Possible Ontologies: How Reality Constrains Building Relevant Ontologies

Dr. Martin Hepp
DERI, University of Innsbruck

Logic and Information Workshop
Schloss Münchenwiler, November 23, 2006

http://www.heppnetz.de

Martin Hepp
martin.hepp@deri.org

About the speaker: Martin Hepp

- **Senior Researcher** and Cluster Leader „Semantics in Business Information Systems" at **DERI**, University of Innsbruck.
- Ph.D. in Management Information Systems, Bayerische Julius-Maximilians-Universität, Würzburg, Germany (2003); M.B.A., ditto, Würzburg, Germany (1999)

See **http://www.heppnetz.de** for current papers and presentations.

Prolog

- For 10+ years, formal ontologies have been a research topic in Computer Science.
- There are 10+k scientific publications on the Semantic Web.
- Formalism, tools, and infrastructure are pretty mature and widely available.
- Still, there are very little „real“ ontologies.
- Why?

“It’s because the stupid business people have not yet realized the enormous potential of ontologies”!

Outline

- Problems of Building Ontologies
  - Conceptual dynamics
  - Resources consumption
  - Communication between creators and users
  - Intellectual Property Rights
- Solutions
  - Automation/Semi-Automatic Creation
  - Enhanced Wikis as Collaborative Ontology Engineering Tools
- Example: The Making of eClassOWL

Ontology for this talk ☺

- Ontologies in my understanding are
  - community agreements over
  - domain theories, reached by specifying
    - concepts,
    - relations,
    - attributes,
    - axioms, and
    - (ontologically relevant) instances

  by formal and/or informal means.

I. Problems of Building Ontologies

I. Problems of Building Ontologies

1. Conceptual Specificity and Size
2. Conceptual Dynamics
3. Resources vs. Benefits
4. Ontology Economics, esp. Network Externalities
5. Communication between Creators and Users
6. Intellectual Property Bottleneck
7. Prediction: Possible Ontologies

[cf. 5]

1. Conceptual Specificity and Size

- Minimal Ontological Commitment
- Expressivity/Specificity Dilemma
- Use cases/examples used by the ontology community as a justification for their works require the existence of rather specific domain ontologies.

Size of Popular Informal Content Standards

- **UNSPSC,** http://www.unspsc.org
  - 20,700 classes, no properties
  - 55 top-level categories
- **eCl@ss,** http://www.eclass.de
  - 25,000 classes, 5,500 properties
  - 25 top-level categories
- **eOTD,** **http://**www.eotd.org
  - 59,000 classes, 34,000 properties
  - 79 top-level categories
- **RNTD,** http://www.rosettanet.org/technicaldictionary
  - 789 classes, 3,600 properties
  - 1 top-level categories

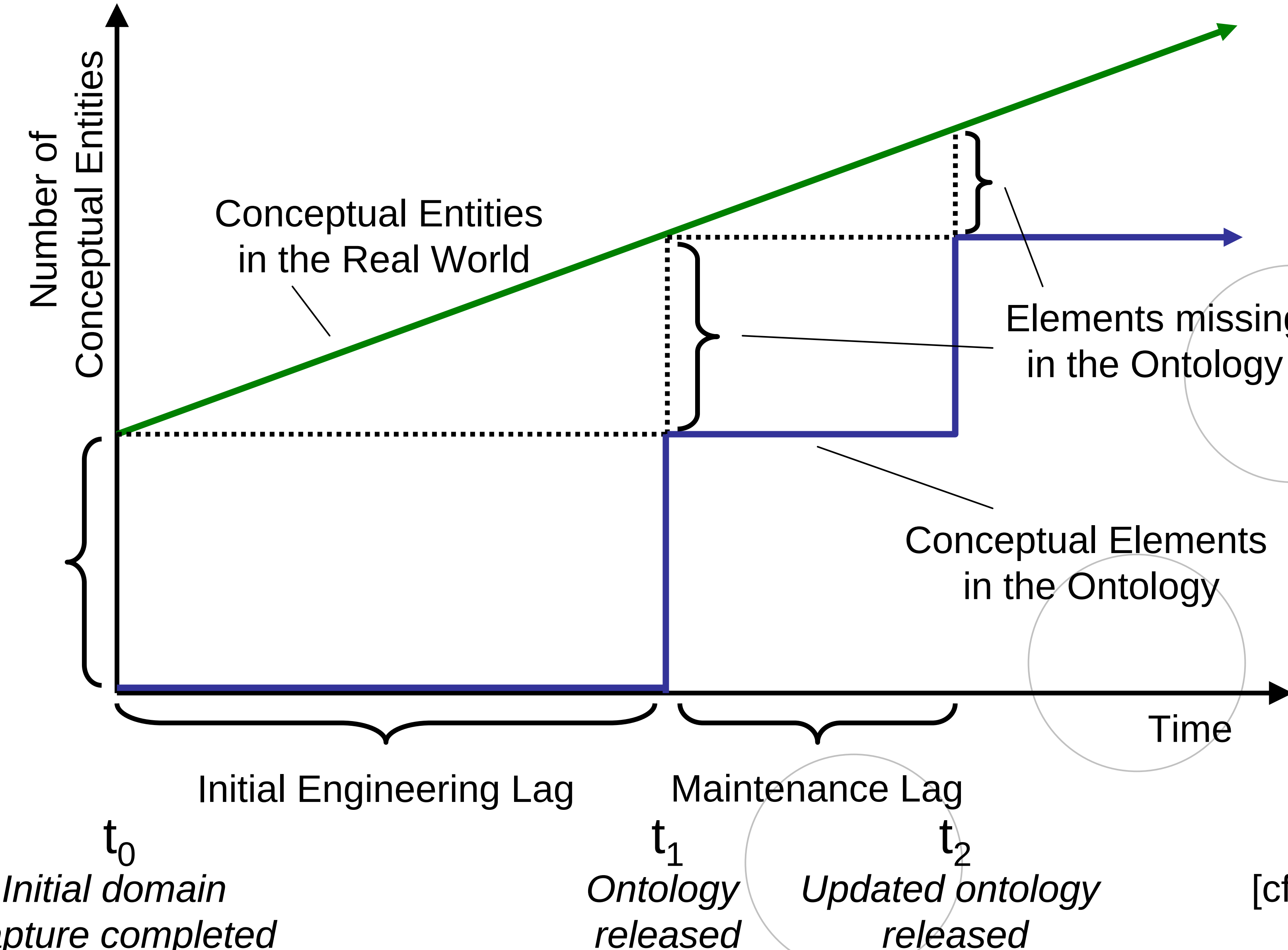
But still: very incomplete coverage!

[cf. 1, 3]

- In practice, we derive the conceptual elements of ontologies, i.e.
  - concepts,
  - relations,
  - attributes,
  - axioms, and
  - (ontologically relevant) instances

from observed *phenomenons*.

Conceptual Dynamics vs. Ontology Engineering Lag



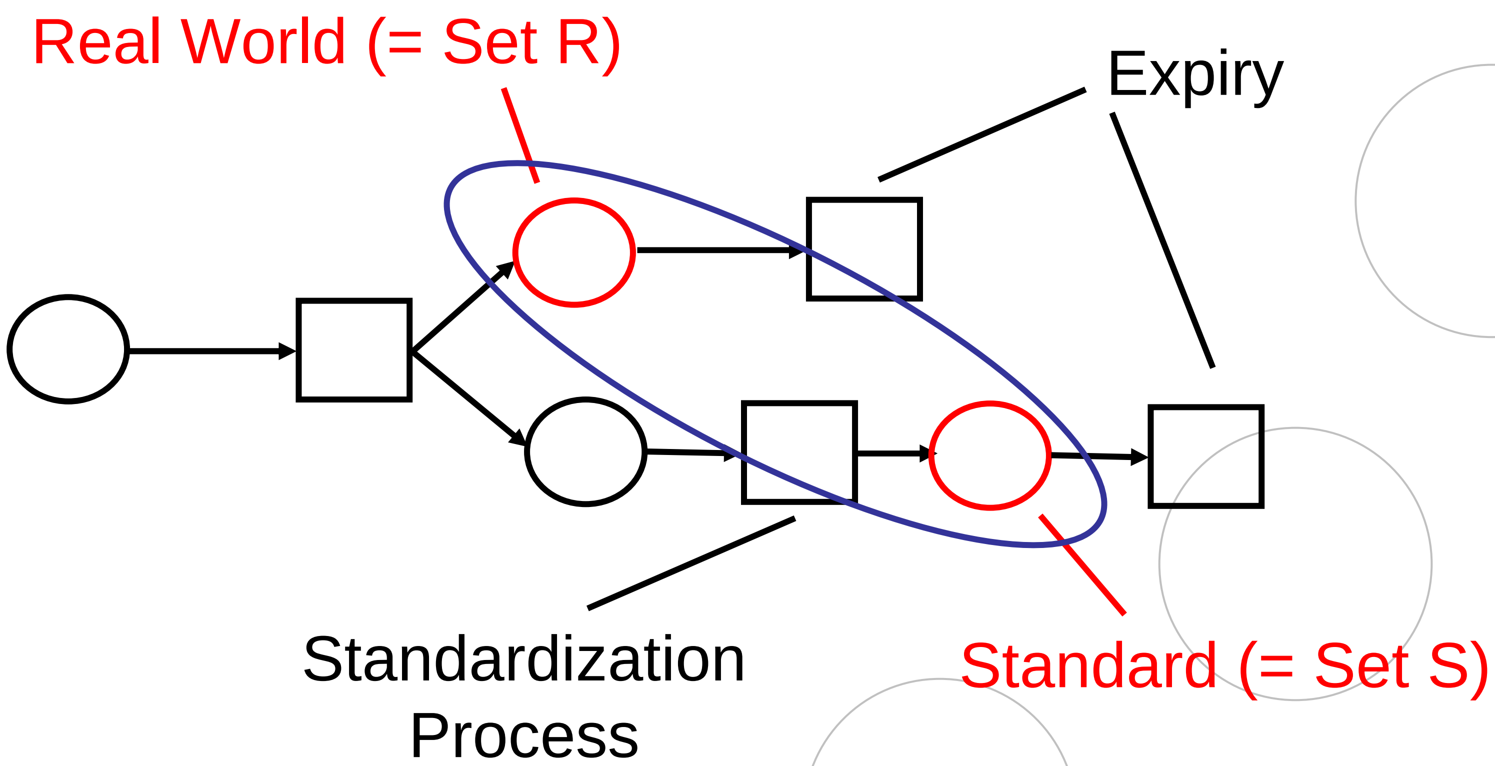
[cf. 5]

Change Dynamics in Business Vocabularies

	Release	Previous release	New classes per 30 days	Mean	Modified classes per 30 days	Mean
eCl@ss	5.0	4.1	865.0	279.6	157.4	1271.4
	5.0SP1	5.0	47.8		10.2	
	5.1beta	5.0SP1	131.6		4918.0	
	5.1de	5.1beta	74.1		0.0	
eOTD	10-01-2003	01-17-2003	6.1	6.2	0.0	0.0
	11-01-2003	10-01-2003	4.8		0.0	
	03-01-2004	11-01-2003	18.3		0.0	
	06-01-2004	03-01-2004	1.6		0.0	
	08-01-2004	06-01-2004	0.0		0.0	
RNTD	2.0	1.4	0.7	1.3	6.4	1.5
	3.0	2.0	2.4		1.0	
	3.1	3.0	0.0		0.1	
	3.2	3.1	0.0		0.0	
	4.0	3.2	3.4		0.0	
UNSPSC	6,0315	5,1001	907.8	233.9	135.6	47.5
	6,0501	6,0315	304.5		53.0	
	6,0801	6,0501	97.5		15.0	
	6,1101	6,0801	69.1		50.2	
	7,0401	6,1101	13.8		29.4	
	7,0901	7,0401	10.8		2.0	

This is only the amount of *actual additions* – the true need for new concepts will be much bigger!

