  - *Title*
  - *Creator*
  - *Subject*
  - *Description*
  - *Publisher*
  - *Contributor*
  - *Data*
  - *Type*
  - *Format*
  - *Identifier*
  - *Relation*
  - *Source*
  - *Language*
  - *Coverage*
  - *Rights*

Metadata Schema in HealthFinland

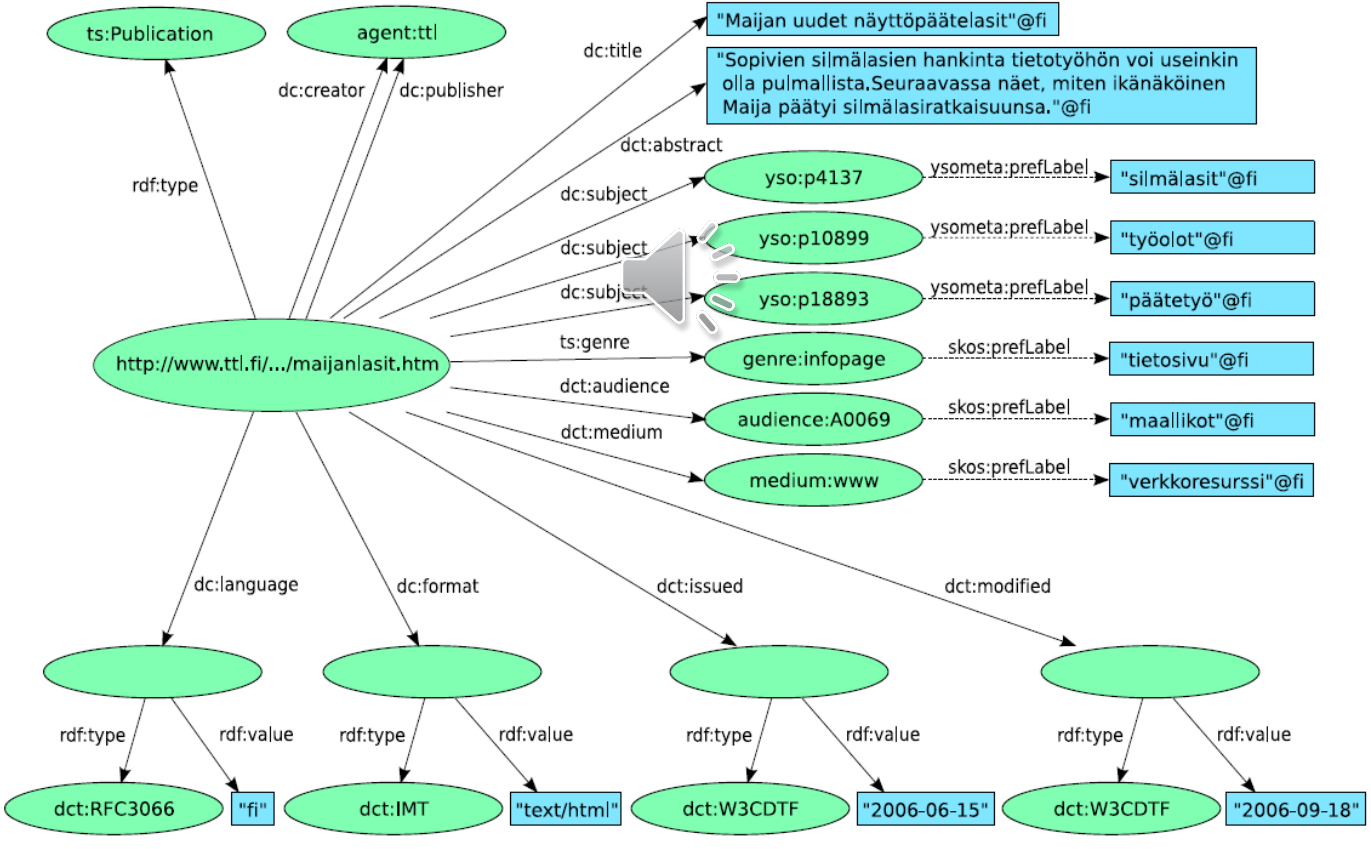
Table 1. HEALTHFINLAND Metadata Schema. Obligatory fields are marked in **bold**. Cardinalities are presented in the column C.

	Name	QName	C	Value type	Value range
General metadata	**Identifier**	**dc:identifier**	1	URI	
	Locator	ts:url	0..1	URL	
	Title	**dc:title**	1[a]	Free text	Non-empty string.
	Abstract	**dcterms:abstract**	1[a]	Free text	Non-empty string.
	Language	**dc:language**	1..*	String	RFC 3066
	Publication time	**dcterms:issued**	1	String	W3CDTF (ISO 8601)
	Acceptance time	dcterms:dateAccepted	0..*	String	W3CDTF (ISO 8601)
	Modification time	dcterms:modified	0..*	String	W3CDTF (ISO 8601)
	Publisher	**dc:publisher**	1..*	Instance	foaf:Organization
	Creator	dc:creator	0..*	Instance	foaf:Organization, foaf:Person or foaf:Group
Content classification	**Subject**	**dc:subject**	1..*	Concept	YSO, MeSH and HPMulti Ontologies
	Audience	**dcterms:audience**	1..*	Concept	Audience Ontology
	Genre	**ts:genre**	1..*	Concept	Genre Ontology
	Presentation type	**dc:type**	1..*	Concept	DCMI Type vocabulary
	Format	**dc:format**	1	String	IANA MIME types
	Medium	**dcterms:medium**	1	Concept	Medium Ontology
	Spatial coverage	dcterms:spatial	0..*	String or concept	DCMI Point, DCMI Box or Location Ontology
	Temporal coverage	dcterms:temporal	0..*	String or concept	W3CDTF, DCMI Period or Time Ontology
Relations	Part of	dcterms:isPartOf	0..*	Document	URI
	Rights	dc:rights	0..*	Free text or document	URI or textual description
	Source	dc:source	0..*	Free text or document	URI (e.g., ISBN) or bibliographical reference
	Reference	dcterms:references	0..*	Free text or document	URI (e.g., ISBN) or bibliographical reference
	Translation of	ts:isTranslationOf	0..*	Document	URI
	Format of	dcterms:isFormatOf	0..*	Document	URI

[a] Multilingual values are allowed, but only one value in each language.

