Making sense of Web content with Knowledge Patterns

Valentina Presutti
STLab ISTC-CNR, Italy

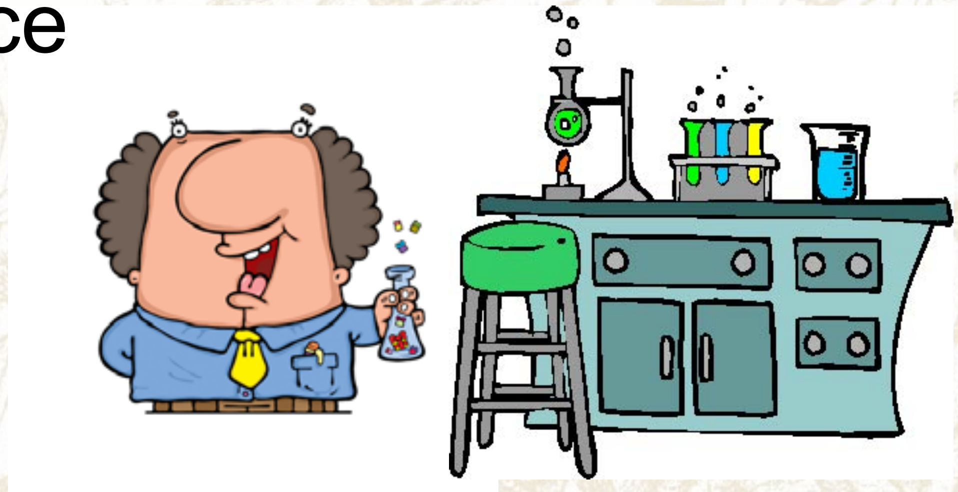
October 31st, 2012 Linköping

Heterogeneity: strength and weakness of the Web

- The Web is the g t source of open infor r
- Such on is iety of co
  – div evoluti

Semantic Web as an empirical science

- A large set of realistic data, created by large communities of practice
- We can perform experiments on it
- Semantic Web can be founded as an empirical science, as a branch of web science



Objects of an empirical science

- An empirical science needs clear research objects,
  - e.g. cells, proteins, or membranes are types of research objects in different branches of biology.
- and develops procedures for making patterns emerge out of the research objects

- Web of data, social network data, bibliographical, musical, and multimedia data, RDFa, Microformats, etc., provide an empirical basis to the Semantic Web, and indirectly to knowledge engineering

Identifying, selecting, constructing patterns from SW research objects

Two main problems

- Two main problems

1. *The knowledge soup problem*
   - The web of data is a knowledge soup because of the heterogeneous semantics of its datasets
   - Since people maintain and encode heterogeneous knowledge, how can formal knowledge be derived from the soup of triplified data?

Identifying, selecting, constructing patterns from SW research objects

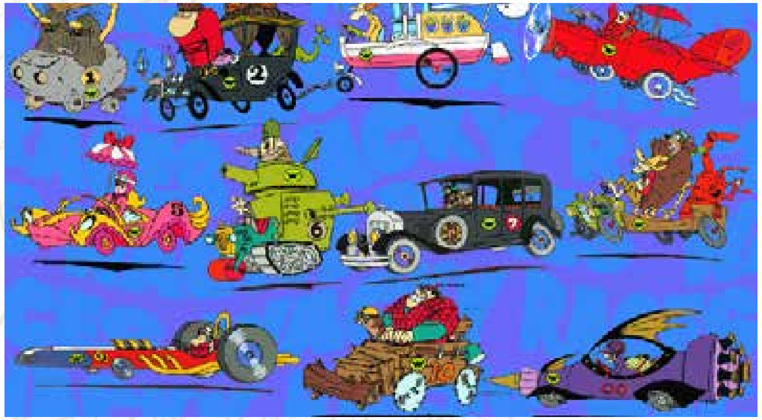
Two main problems

- Two main problems

2. The knowledge boundary problem

- How to establish the boundary of a set of triples that makes them meaningful i.e. relevant in context, so that they constitute a knowledge pattern?
- How the very different types of data (e.g. natural language processing data, RDFa, database tables, etc.) that are used by Semantic Web techniques contribute to carve out that boundary?