[cf. 1]


Real World (= Set R)
Expiry
Standardization Process
Standard (= Set S)

[cf. 6]

Simulation Run

Raw Data

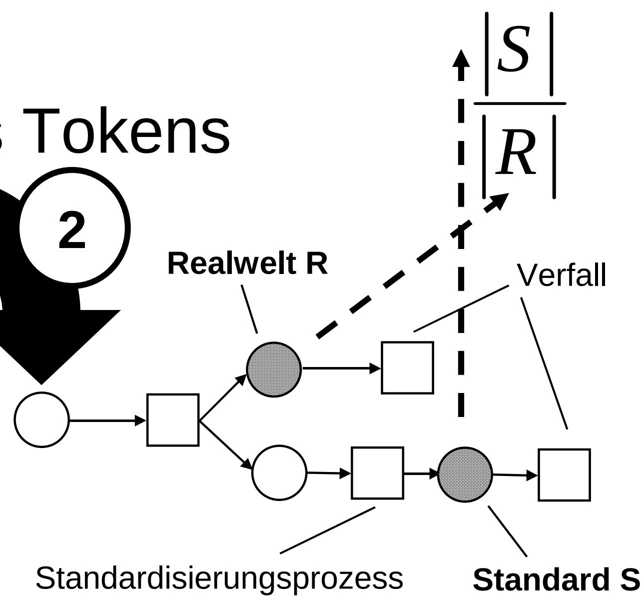
Transformation

1

ID	Description	Date of Birth	Life-span
1	Deposit Tokens	01.01.2003	365 days
2	CBT for MS Office 2003	01.01.2003	180 days

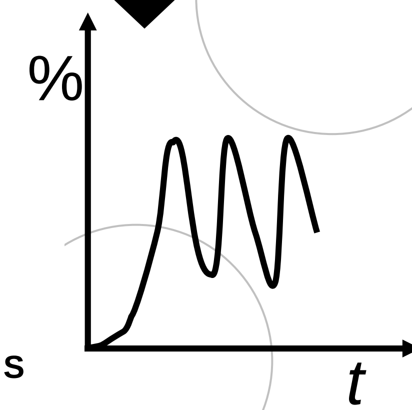
as Tokens

2



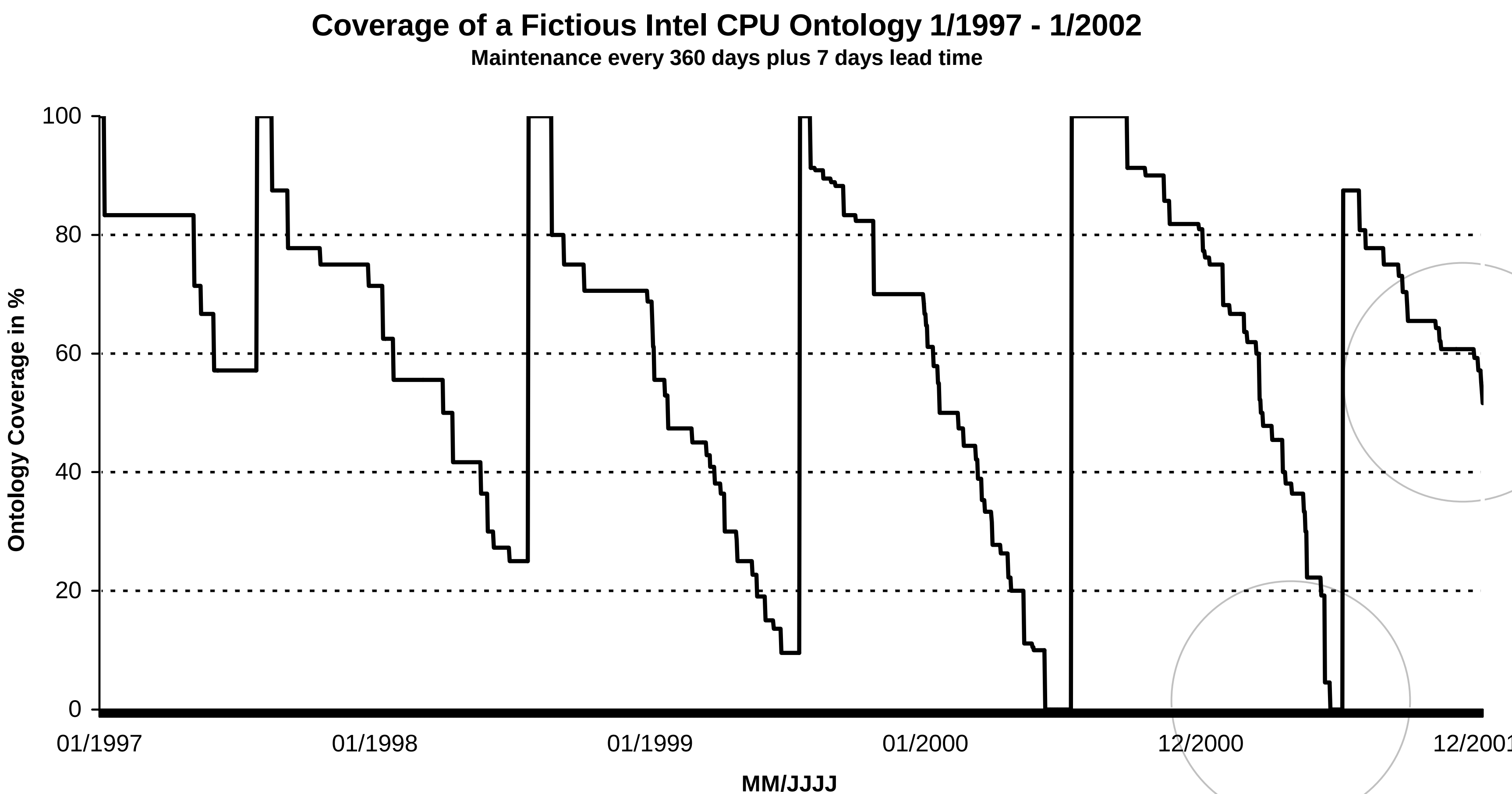
Interpretation

3



[cf. 6]

Results: Typical Update Cycles Result in Insufficient Coverage



[cf. 5, 6]

3. Resources vs. Benefits

- Is it reasonable to build a particular ontology from a resources point of view?
- Economic Perspective
  - Does the gain in automation made possible by the ontology justify the resources necessary to yield the ontology?
- Technical Perspecive
  - Do the problems that we can solve with the help of the ontology outweigh the problems that we must master in order to create it?

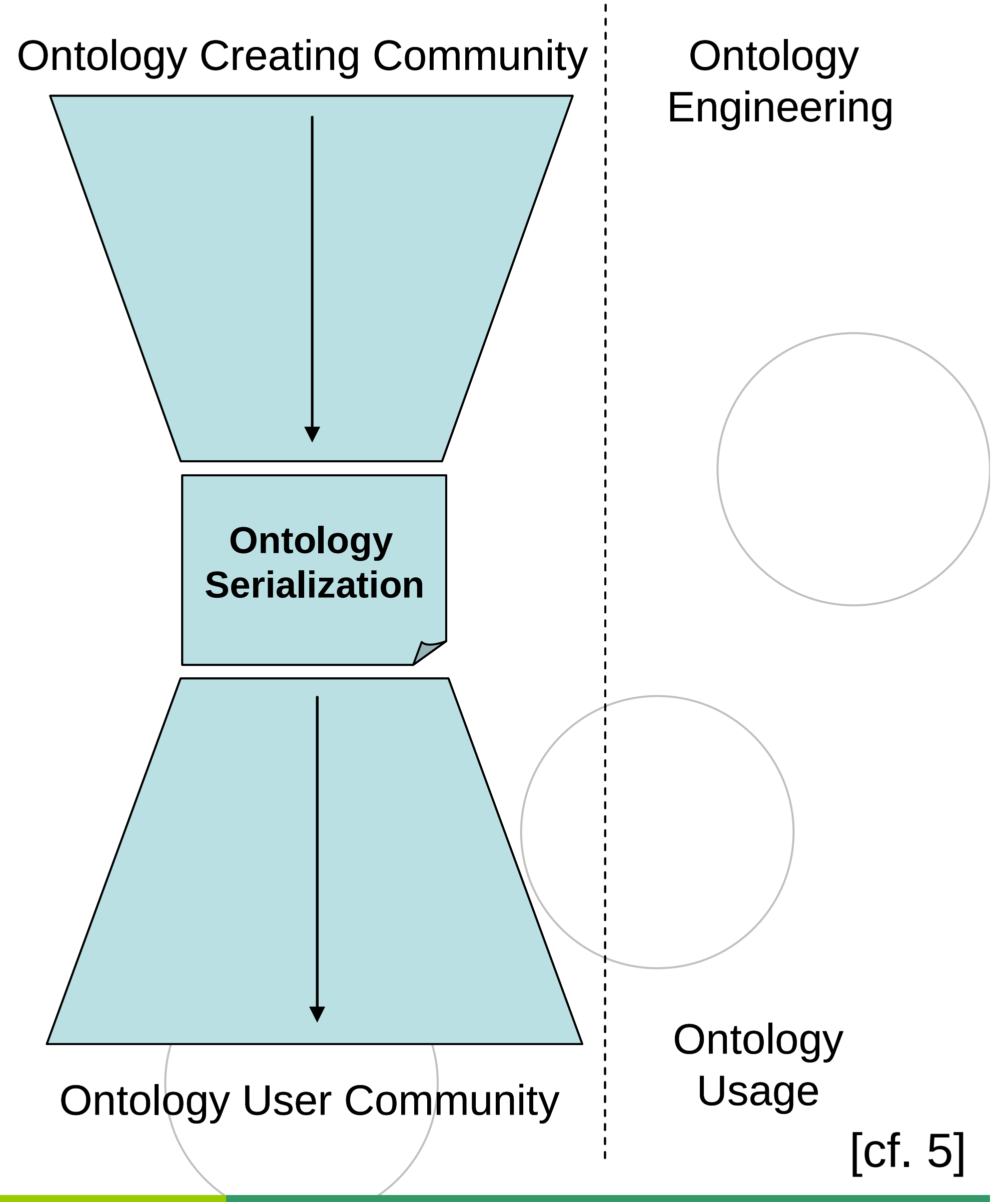
[cf. 5]