HealthFinland portal:
Maija's eyeglasses – PDF document on the web

Maija's eyeglasses: metadata in RDF form



Ontology Level

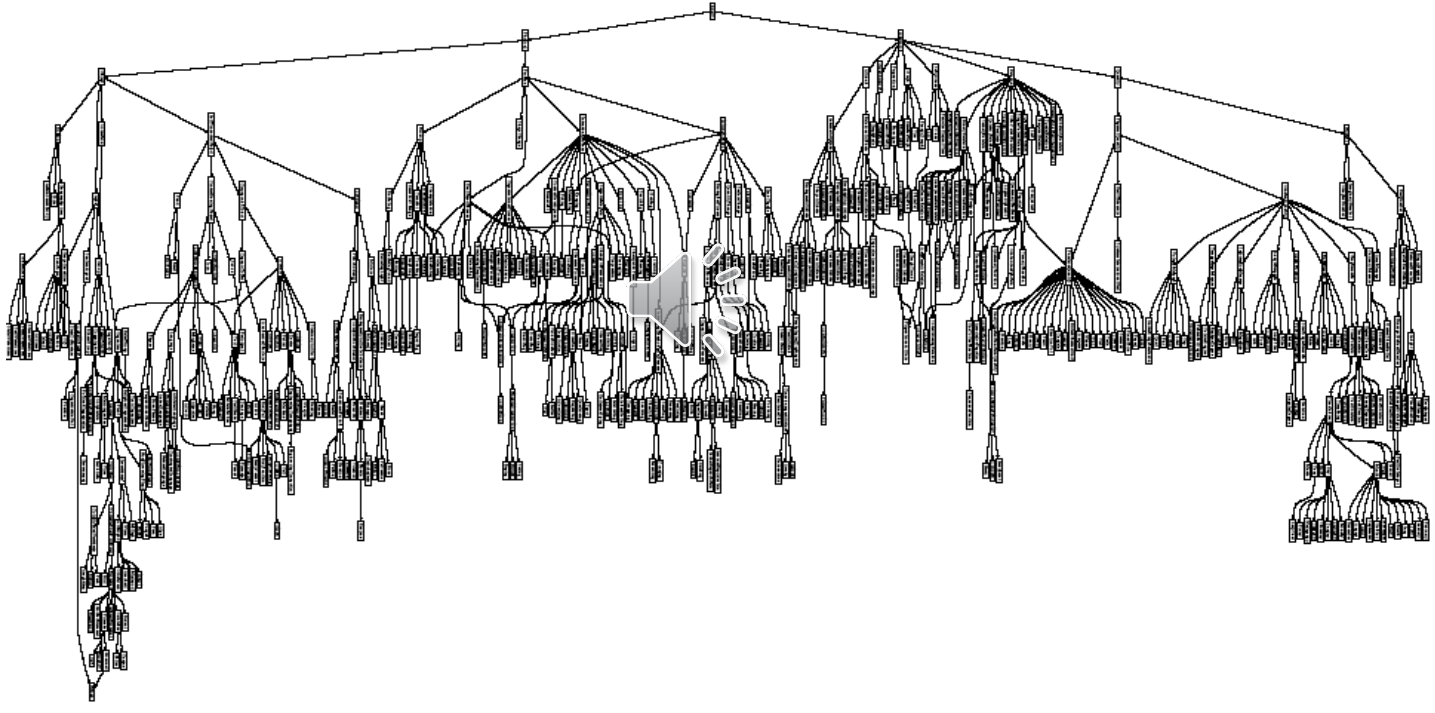
What is an ontology?

"An ontology is an explicit specification of a conceptualization ...definitions need to be couched in some common formalism" (Gruber, 1993)

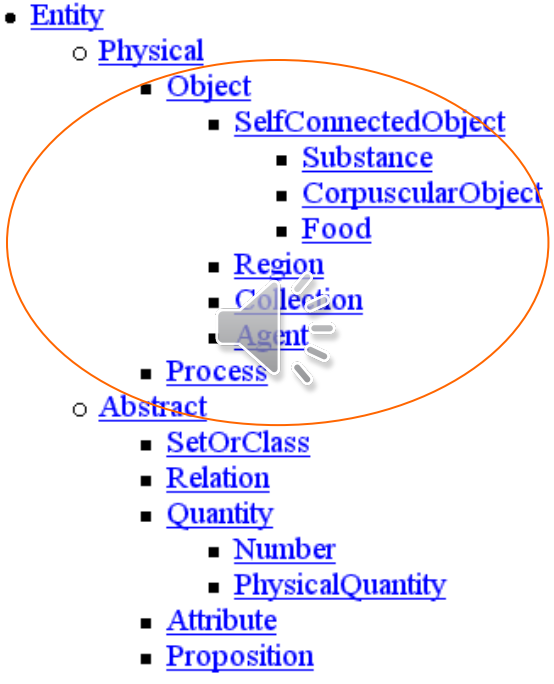
- *Explicit: machine can understand*
- *Formal: precisely defined*
- *Common (shared): communication is possible*

- **Defines the concepts/objects and their relations in a given application domain**
- **A first requirement for the humans and machines to understand each other**