How do we recognize situations?



Foreground and background

- People tend to remember things because they stick out from the background (“profiling”)
  - but what makes a background as such?
- Expectations create scenarios
  - even things that are not there become part of the scenario if activated by an expectation
- Cf. Gestalt psychology (Köhler, Langacker, etc.)

Schema-based memory

- People tend to remember items that fit into a schema.
  - Things that are associated through some functional similarity (cf. Gibson's affordances)
- Schemata seem to be learnt mostly inductively
  - blocks world, repeated verbalization of invariant scenes, peek-a-boo, etc. Cf. Deb Roy's TED talk
- Schema similar to (conceptual) frame, script, knowledge pattern, etc.

Origin of modern frames and knowledge patterns

- «When one encounters a new situation (or makes a substantial change in one's view of the present problem) one selects from memory a structure called a **Frame**. This is a remembered framework to be adapted to fit reality by changing details as necessary … a frame is a data-structure for representing a stereotyped situation» (Minsky 1975)
- Frames, schemas, scripts … «These large-scale knowledge configurations supply top-down input for a wide range of communicative and interactive tasks … the availability of **global patterns of knowledge** cuts down on non-determinacy enough to offset idiosyncratic bottom-up input that might otherwise be confusing» (Beaugrande, 1980)

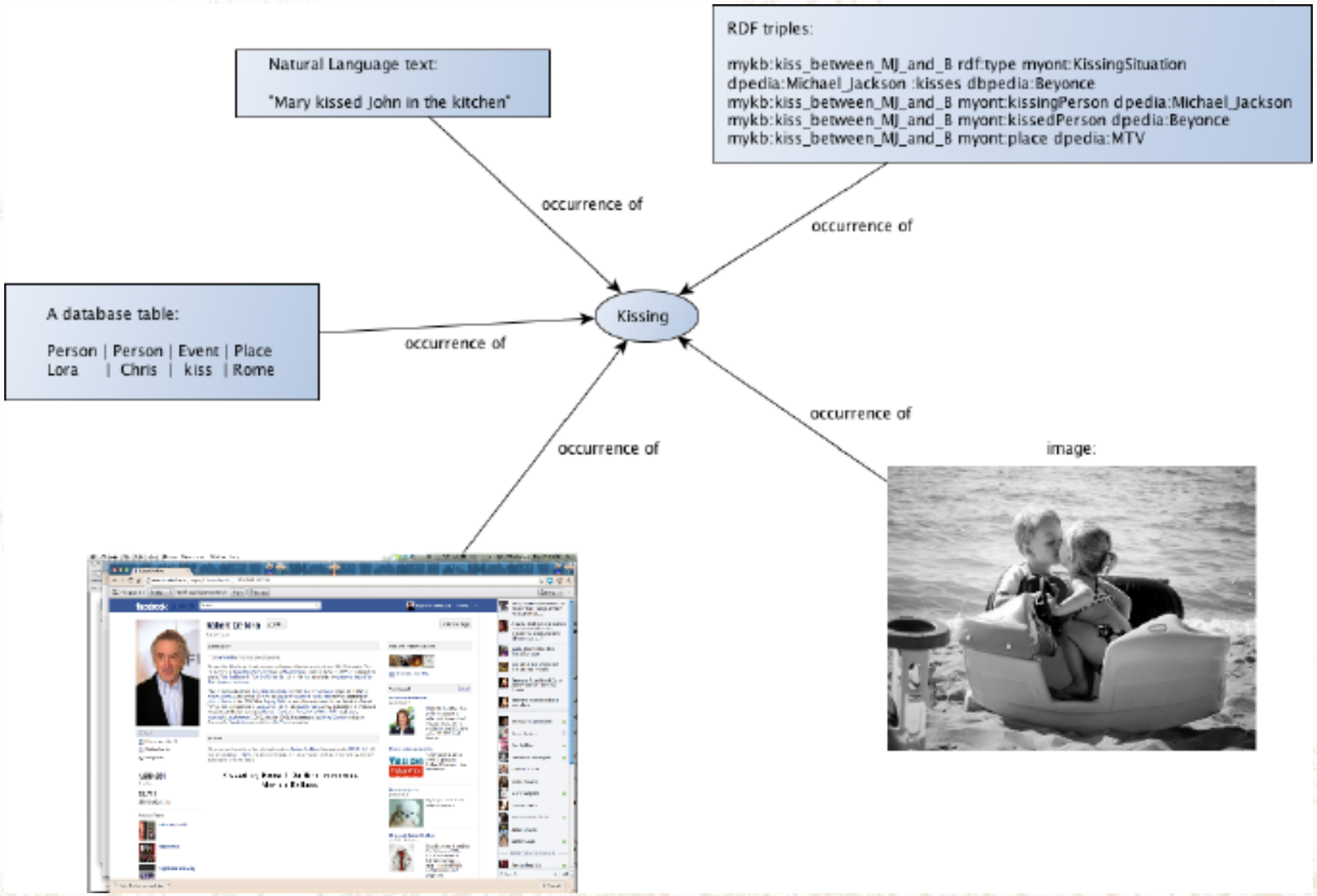
Knowledge patterns (KP)