4. Ontology Economics, esp. Network Externalities

- Economics of Ontologies
- Distribution of Incentives
- Network Externalities

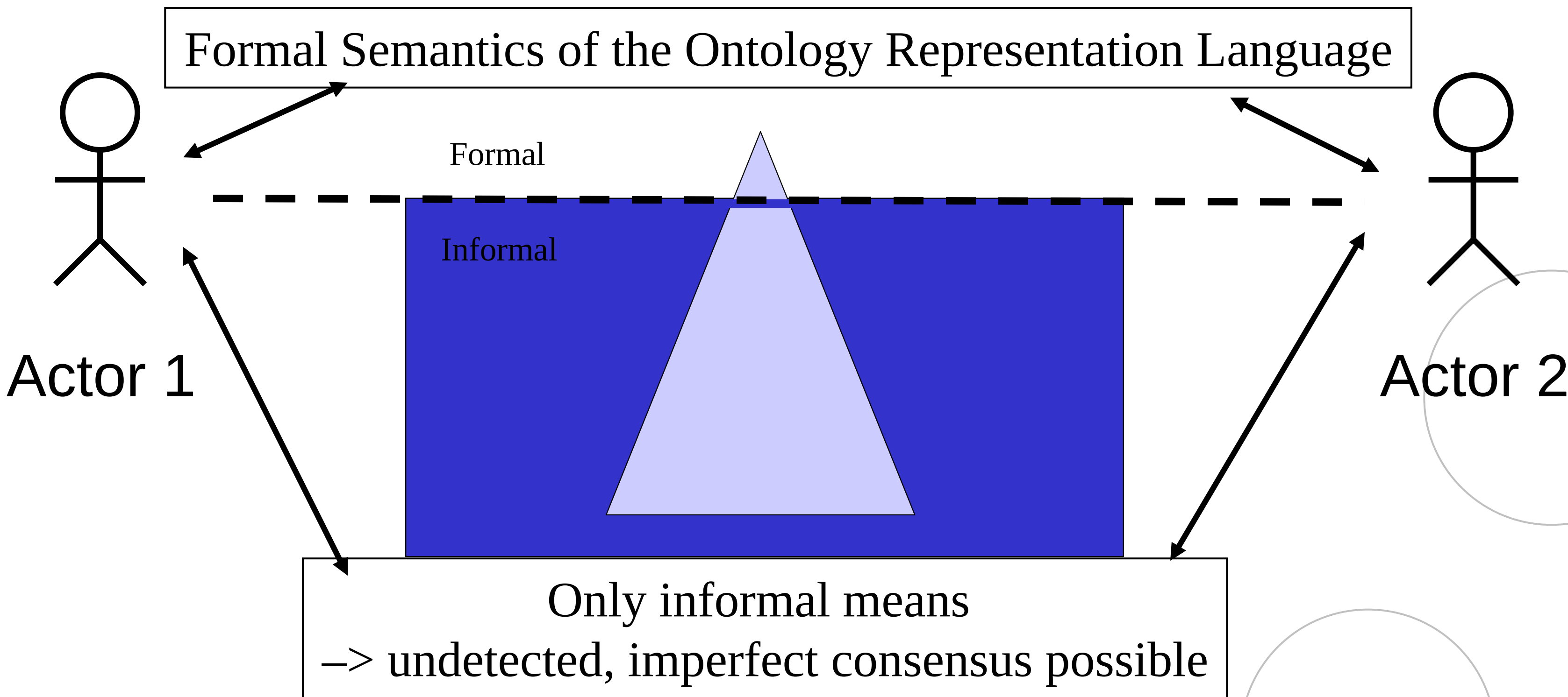
5. Communication between Creators and Users

- Currently, most ontologies are built by a small (worst case: n=1) elite but meant for a wider user community.
- Annotation data / expressing queries
- Two ways of committing to an ontology
  a) by trust
  b) by reviewing the formalization
- Solution b) constrained by
  - skills (e.g. education in logic)
  - resources



[cf. 5]

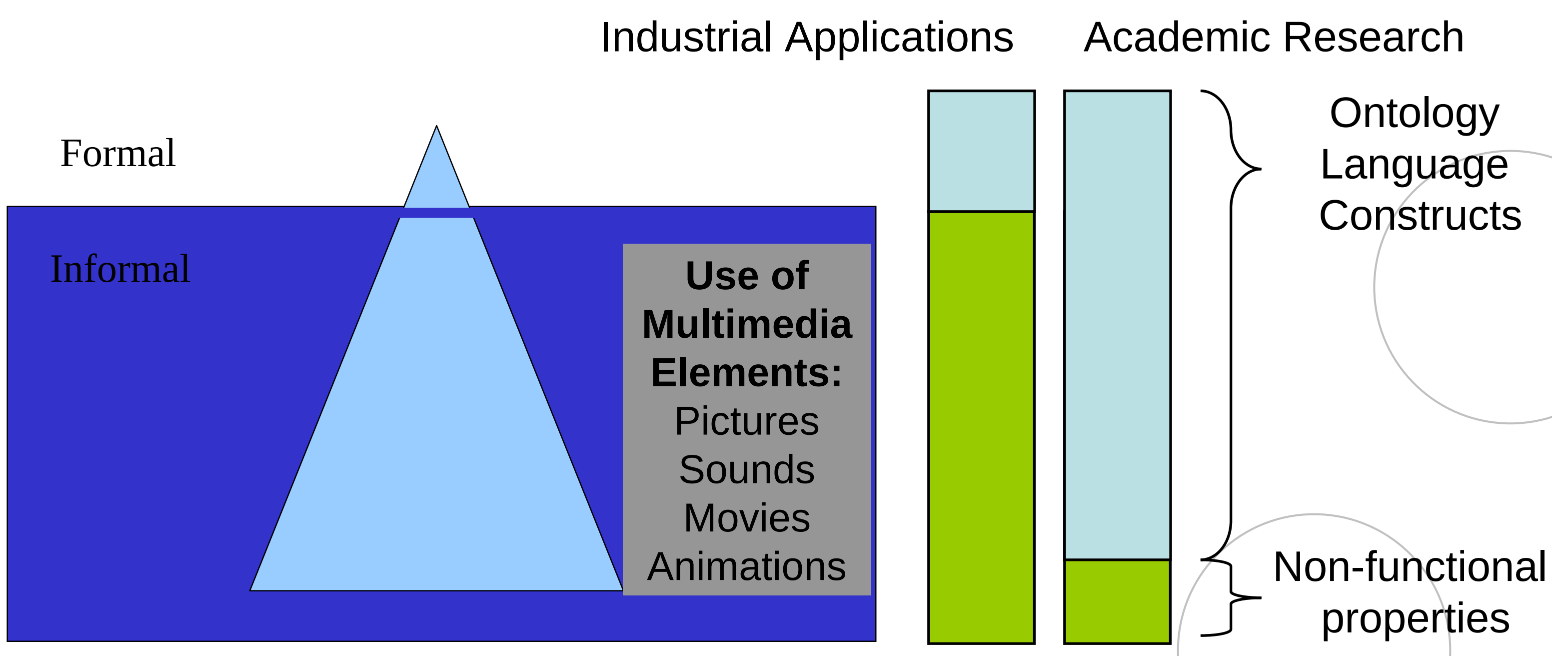
How to Maintain the Community Treaty?



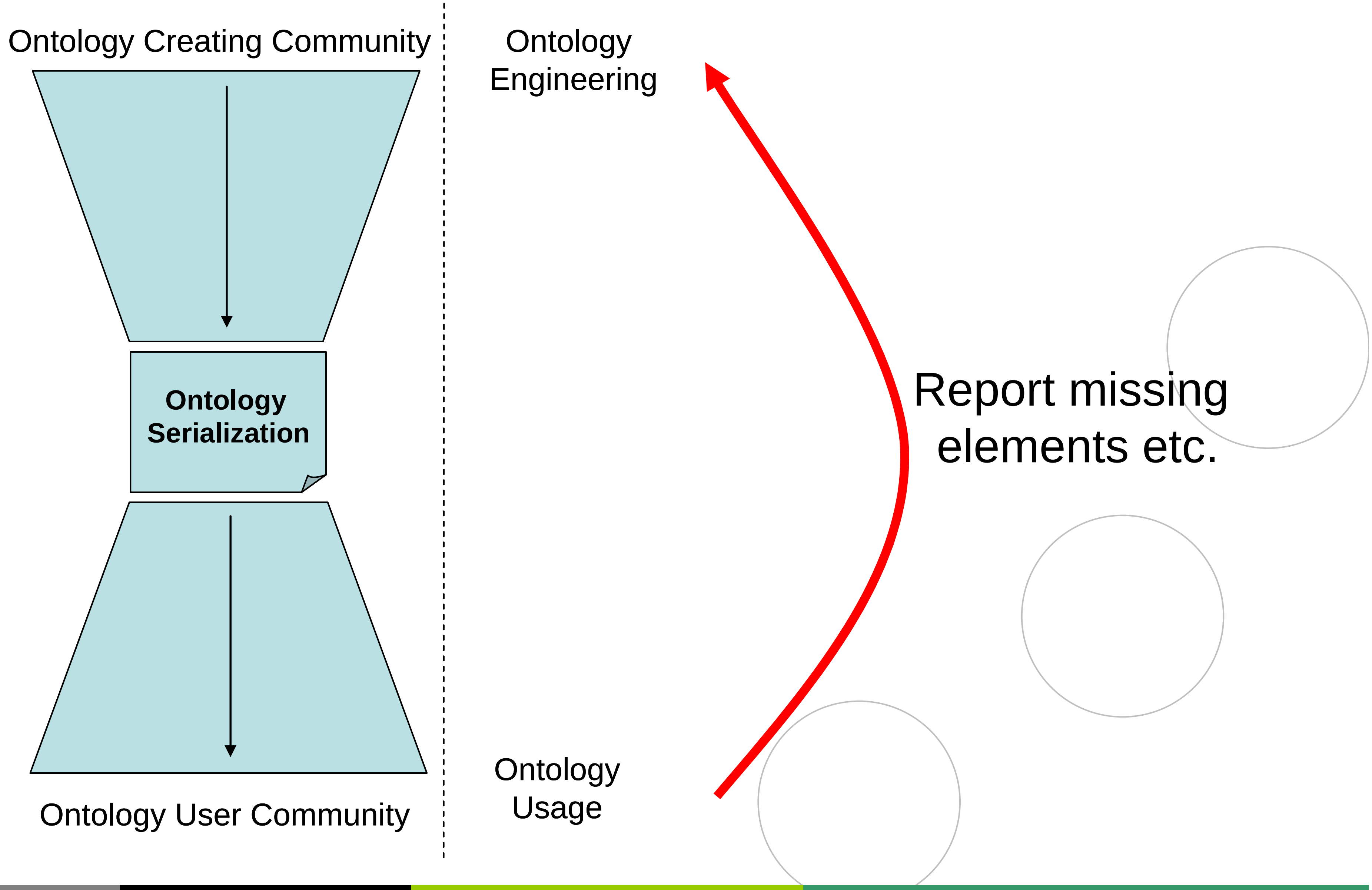
NFP dc:description is a bit simplistic
for the complexity of this problem!