Standard Upper Merged Ontology SUMO

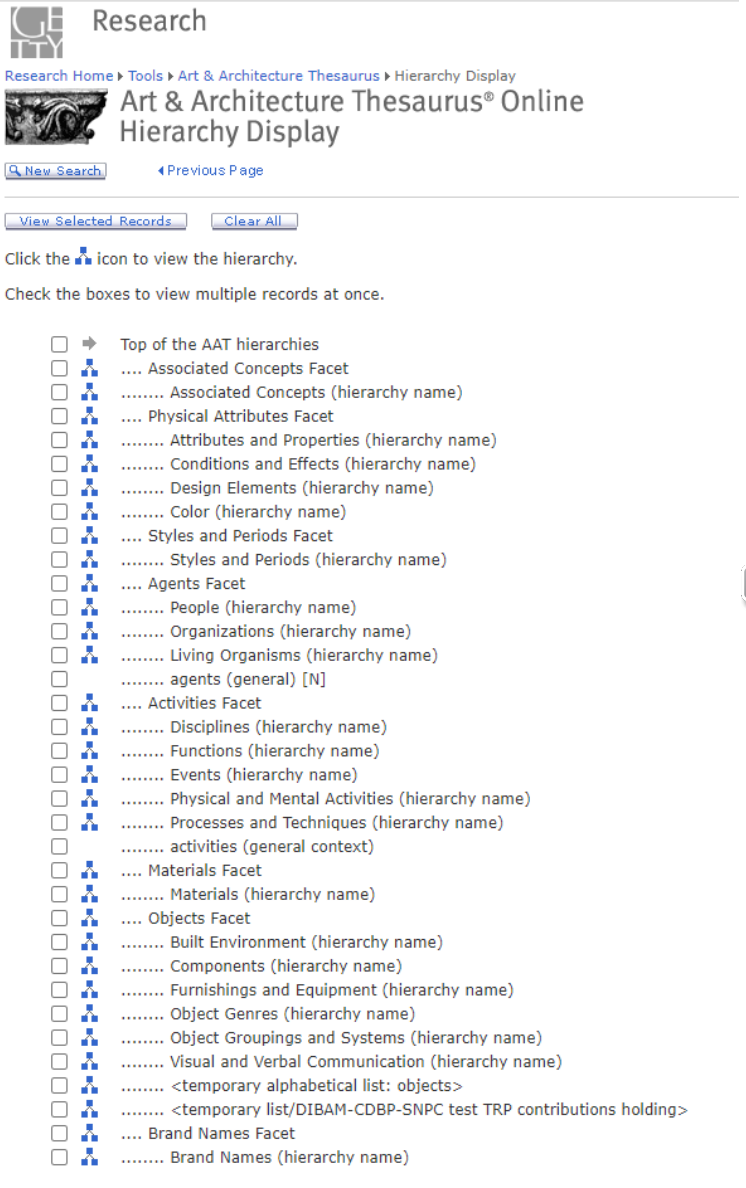


SUMO principal distinctions



SUMO Object:

- Object
  - SelfConnectedObject
    - Substance
      - PureSubstance
        - ElementalSubstance
          - Metal
          - Atom
          - SubatomicParticle
            - AtomicNucleus
            - Electron
            - Proton
            - Neutron
        - CompoundSubstance
          - Water
          - Molecule
      - Mixture
        - Solution
        - Mineral
        - BodySubstance
      - BiologicallyActiveSubstance
        - Nutrient
        - Hormone
    - CorpuscularObject
      - OrganicObject
        - Organism
        - AnatomicalStructure
      - Artifact
      - ContentBearingObject
    - Food
  - Region
    - GeographicArea
    - AstronomicalBody
    - Hole
  - Collection
    - Group
      - GroupOfPeople
      - Organization
  - Agent
    - Organism
    - Group
    - GeopoliticalAgent
    - SentientAgent



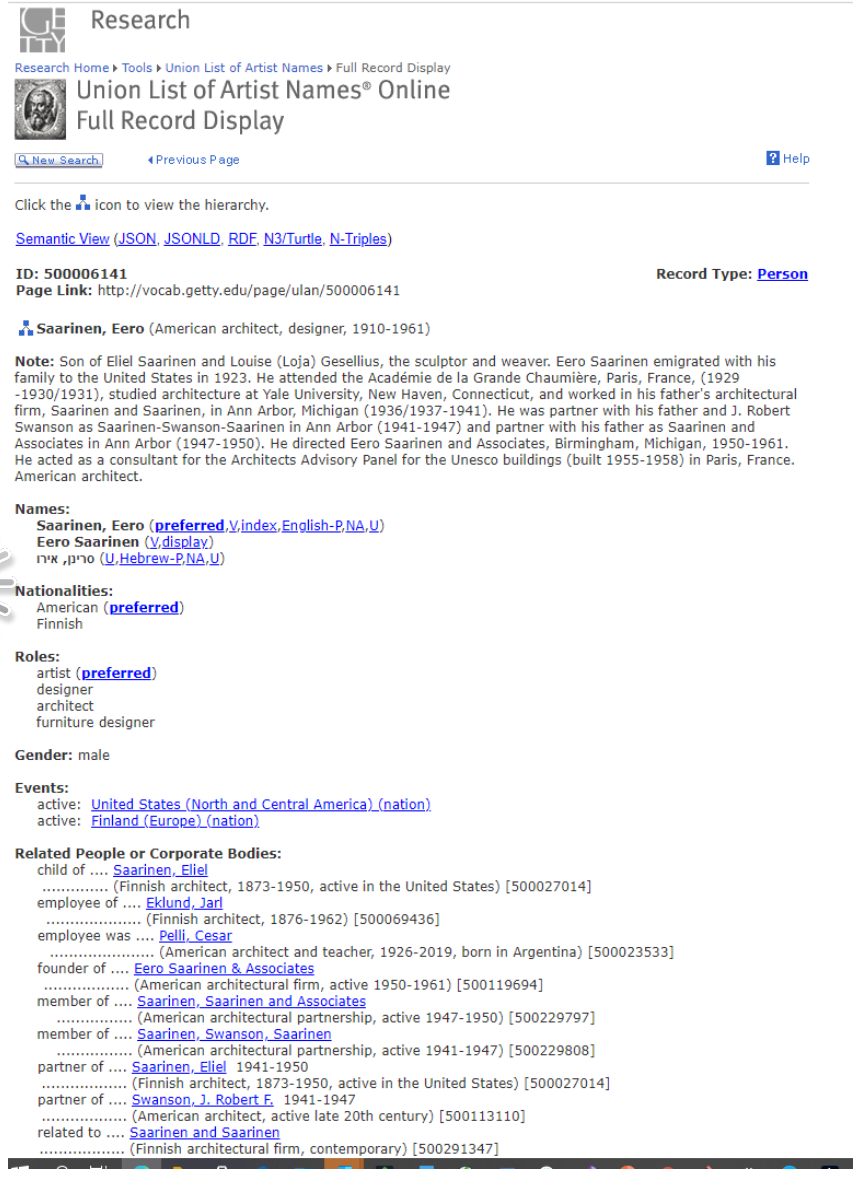
AAT Art & Architecture Thesaurus

- maintained by Getty Research Intstitute
- 7 main classes, 125 000 concepts

Universal List of Artist Names ULAN

- Over 300 000 artists with 720 000 names as Linked Open Data (2018)

Example: Eero Saarinen data



Research

Research Home ▸ Tools ▸ Union List of Artist Names ▸ Full Record Display

Union List of Artist Names® Online
Full Record Display

New Search | Previous Page | Help

Click the icon to view the hierarchy.