- We suggest the usage of frames as the primary research objects over the Semantic Web, as opposed to simple concepts or binary relations, and we call them knowledge patterns.

Knowledge patterns (KP) are cognitively and pragmatically relevant conceptual structures capturing a piece of generic ontological or procedural knowledge. A knowledge pattern logically formalizes conceptual structures as composed of concepts and relations between them.

- KPs are an abstraction of data structures like frames in linguistics and artificial intelligence, microformats and microdata in Web technologies, association rules and patterns in data mining technologies, and ontology design patterns.

Knowledge Pattern: manifestations



How many KP?

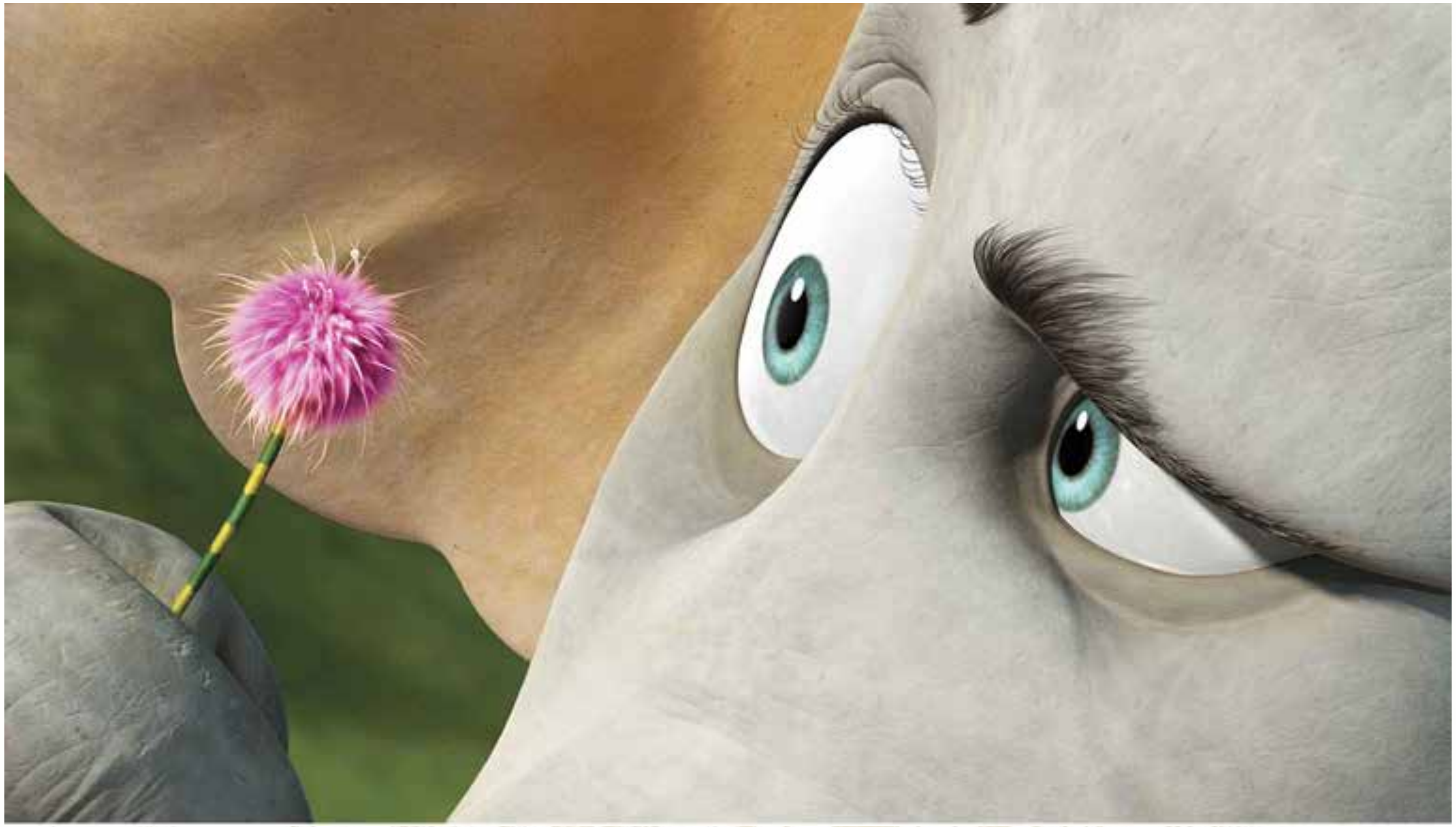
We (STLab) are researching on

- collecting
    - reengineering
        - aligning
            - and using

knowledge patterns as keys for accessing meaning of the Web

STLab research on KP

- Ontology design patterns
  - From linguistic frames, business models, database models, foundational and domain ontologies, etc.
- Pattern-based ontology design
- KP detection and discovery on linked data
- Frame-based machine reading and ontology learning
- KP-based knowledge extraction
  - Automatic entity typing, automatic link typing
- KP-based exploratory search
- …