If you cut the link between an ontology and the community,
the ontology stops being an ontology.

Ontologies: Formal vs. Informal



The insufficient back-channel

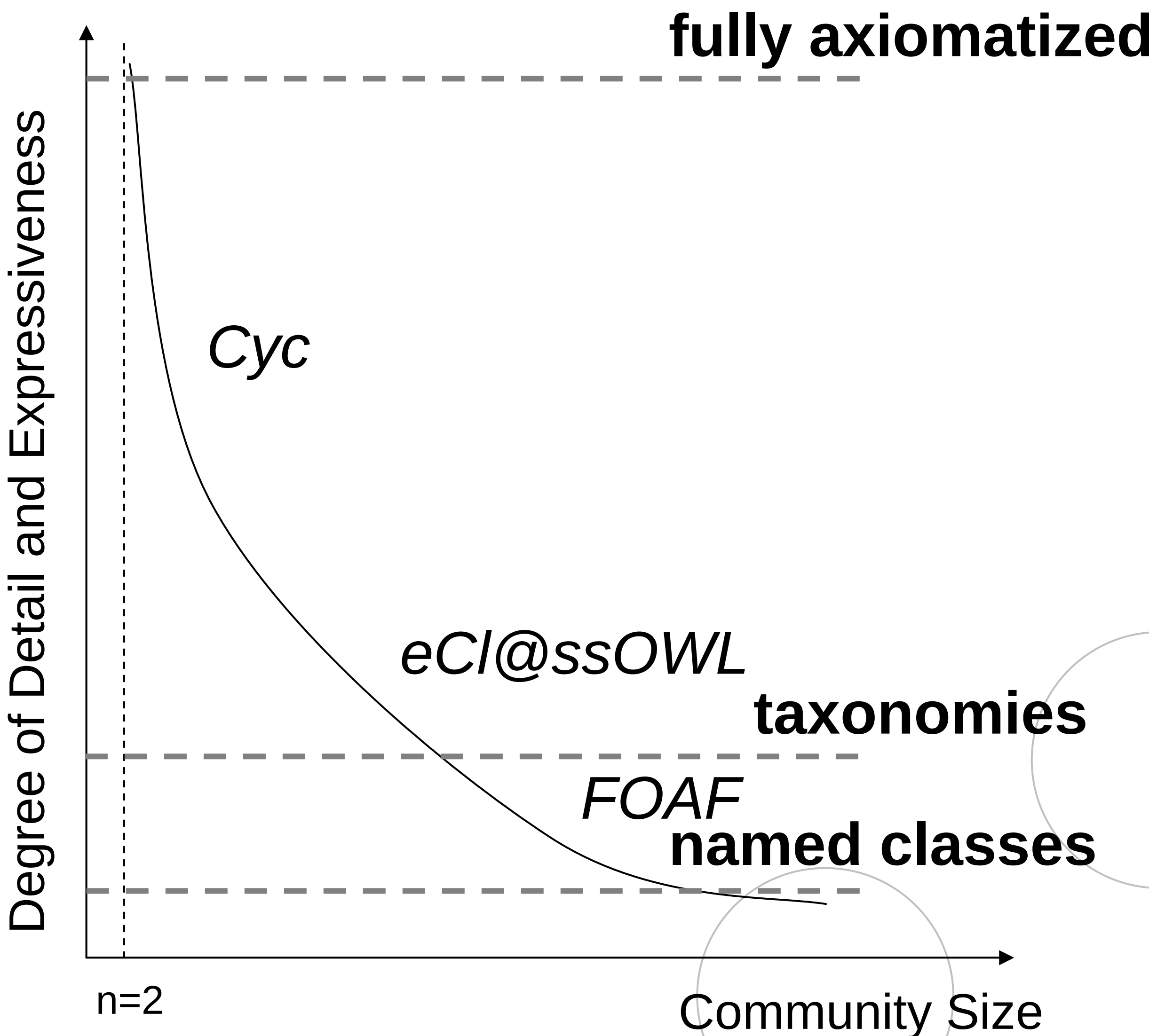


6. Intellectual Property Bottleneck

- We will need to build ontologies that represent existing standards.
- However, standards are often subject to Intellectual Property Rights.
- It is thus not trivial to establish the legal framework for deriving ontologies from relevant standards.
- Examples: UNSPSC, eCl@ss, X12, ISO 639,…
- Same holds for contributions to collaboratively built ontologies.

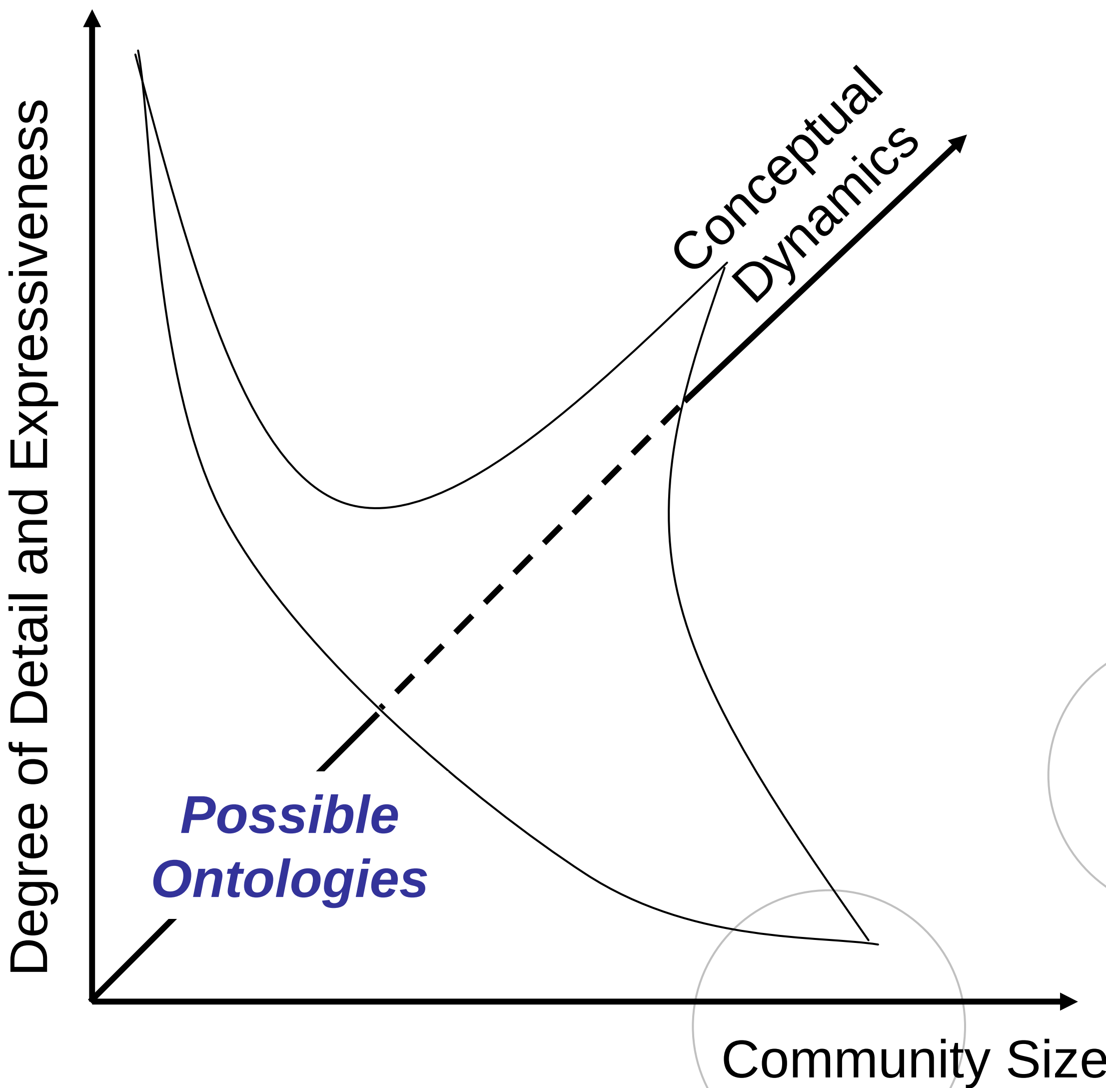
[cf. 5]

7. Prediction: Possible Ontologies (1)



[cf. 5]

7. Prediction: Possible Ontologies (2)



[cf. 5]

Prediction: The Ontology Divide

- Shallow, public-domain, license-free collections of named classes + simple relations
- Narrow, rich, program-like ontologies, developed and licensed like software is licensed today