Semantic View (JSON, JSONLD, RDF, N3/Turtle, N-Triples)

ID: 500006141 **Record Type: Person**
Page Link: http://vocab.getty.edu/page/ulan/500006141

Saarinen, Eero (American architect, designer, 1910-1961)

Note: Son of Eliel Saarinen and Louise (Loja) Gesellius, the sculptor and weaver. Eero Saarinen emigrated with his family to the United States in 1923. He attended the Académie de la Grande Chaumière, Paris, France, (1929 -1930/1931), studied architecture at Yale University, New Haven, Connecticut, and worked in his father's architectural firm, Saarinen and Saarinen, in Ann Arbor, Michigan (1936/1937-1941). He was partner with his father and J. Robert Swanson as Saarinen-Swanson-Saarinen in Ann Arbor (1941-1947) and partner with his father as Saarinen and Associates in Ann Arbor (1947-1950). He directed Eero Saarinen and Associates, Birmingham, Michigan, 1950-1961. He acted as a consultant for the Architects Advisory Panel for the Unesco buildings (built 1955-1958) in Paris, France. American architect.

Names:
Saarinen, Eero (**preferred**,V,index,English-P,NA,U)
Eero Saarinen (V,display)
סרינן, אירו (U,Hebrew-P,NA,U)

Nationalities:
American (**preferred**)
Finnish

Roles:
artist (**preferred**)
designer
architect
furniture designer

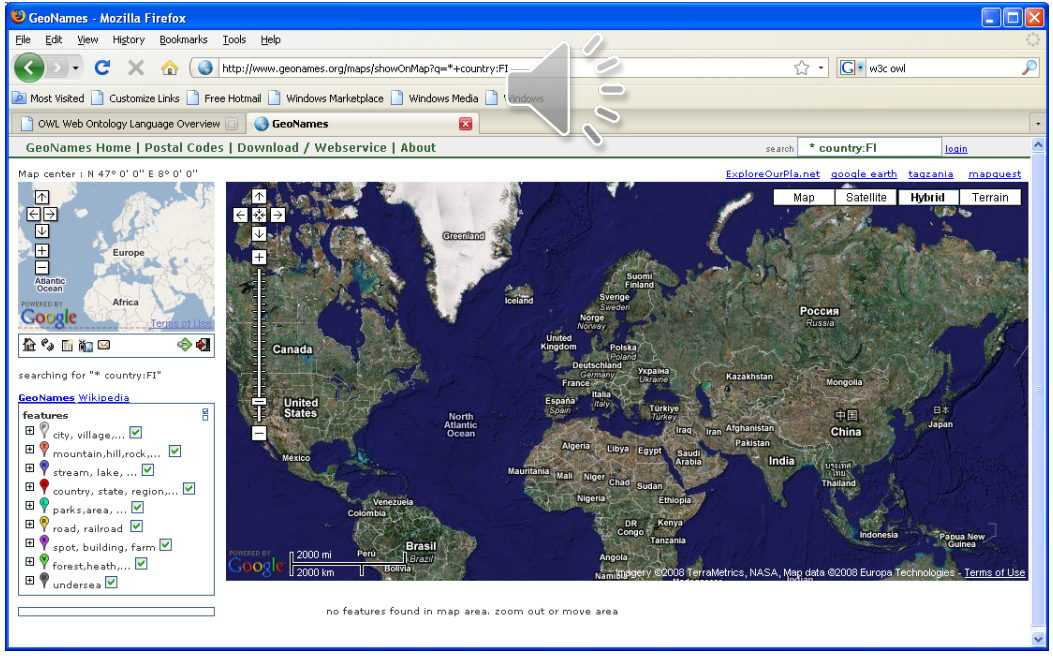
Gender: male

Events:
active: United States (North and Central America) (nation)
active: Finland (Europe) (nation)

Related People or Corporate Bodies:
child of .... Saarinen, Eliel
..... (Finnish architect, 1873-1950, active in the United States) [500027014]
employee of .... Eklund, Jarl
..... (Finnish architect, 1876-1962) [500069436]
employee was .... Pelli, Cesar
..... (American architect and teacher, 1926-2019, born in Argentina) [500023533]
founder of .... Eero Saarinen & Associates
..... (American architectural firm, active 1950-1961) [500119694]
member of .... Saarinen, Saarinen and Associates
..... (American architectural partnership, active 1947-1950) [500229797]
member of .... Saarinen, Swanson, Saarinen
..... (American architectural partnership, active 1941-1947) [500229808]
partner of .... Saarinen, Eliel 1941-1950
..... (Finnish architect, 1873-1950, active in the United States) [500027014]
partner of .... Swanson, J. Robert F. 1941-1947
..... (American architect, active late 20th century) [500113110]
related to .... Saarinen and Saarinen
..... (Finnish architectural firm, contemporary) [500291347]