LET'S HAVE A CLOSER LOOK…

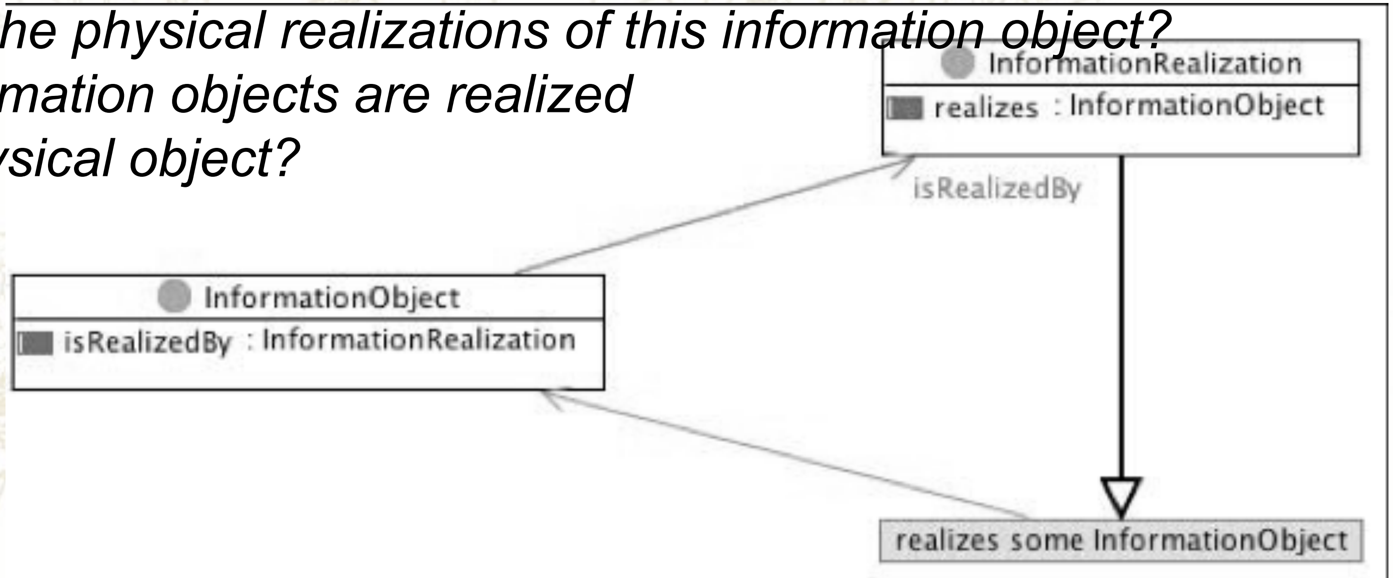
Top-down: expertise patterns

- Evidence that units of expertise are larger than what we have from average linked data triples, or ontology learning
  - Cf. cognitive scientist Dedre Gentner: "uniform relational representation is a hallmark of expertise"
  - We need to create expertise-oriented boundaries unifying multiple triples
  - "Competency questions" are used to link ontology design patterns to requirements:
    - Which objects take part in a certain event?