II. Approaches and Solutions

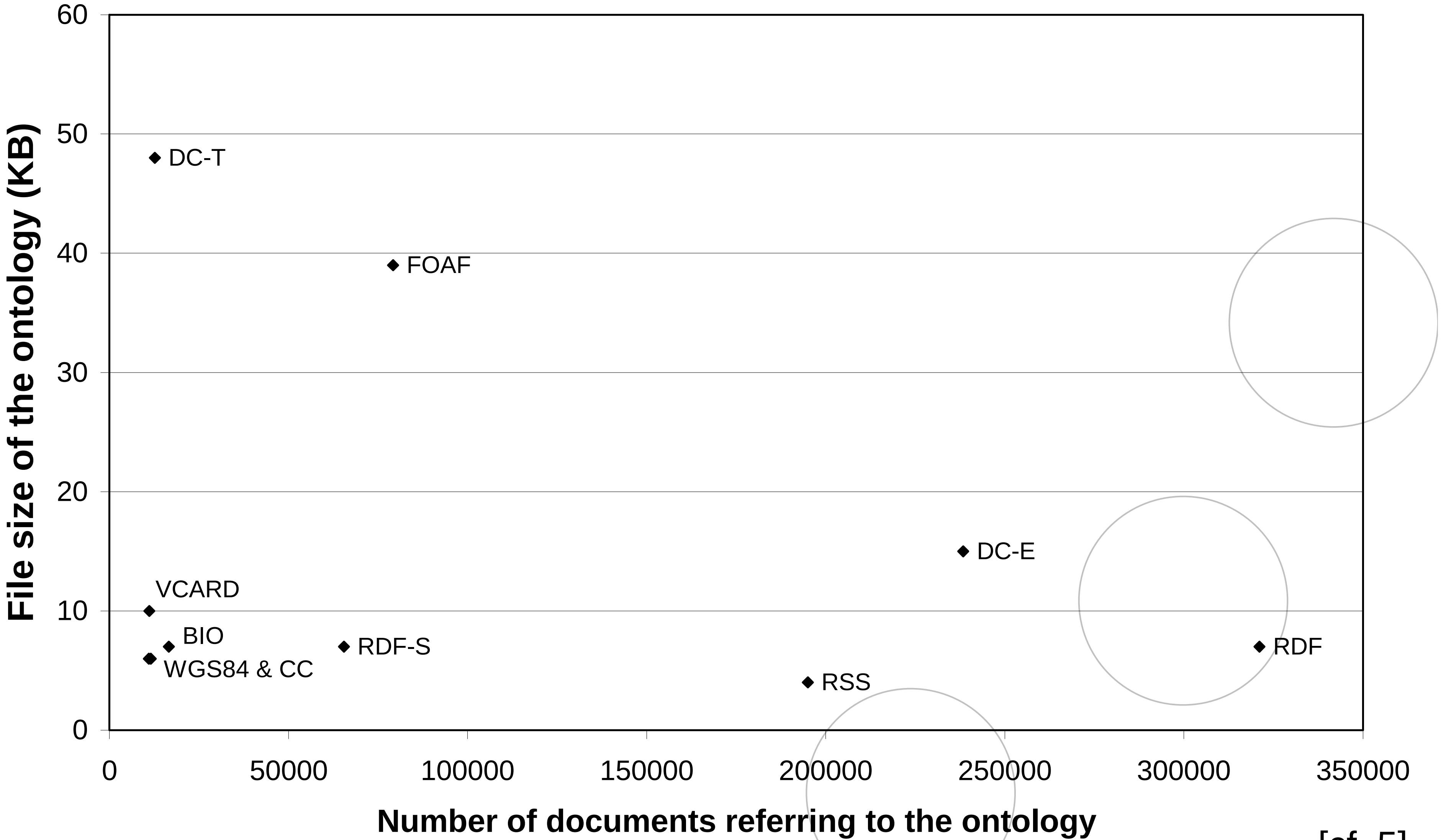
II. Approaches and Solutions

1. General Guidelines
2. Semi-automatic creation of ontologies from existing consensus
   - Gen/Tax Methodology and tooling
3. Combine statistical approaches with declarative approaches
   - E.g. predict missing concepts or relations
4. Community-driven Ontology Engineering
   - Re-use Wikipedia consensus
   - Wiki-like ontology engineering environments with low entrance barriers

1. General Guidelines: What does it take to create a relevant ontology?

- Small
- Well-documented
  - labels/descriptions
  - retrievable online documentation
- Immediate, obvious benefits
  - Ontology familiarization costs lower than perceived utility

Ten most successful Web ontologies (Nov. 2005)



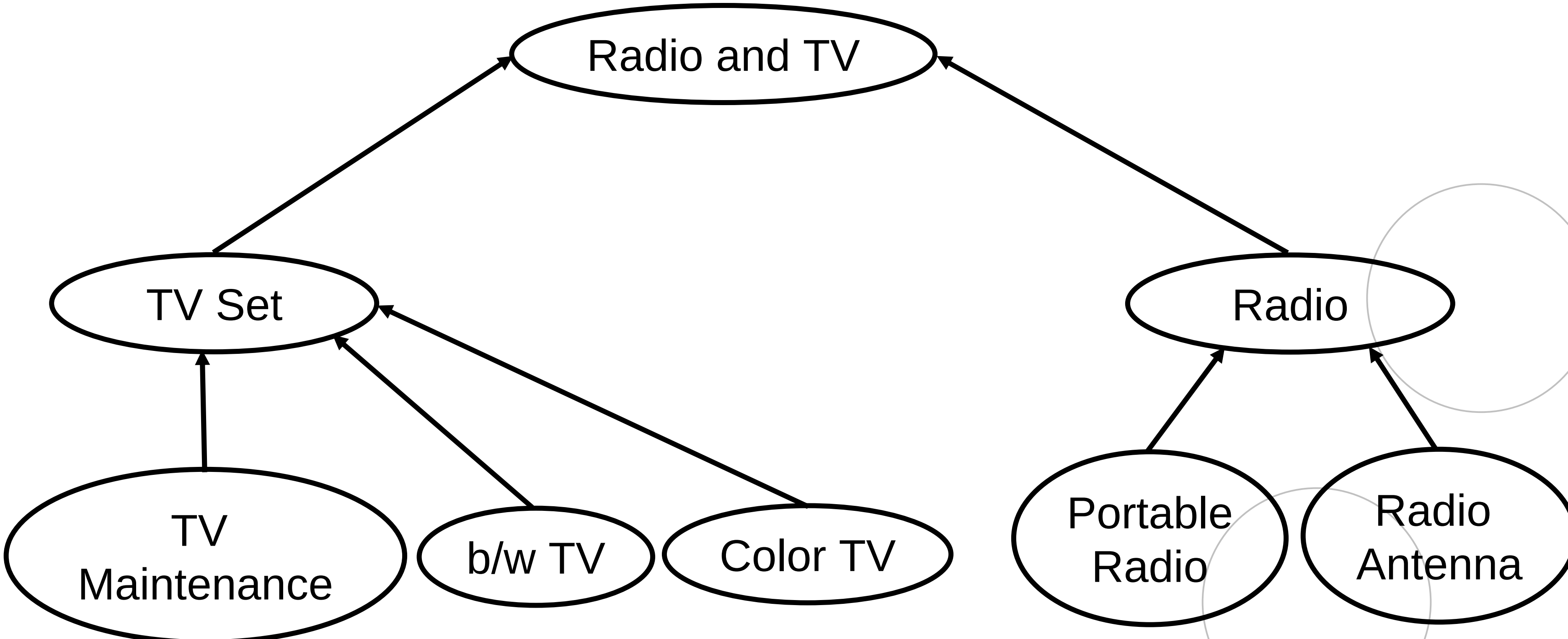
[cf. 5]

2. Semi-automatic creation of ontologies from existing consensus

- Idea: Re-use classfications, vocabularies, standards for building ontologies
- Problems: Formalizing requires interpretation

[cf. 3]

Typical Problem



[cf. 3]

Concepts in Hierarchical Standards

1

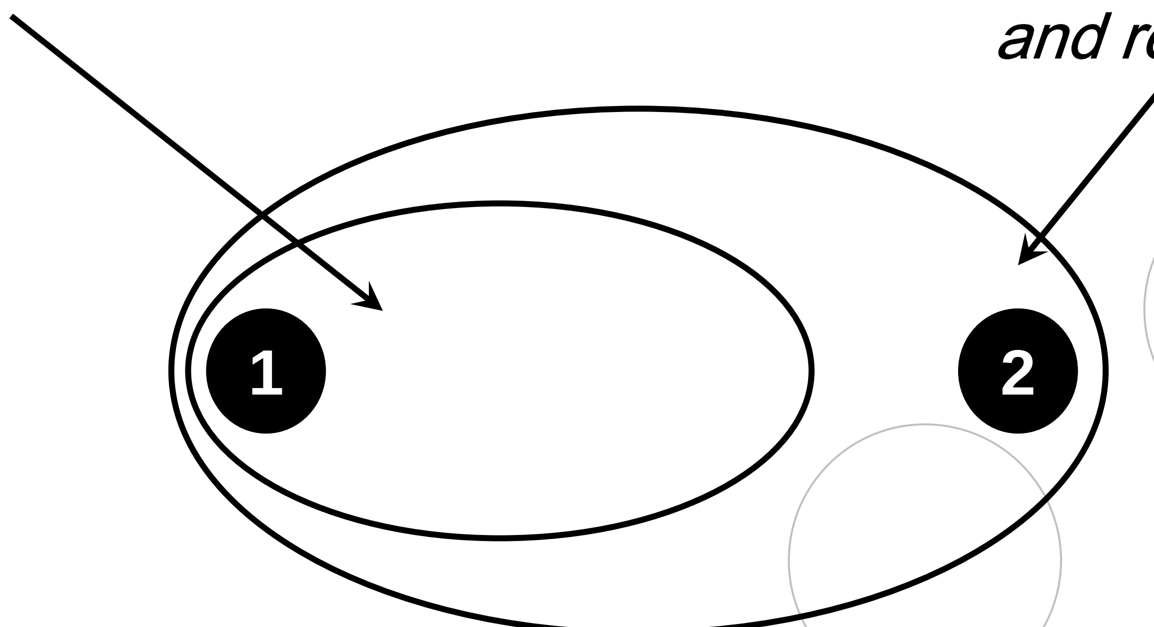
Generic concept

(Example: “Home appliances”)

2

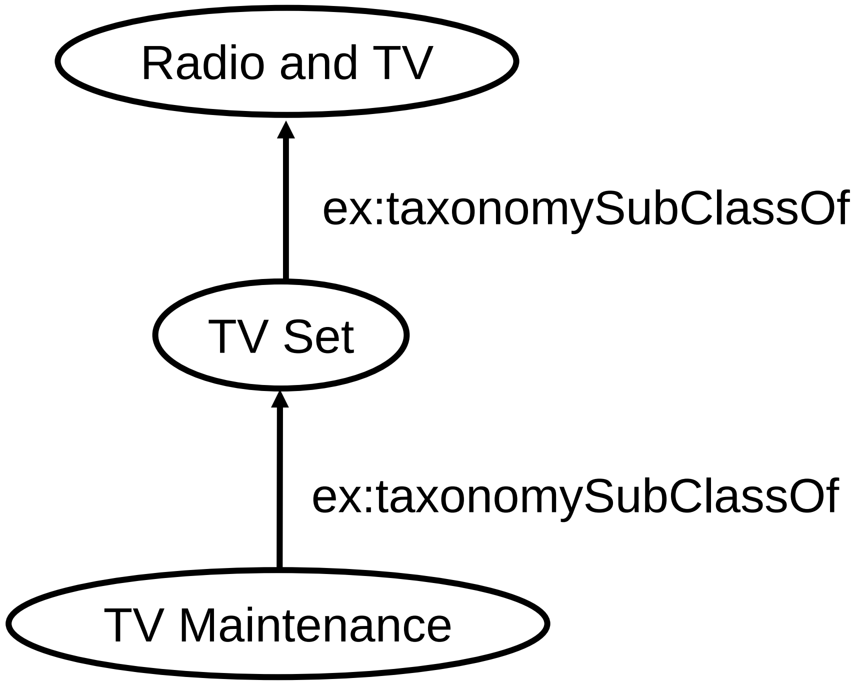
Taxonomic Concept

(Example: “Home appliances *and related goods*”)