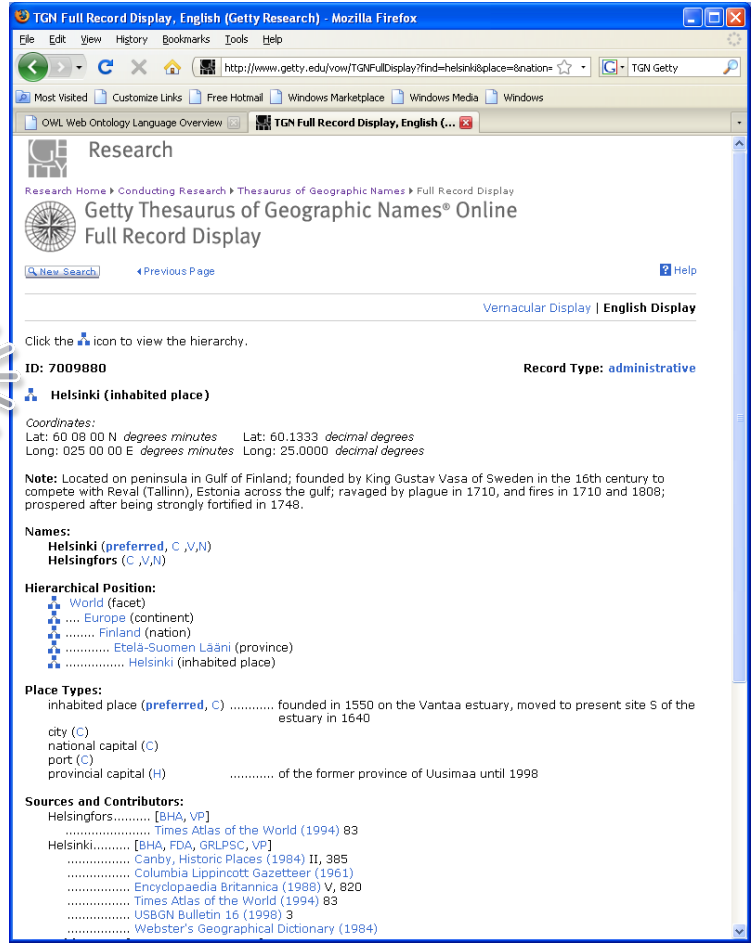
Geonames

- Classes: 9 feature classes, 645 feature codes
- Instances:
  - *8 million geographical names, 6.5 million unique features, 2.2 million populated places, 1.8 million alternate names*
  - *Registries and Wiki used for populating the ontology*



TGN Thesaurus of Geographical Names

- 912,000 records
- 1.1 million names, place types, coordinates, and descriptive notes
- Places important for the study of art and architecture
- Available in a Linked Open Data service: Getty Thesaurus of Geographic Names (Getty Research Institute)



W3C Standards for Semantic Web Ontologies/Vocabularies

RDF Schema

- Class and property hierarchies

SKOS Simple Knowledge Organization System

- Light-weight semantics
- E.g., for representing existing glossaries, thesauri, and classifications

OWL Web Ontology Language

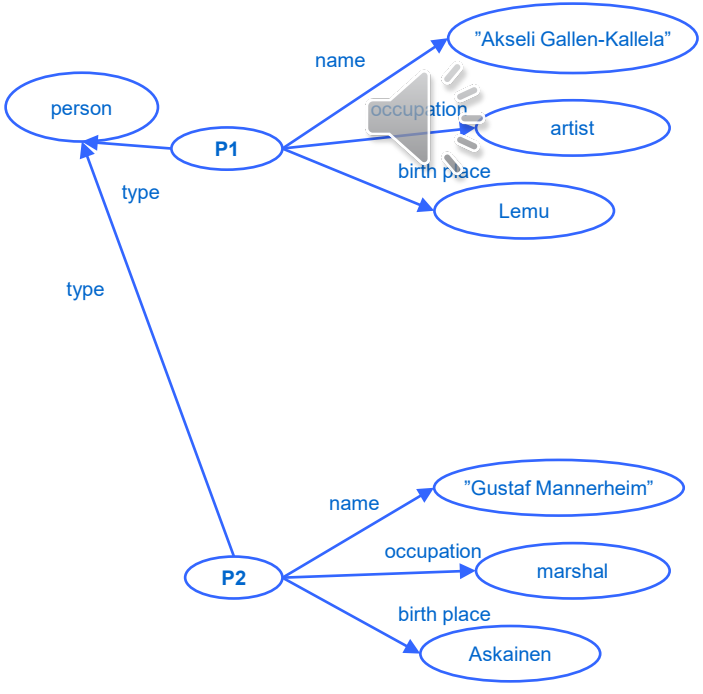
- Rich semantics based on logic
- Supports more advanced reasoning

Metadata + Ontologies = Linked Data (Web of Data)

Enriching Data by Data Linking through Shared Ontologies: An Example

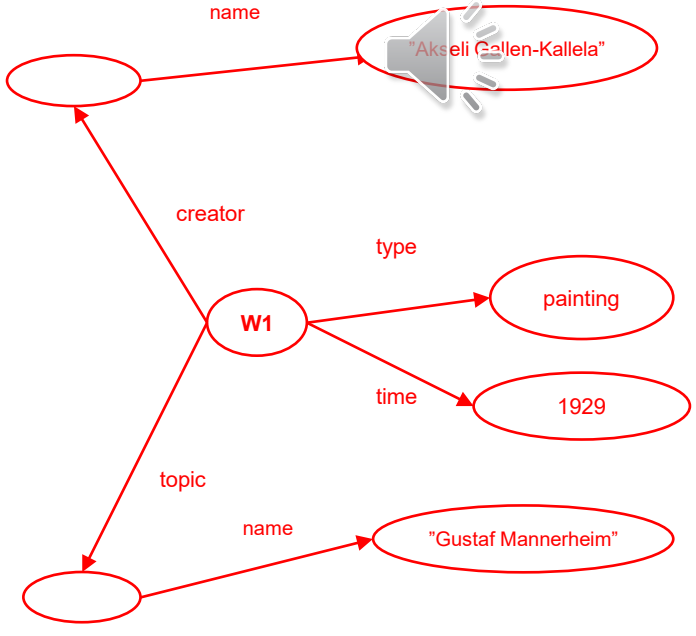
Biography Centers and Libraries Have Databases of about Historical People

person	name	occupation	birth place	…
P1	Akseli Gallen-Kallela	artist	Lemu	
P2	Gustaf Mannerheim	marshal	Askainen	
…				