[cf. 3,4]

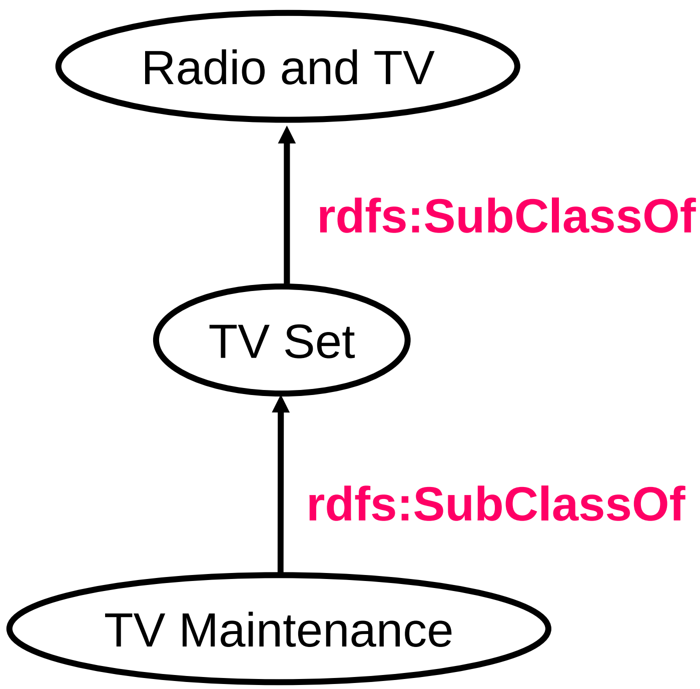
Good Representation (but impossible in OWL)



FORALL
A ex:taxonomySubClassOf B AND
B ex:taxonomySubClassOf C
→ A ex:taxonomySubClassOf C

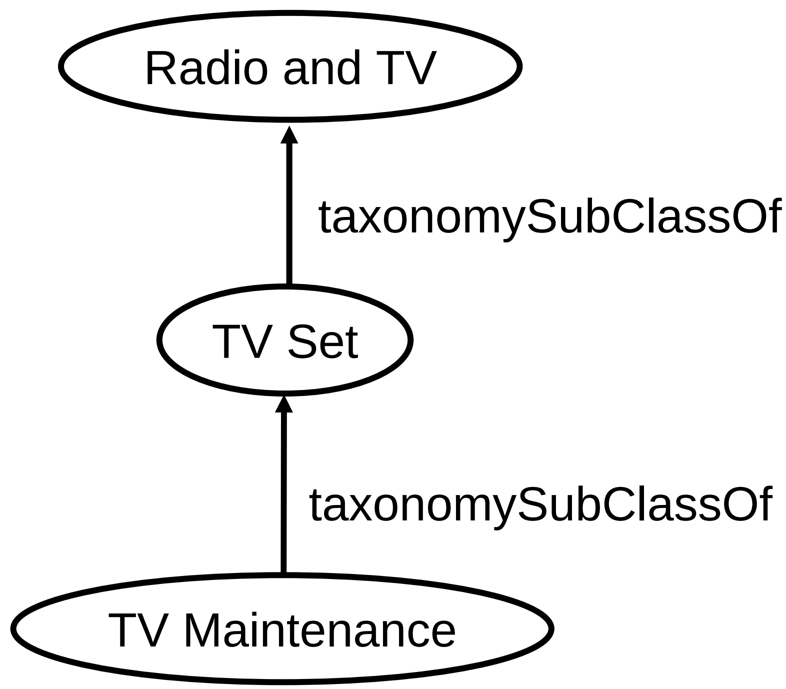
[cf. 3,4]

Naïve Approach
(leads to ontologies limited in use)



[cf. 3,4]

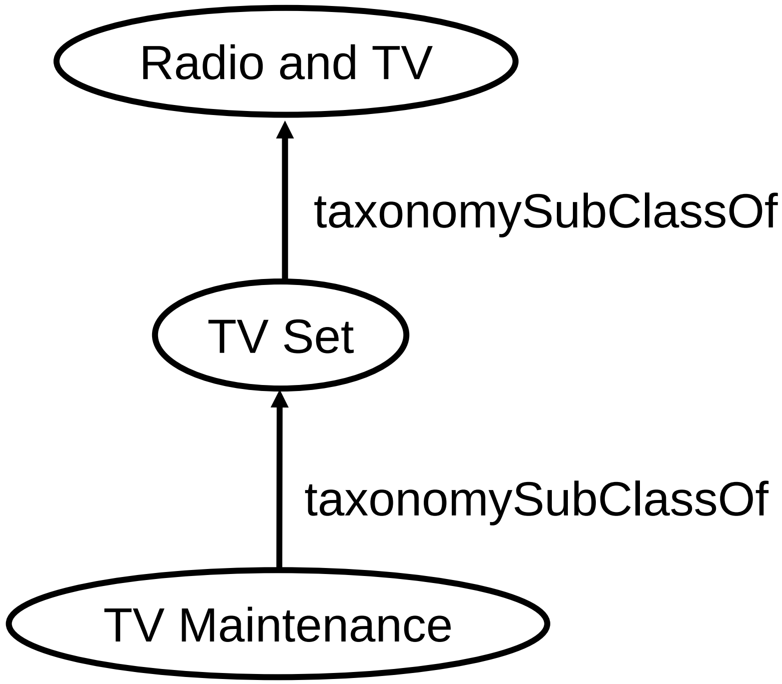
Workaround 1 in OWL: Lack of Transitivity and OWL Full



<owl:AnnotationProperty rdf:about=“ex:taxonomySubClassOf"/>

[cf. 3,4]

Workaround 2 in OWL: Categories as Instances



<owl:TransitiveProperty rdf:about="ex:taxonomySubClassOf"/>

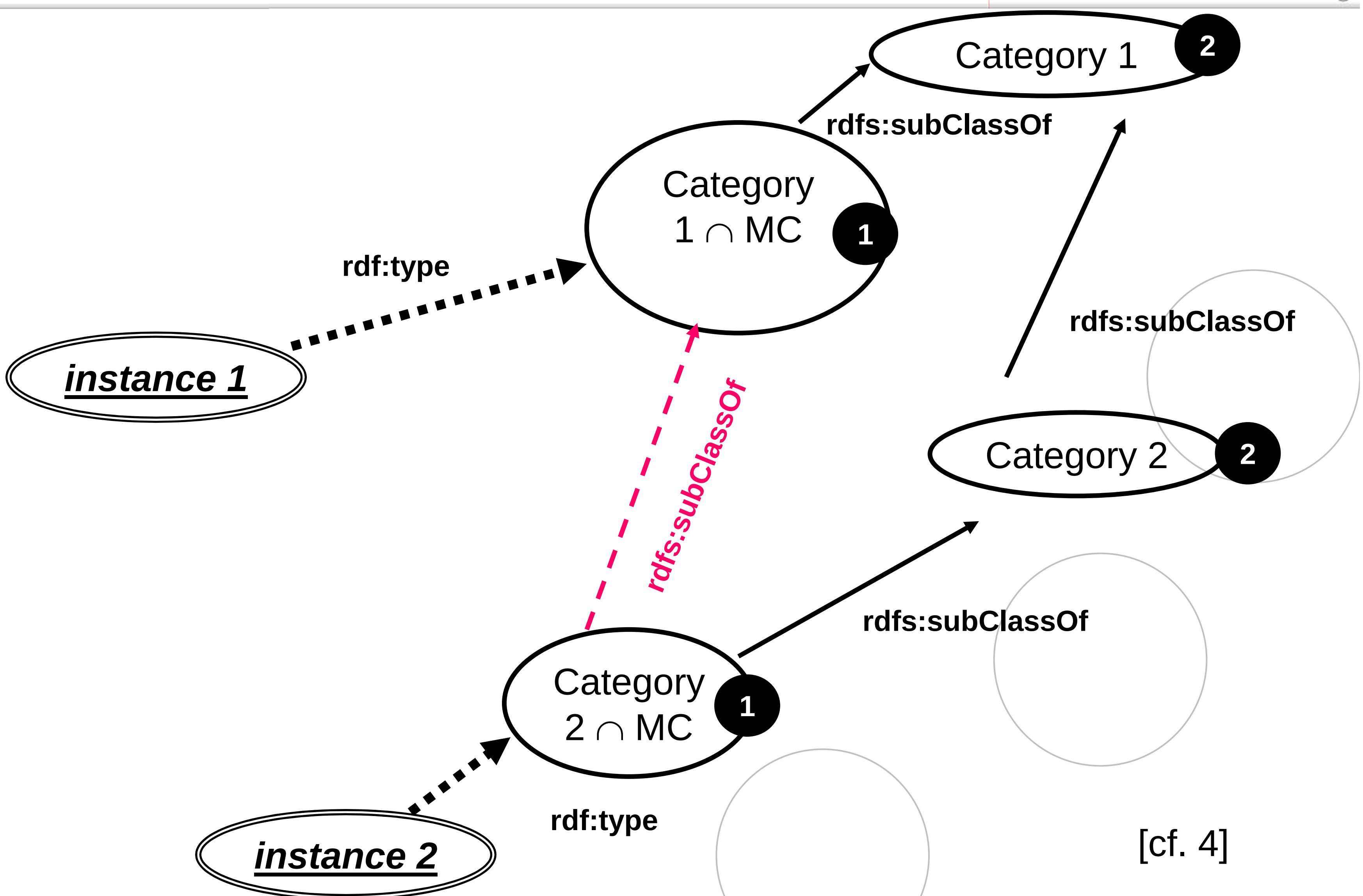
[cf. 3,4]

The Gen/Tax Approach

- We (=humans) interpret the original categories
  - in the original context of the hierachichal ordering and
  - in the desired target context of useful ontology classes
- We check several properties of these concepts with regard to the original hierarchy by statistical means.
- We represent each original categories from the input specification as TWO ontology classes
  - gen: Generic class in the target context
  - tax: Taxonomic class in the original context
- Where consistent, we represent the original hierarchical relations by subclassOf relations.

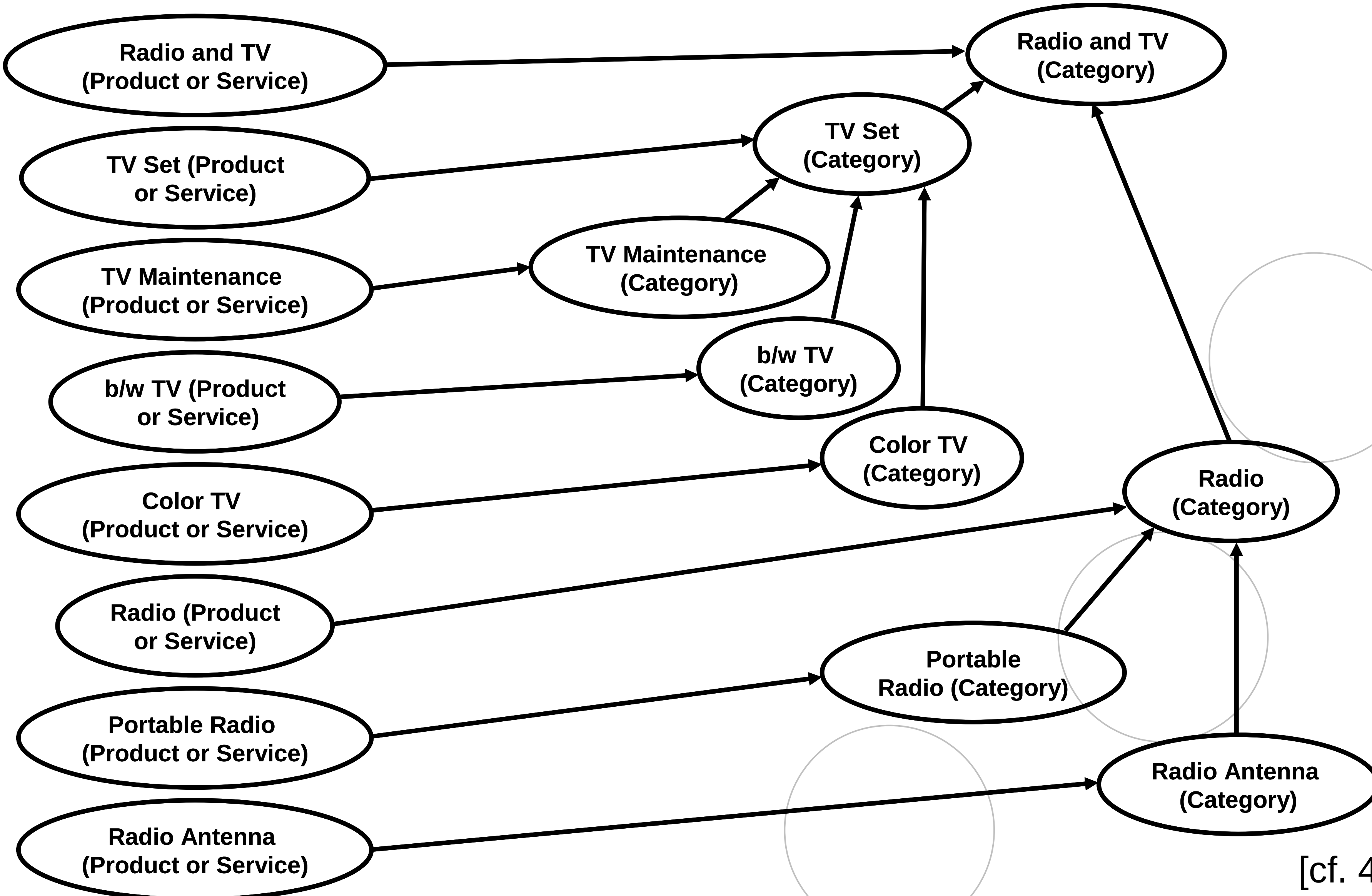
[cf. 4]

Gen/Tax Approach



[cf. 4]

Example


Radio and TV
(Product or Service)
TV Set (Product
or Service)
TV Maintenance
(Product or Service)
b/w TV (Product
or Service)
Color TV
(Product or Service)
Radio (Product
or Service)
Portable Radio
(Product or Service)
Radio Antenna
(Product or Service)
Radio and TV
(Category)
TV Set
(Category)
TV Maintenance
(Category)
b/w TV
(Category)
Color TV
(Category)
Radio
(Category)
Portable
Radio (Category)
Radio Antenna
(Category)

[cf. 4]

Gen/Tax Evaluation

- Advantages
  - High degree of automation
  - Minimal requirements on the ontology formalism
  - Reasoner can be used to exploit the original hierarchical order
- Disadvantages
  - increased number of classes (2 per category)
  - often no subsumption hierarchy for the generic concepts

3. Combine statistical approaches with declarative approaches

- E.g. predict missing concepts or relations

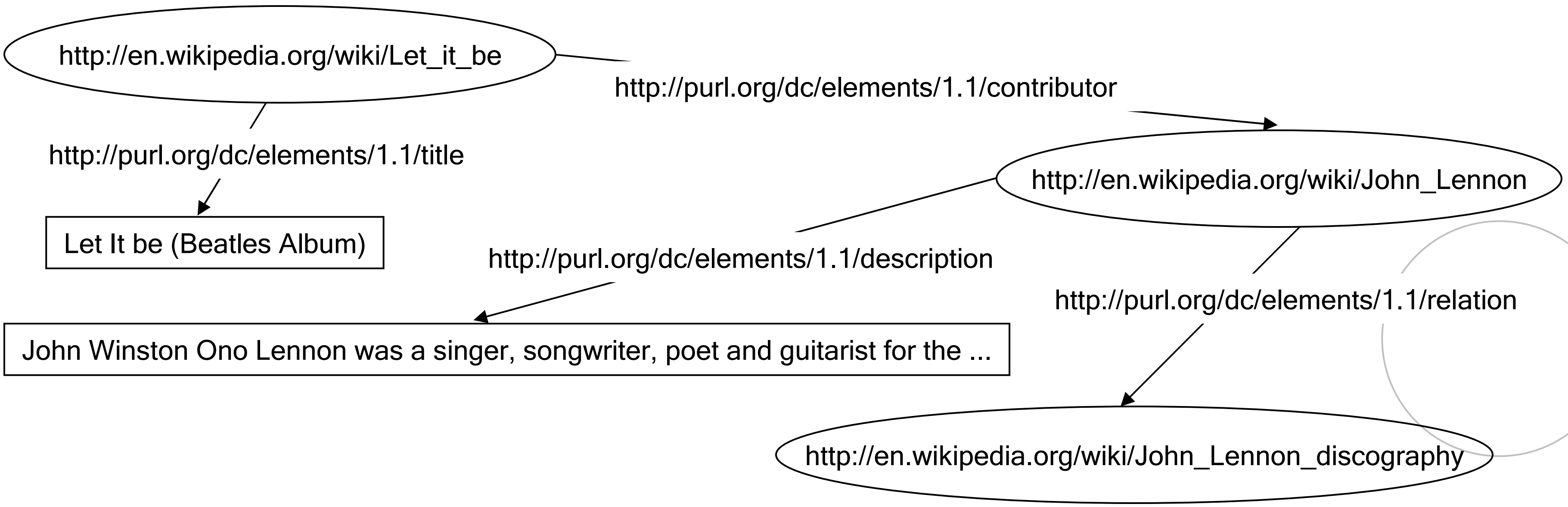
4. Community-driven Ontology Engineering

- Re-use Wikipedia consensus
- Wiki-like ontology engineering environments with low entrance barriers

- Observation: Though never meant as a „public space for conceptualizing domain theories“, Wikipedia is likely the largest set of consensually defined concepts identified by an URI
  - more than 1 Mio entries in the English version
- Idea: Use culture and technology of Wikis for collaborative ontology engineering for the masses.

[cf. 2]

Reusing Wikipedia Entries as Conceptual Entities



[cf. 2]

Analysis: Are the masses ☺ able to agree upon the definition of conceptual entities?

Methodology: We compared whether the concept or entity identified by the URI has changed significantly between the very first version and the current versions, in the sense that a layman annotation of a Web resource or a layman statement about the initial concept would hold for the first version but not for the current or vice versa. We distinguished the following cases:

- **Case 1a:** No significant change in meaning; the entry has been a stable, regular concept from its very first version to the current one.
- **Case 1b:** The entry has always been a Wiki "disambiguation page". It refers to a stable concept (i.e. all homonyms that could be referred to by this name).
- **Case 2:** A minor change in meaning has occurred. An example is that "Gloucester Courthouse" initially referred to the town and now refers to the "census designated place", which is still the same for many purposes.
- **Case 3a:** There was a major change in meaning.
- **Case 3b:** The URI was a regular entry in the beginning but turned into a disambiguation page later.

[cf. 2]

Results: Amount of URIs in the sample (n=100) that have turned into disambiguation pages

Disambiguation pages		
URI refers to a regular concept	**URI has always been a disambiguation page**	**URI became a disambiguation page during its lifespan**
92	5	3

[cf. 2]

Results: Amount of significant changes in meaning between an initial and the current version of Wikipedia entries

Significant changes in meaning between initial and current version of Wikipedia entries				
Case 1: None		Case 2: Minor	Case 3: Major	
1a: Stable, regular concept	1b: Always a disambiguation page	Slight change in meaning	3a: Major change in meaning	3b: URI became a disambiguation page
89	5	1	2	3

[cf. 2]

III. Example: The Making of eClassOWL

Goal

- Create an OWL Lite/DLP Ontology that preserves as much of the original semantics as possible
- Based on **Gen/Tax** (version 1)