Museums Catalogue Paintings

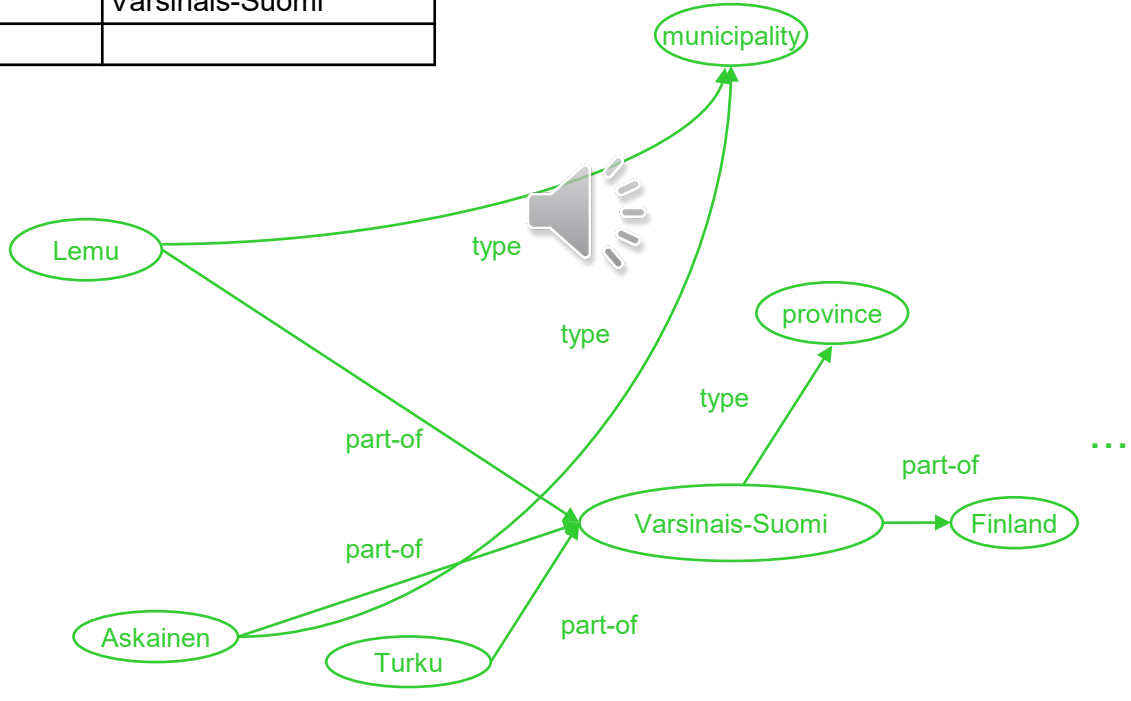
Work	name	creator	time	Topic	…
W1	Portrait of Mannerheim	Akseli Gallen-Kallela	1929	Gustaf Mannerheim	
W2	Aino Triptych	Akseli Gallen-Kallela	1891	Aino, Kalevala	
…					



…

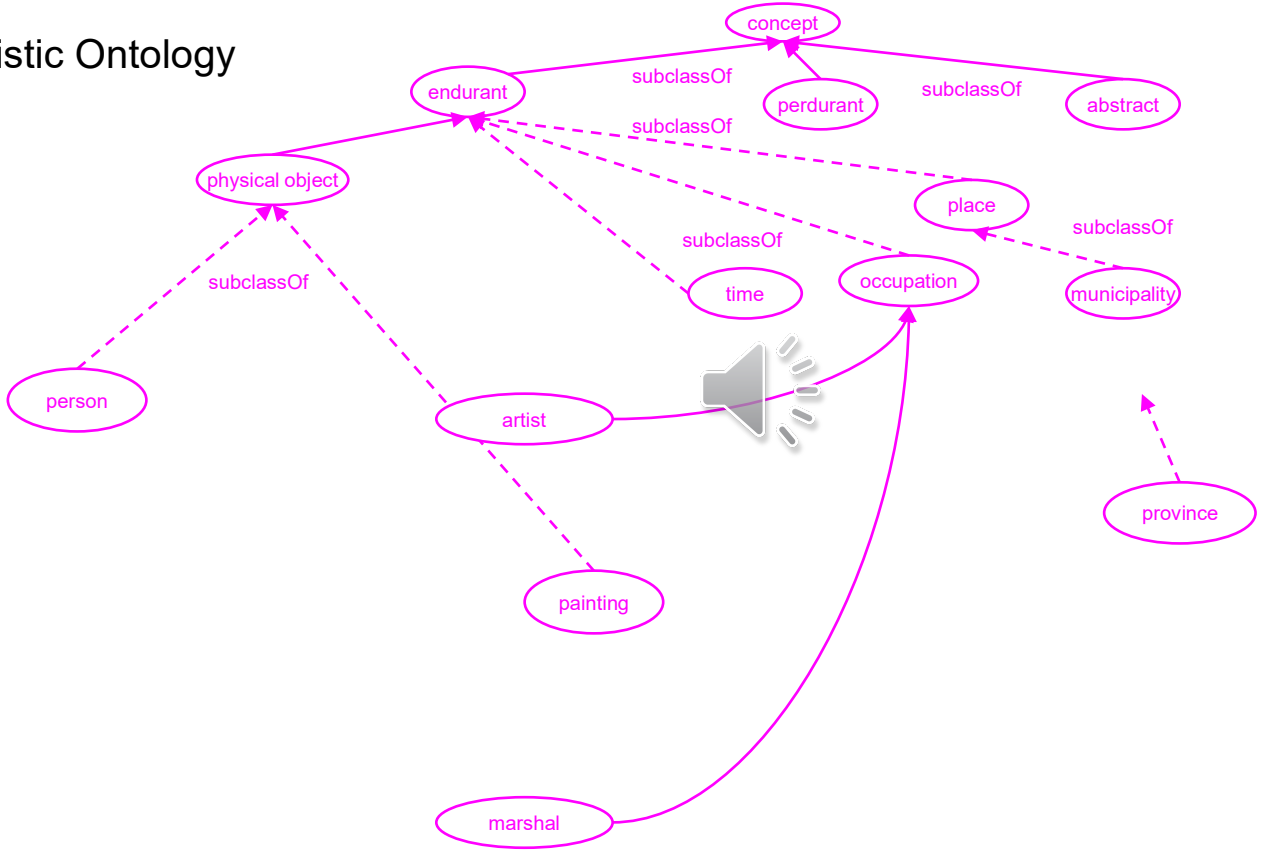
Land Surveys Maintains Place Registries

municipality	province
Askainen	Varsinais-Suomi
Helsinki	Uusimaa
Lemu	Varsinais-Suomi
Turku	Varsinais-Suomi
…	

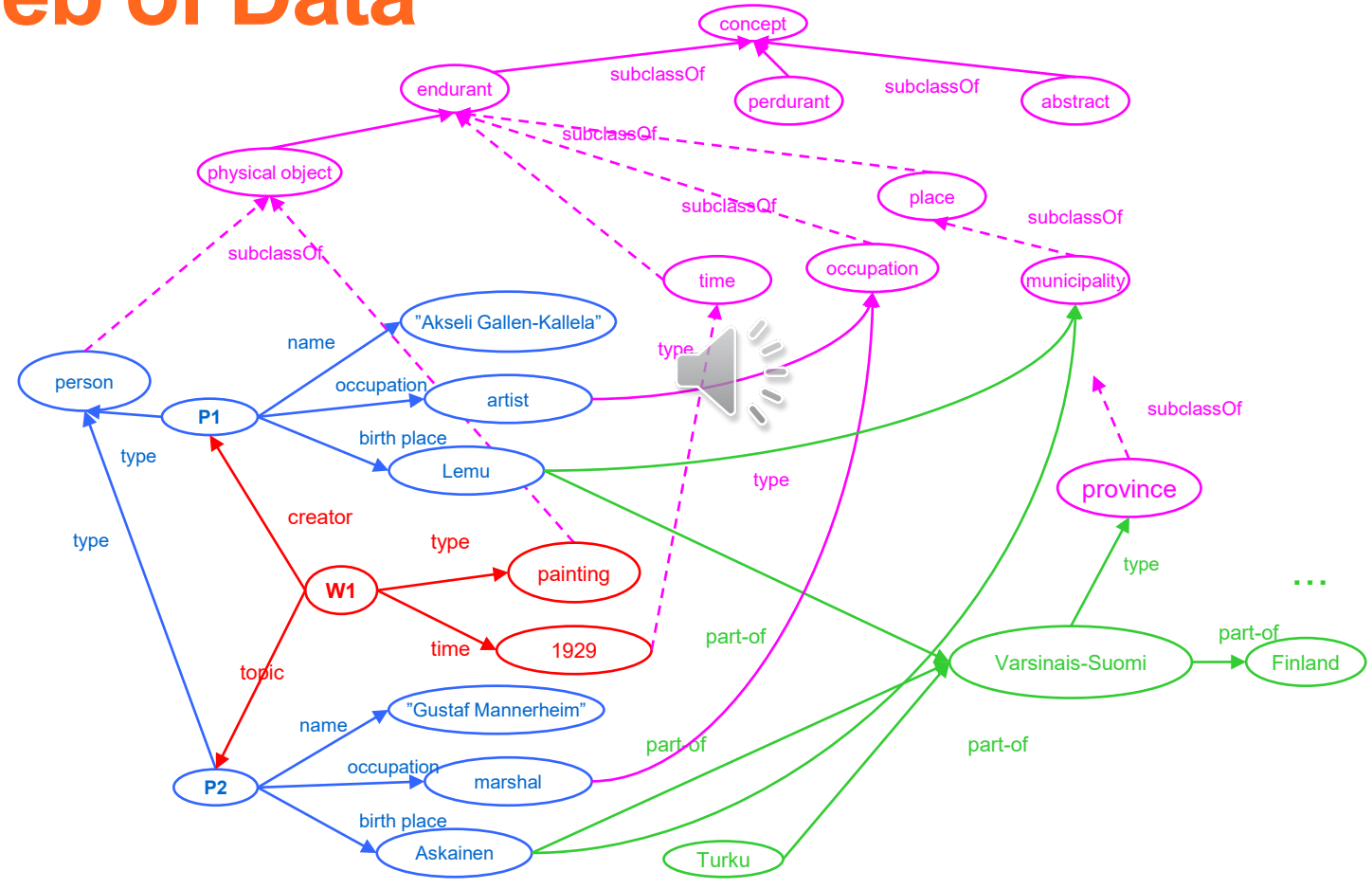


National Library Builds Ontologies

Holistic Ontology



Semantic RDF Graph Combines All Data: Web of Data



Two Key Challenges in Aggregating Data by Data Linking

Ontogies used is metadata descriptions must be shared by collaborating parties

- **Otherwise the data just does not link properly!**
- **Multiple concepts for the same thing emerge**

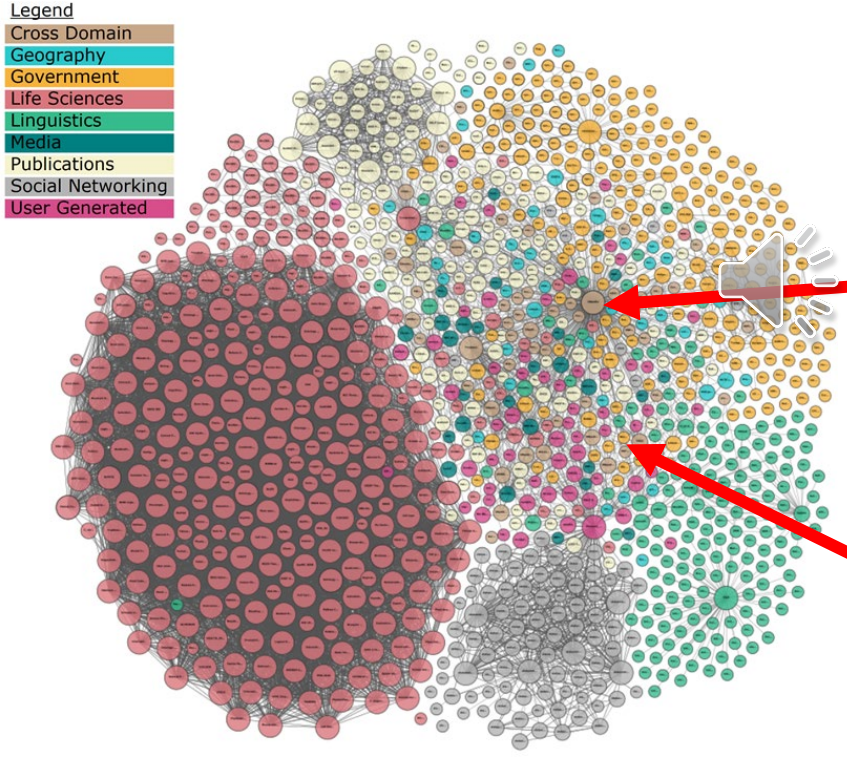
Metadata models have to be aligned

- E.g. two libraries providing data about books in different forms

- Otherwise the data is not interoperable

Web of Data: Linked Open Data Cloud

Human Knowledge on the Semantic Web



Wikipedia

WarSampo

LODStats 2018 (http://stats.lod2.eu/):
10 000 datasets,150 billion triples

Application Example: WarSampo – Finnish WW2 on the Semantic Web

- https://vimeo.com/212249404

Rule Level - Logic

The Idea of Rules

- Semantic web semantics is based on **logic**
- Logic = “new” information can be derived from old by (rule-based) reasoning