eClassOWL

The Products and Services Ontology

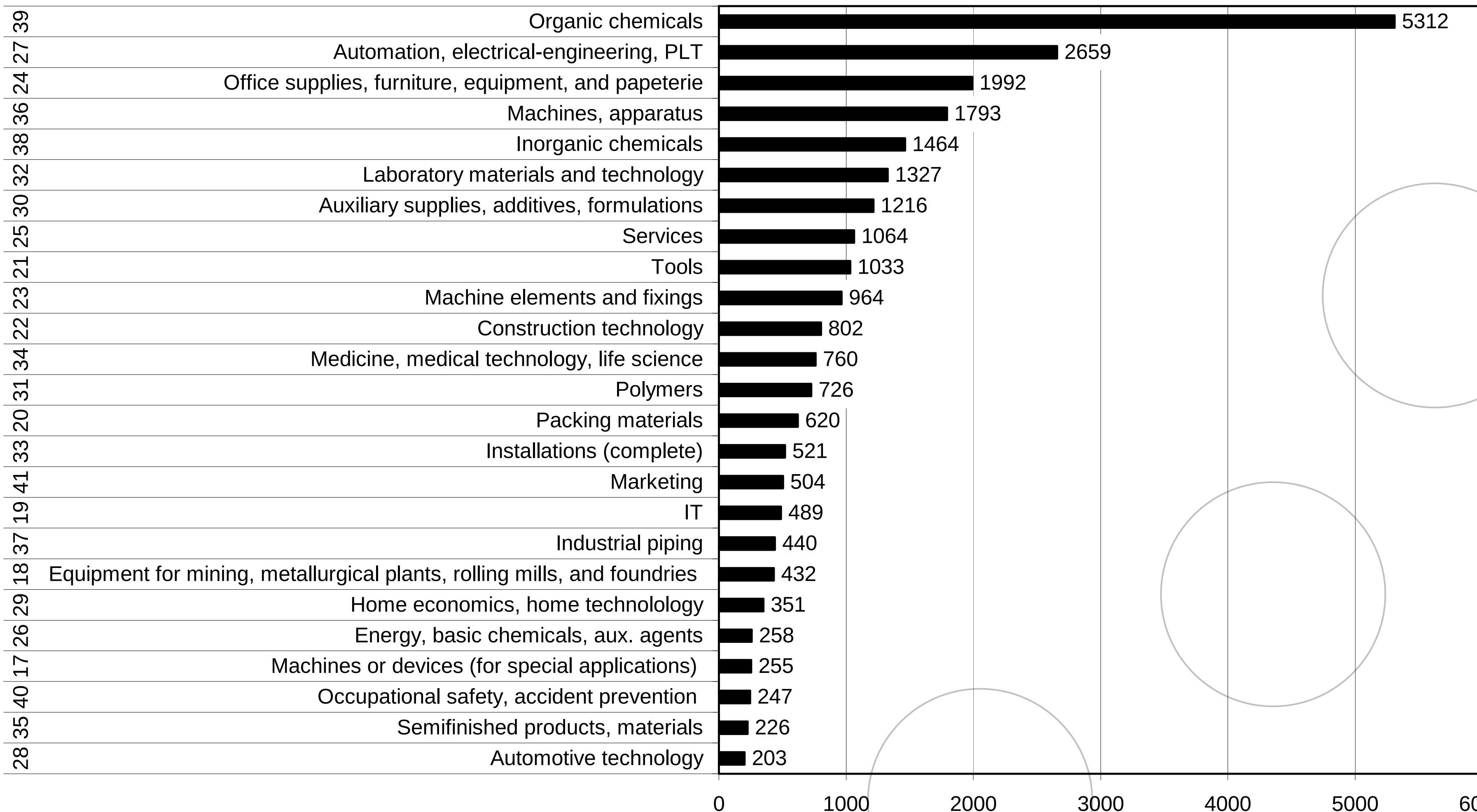
http://www.heppnetz.de/eclassowl/

eClassOWL Content

Type of Element	Number
Product categories	25,658 (each represented by 3 ontology classes, see below)
Properties for describing products and services instances	5,525
DatatypeProperties	*3,232*
ObjectProperties	*2,293*
Value Instances	4,544
Product categories that have a property recommendation	21,100
Total number of class-property statements	403,859
ObjectProperties that have a value recommendation	2,293
Total number of property-value statements	10,000

eClassOWL: Scope

eCl@ss 5.1de: Total Number of Classes by Top-level Categories



[cf. 1,6]

Summary of challenges

- Import of CSV in RDBMS
- Base URI and concept identifiers
- Detection of ObjectProperties
- Conversion of eCl@ss datatypes into xsd types
- Capturing property and value recommendations
- XML character encoding issues (&, ‘, “,)

[cf. 3]

Split in three files

eclass.owl

```
3 classes per concept
DatatypeProperties
ObjectProperties
instances for values
```

eclassClassesProperties.owl

```
<owl:AnnotationProperty rdf:about="&pcs;recommendedProperty"/>
<owl:Class rdf:ID="C_AKJ647001">
        <pcs:recommendedProperty rdf:resource="&pcs;P_AAA001001"/>
        <pcs:recommendedProperty rdf:resource="&pcs;P_AAA003001"/>
</owl:Class>
```

eclassPropertiesValues.owl

```
<owl:AnnotationProperty rdf:about="&pcs;recommendedValue"/>
<owl:ObjectProperty rdf:ID="P_XYZ001001">
        <pcs:recommendedValue rdf:resource="&pcs;V_AAA028001"/>
</owl:ObjectProperty>
</rdf:RDF>
```

[cf. 3]

Experiences

- Taxonomies should be imported into RDBMS, since we need iterative queries.
- Writing a script was a bit a tedious, but much reuse can be expected.
- Eventually, fully mechanized transcript was possible.
- Resulting ontologies are very big
  - file size > 25 MB

[cf. 3,4]

Versioning

- There are sometimes slight changes in meaning between releases.
- It is not always possible to automatically determine identity between concepts in any two releases.
- Thus, we favor explicit sets of identity statements, i.e.
  - owl:sameAs for property values in ecl@ssOWL
  - owl:sameClass for the classes and
  - owl:sameProperty for the properties.

[cf. 3]

Conclusions

- Industrial Categorization Standards contain a lot of valuable concepts, but were created for specific purposes and not with the rigor of knowledge representation in mind.
- Deriving an ontology is often not a one-time activity, but a continuous task.

Conclusions (2)

- When taking taxonomies for an ontology, the interpretation of the taxonomic relationship determines the intension of the concepts and vice versa.
- The interpretation of the taxonomic relationship is an important modeling decision.

Semantic Web tools and APIs: Ready for the real world? (Jena)

Protege.exe

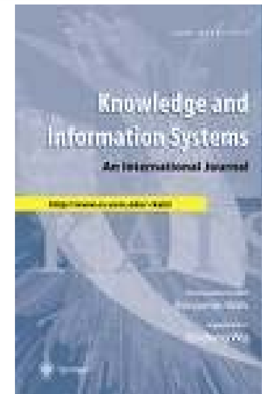
```
CONFIG: Loaded plugin edu.stanford.smi.protege.jot - JOT Plugin
CONFIG: Loaded plugin edu.stanford.smi.protegex.standard_extensions - Graph Widget, Table Widget
CONFIG: Loaded plugin edu.stanford.smi.protege.stringsearch - String Search Tab
CONFIG: Loaded plugin edu.stanford.smi.protegex.pal_tabs - PAL Tabs
CONFIG: Loaded plugin edu.stanford.smi.protegex.ezpal - EZPal Tab
CONFIG: Loaded plugin edu.stanford.smi.protegex.facets_tab
CONFIG: Loaded plugin edu.stanford.smi.protegex.instance_tree
CONFIG: Loaded plugin edu.stanford.smi.protegex.prompt - PROMPT tab
CONFIG: Loaded plugin edu.stanford.smi.protegex.psm - PSM Librarian
CONFIG: Loaded plugin edu.stanford.smi.protegex.rdf_backend - RDF Backend
CONFIG: Loaded plugin edu.stanford.smi.protegex.umls_tab - UMLS Tab
CONFIG: Loaded plugin uk.ac.man.cs.mig.coode.owlviz
CONFIG: Loaded plugin uk.ac.man.cs.mig.coode.owlwizard - OWL Wizard
Starting to load OWL stream at Wed Nov 03 16:52:49 EST 2004
java.lang.OutOfMemoryError

Starting: Wizard Plugin, V0.3, build 11
Starting to load OWL stream at Wed Nov 03 16:52:49 EST 2004
java.lang.OutOfMemoryError

Starting: Wizard Plugin, V0.3, build 11
```

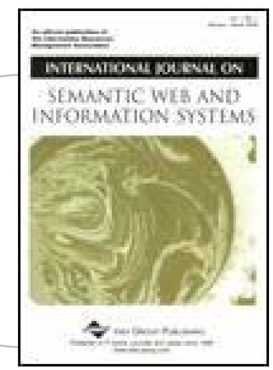
Further Information

- **Martin Hepp:** *Possible Ontologies: How Reality Constraints Building Relevant Ontologies, IEEE Internet Computing (forthcoming)*

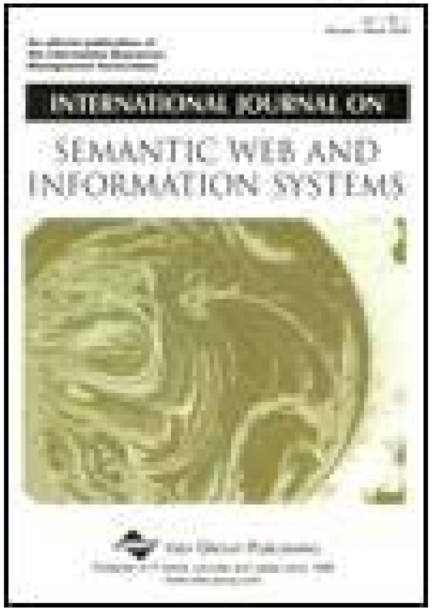
- **Martin Hepp,** Joerg Leukel, and Volker Schmitz: *A Quantitative Analysis of Product Categorization Standards: Content, Coverage, and Maintenance of eCl@ss, UNSPSC, eOTD, and the RosettaNet Technical Dictionary*, accepted for publication in: Knowledge and Information Systems (KAIS), Springer (forthcoming)

- **Martin Hepp:** *Products and Services Ontologies: A Methodology for Deriving OWL Ontologies from Industrial Categorization Standards,* *Int'l Journal on Semantic Web & Information Systems* (IJSWIS), Vol. 2, No. 1, pp. 72-99, January-March 2006.





The International Journal on Semantic Web and Information Systems



http://www.ijswis.org

References

[1] Martin Hepp, Joerg Leukel, and Volker Schmitz: *A Quantitative Analysis of Product Categorization Standards: Content, Coverage, and Maintenance of eCl@ss, UNSPSC, eOTD, and the RosettaNet Technical Dictionary*, accepted for publication in: *Knowledge and Information Systems (KAIS)*, Springer (forthcoming).

[2] Martin Hepp, Daniel Bachlechner, and Katharina Siorpaes: *Harvesting Wiki Consensus - Using Wikipedia Entries as Ontology Elements*, in: Proceedings of the 1st Workshop: SemWiki2006 - From Wiki to Semantics, co-located with the 3rd Annual European Semantic Web Conference (ESWC 2006) , June 12, 2006, Budva, Montenegro.

[3] Martin Hepp: *Products and Services Ontologies: A Methodology for Deriving OWL Ontologies from Industrial Categorization Standards* , in: *Int'l Journal on Semantic Web & Information Systems* (IJSWIS), Vol. 2, No. 1, pp. 72-99, January-March 2006.

[4] Martin Hepp, Daniel Bachlechner, and Katharina Siorpaes: *OntoWiki: Community-driven Ontology Engineering and Ontology Usage based on Wikis*, in: Proceedings of the 2005 International Symposium on Wikis (WikiSym 2005), October 16-18, 2005, San Diego, California, USA.

[4] Martin Hepp: *Representing the Hierarchy of Industrial Taxonomies in OWL: The gen/tax Approach* , in: Proceedings of the ISWC Workshop on Semantic Web Case Studies and Best Practices for eBusiness (SWCASE'05), November 7, Galway, 2005, Irland, pp. 49-56.

[5] Martin Hepp: *Possible Ontologies: How Reality Constraints Building Relevant Ontologies,* , accepted for publication in: *IEEE Internet Computing* (forthcoming)

[6] Martin Hepp: Güterklassifikation als semantisches Standardisierungsproblem, Deutscher Universitätsverlag, Wiesbaden 2003 (in German). http://www.duv.de/index.php?do=show&book_id=5295

Thank you!

Dr. Martin Hepp

http://sebis.deri.org
http://www.heppnetz.de

Martin Hepp
martin.hepp@deri.org