Rule Markup Language RuleML

Standardized XML notation for rules

$$hasParent(?x1,?x2) \wedge hasBrother(?x2,?x3) \Rightarrow hasUncle(?x1,?x3)$$

```
<ruleml:imp>
 <ruleml:_rlab ruleml:href="#example1"/>
 <ruleml:_body>
   <swrlx:individualPropertyAtom  swrlx:property="hasParent">
     <ruleml:var>x1</ruleml:var>
     <ruleml:var>x2</ruleml:var>
   </swrlx:individualPropertyAtom>
   <swrlx:individualPropertyAtom  swrlx:property="hasBrother">
     <ruleml:var>x2</ruleml:var>
     <ruleml:var>x3</ruleml:var>
   </swrlx:individualPropertyAtom>
 </ruleml:_body>
 <ruleml:_head>
   <swrlx:individualPropertyAtom  swrlx:property="hasUncle">
     <ruleml:var>x1</ruleml:var>
     <ruleml:var>x3</ruleml:var>
   </swrlx:individualPropertyAtom>
 </ruleml:_head>
</ruleml:imp>
```

Application Example: MuseumFinland Recommends

Inference rules tell machine about the world

- E.g., that ”student’s cap” is related to ”parties”
- E.g., that entities are related to each other if their superclasses are related to each other

Based on the graph of metadata+ontologies, machine can:

- Reason interesting new relations between museum items, and
- Provide them to end users as recommendation links

Application example: MuseumFinland

Application Domains of Semantic web

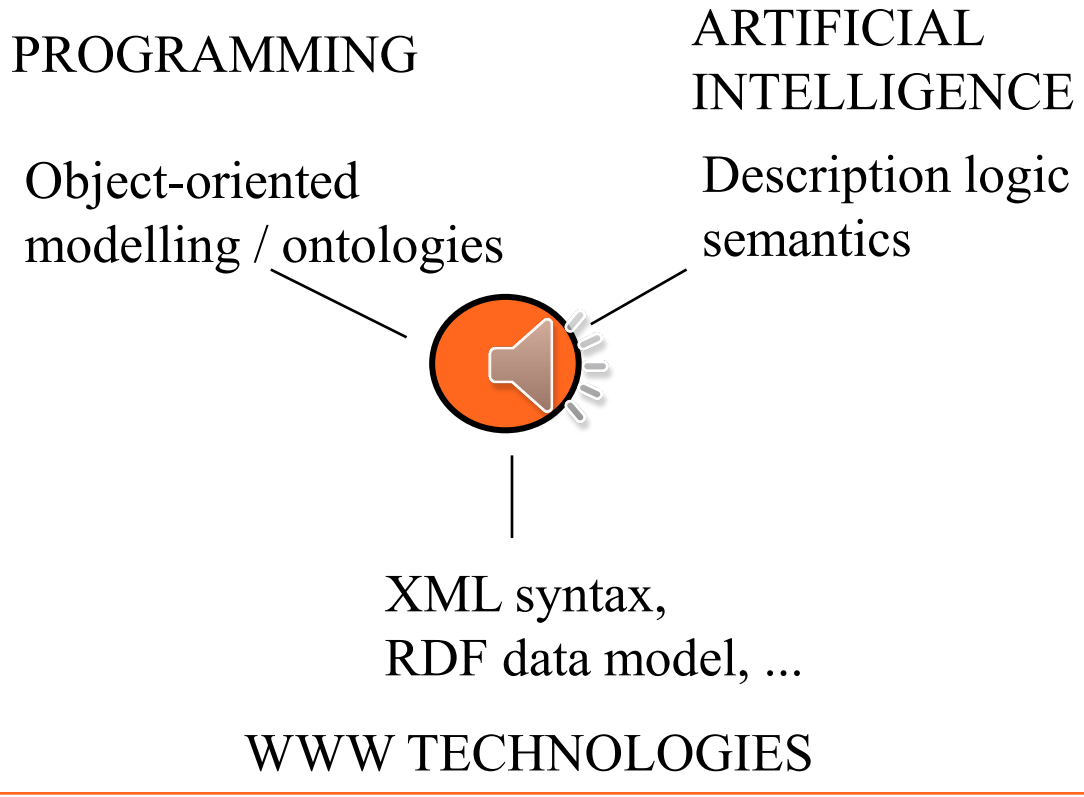
- Semantic portals
- Information retrieval systems
- Recommender systems
- Knowledge management systems
- Personalized systems

…

Examples of applications / domains have been collected here:

https://www.w3.org/2001/sw/sweo/public/UseCases/

What is New? Components of Semantic Web



What is the Semantic web?

Content perspective: A new metadata layer on the Web describing its contents in terms of shared vocabularies, i.e., ontologies

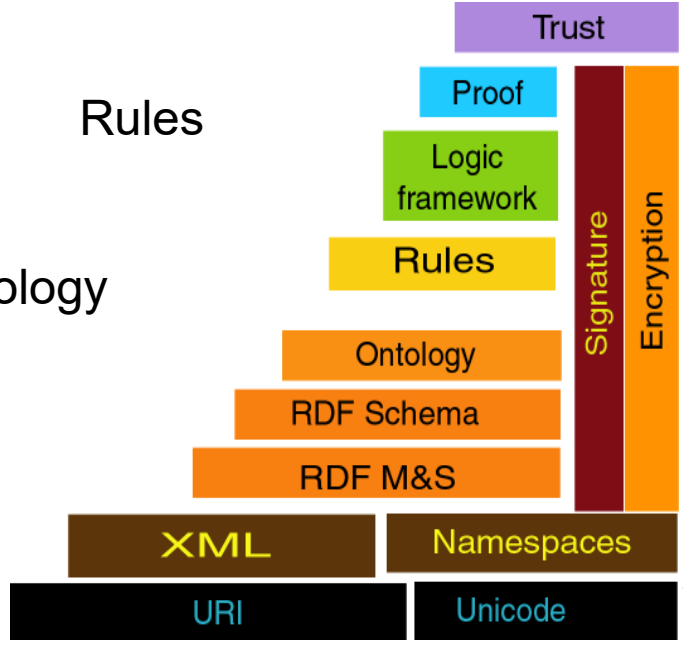
- Web as a global database system
- Web of Pages vs. Web of Data

Application perspective: Machine-understandable web

- The meaning (semantics) of contents accessible to machines
- Enables human usage
  - *Intelligent web services*
  - *Semantic interoperability*

Technological perspective:
Next layers above XML

- W3C standards:
  RDF(S), OWL , SPARQL, etc.



Questions