    - Which tasks should be executed in order to achieve a certain goal?
    - What's the function of that artifact?
    - What norms are applicable to a certain case?
    - What inflammation is active in what body part with what morphology?

Ontology Design Patterns (ODP)

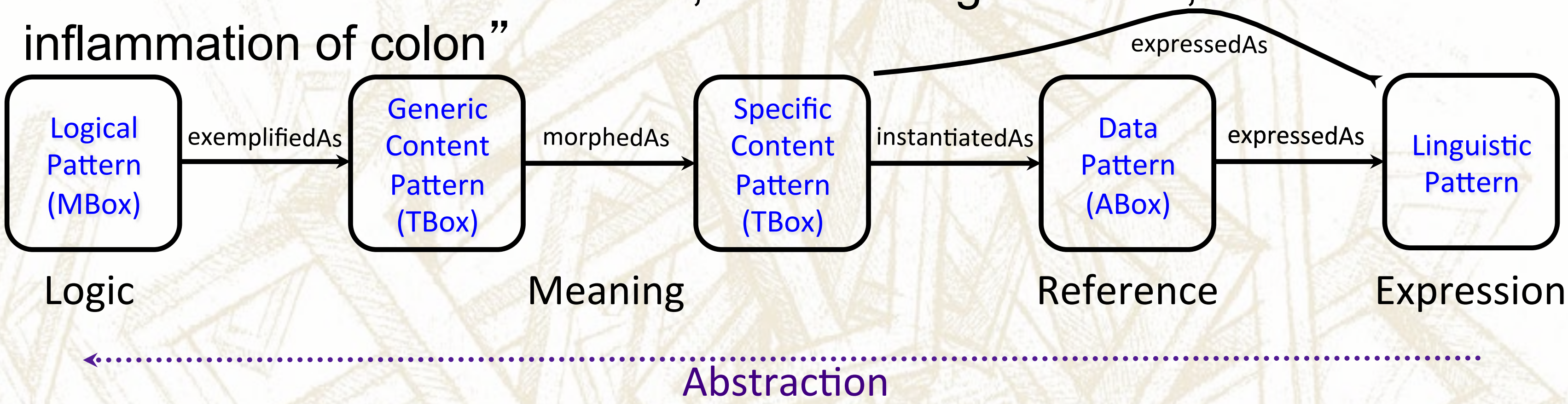
- A Content ODP is always associated with a General Use Case (GUC) expressed using Competency Questions (CQs)
- Example: InformationRealization

What are the physical realizations of this information object?
What information objects are realized by this physical object?



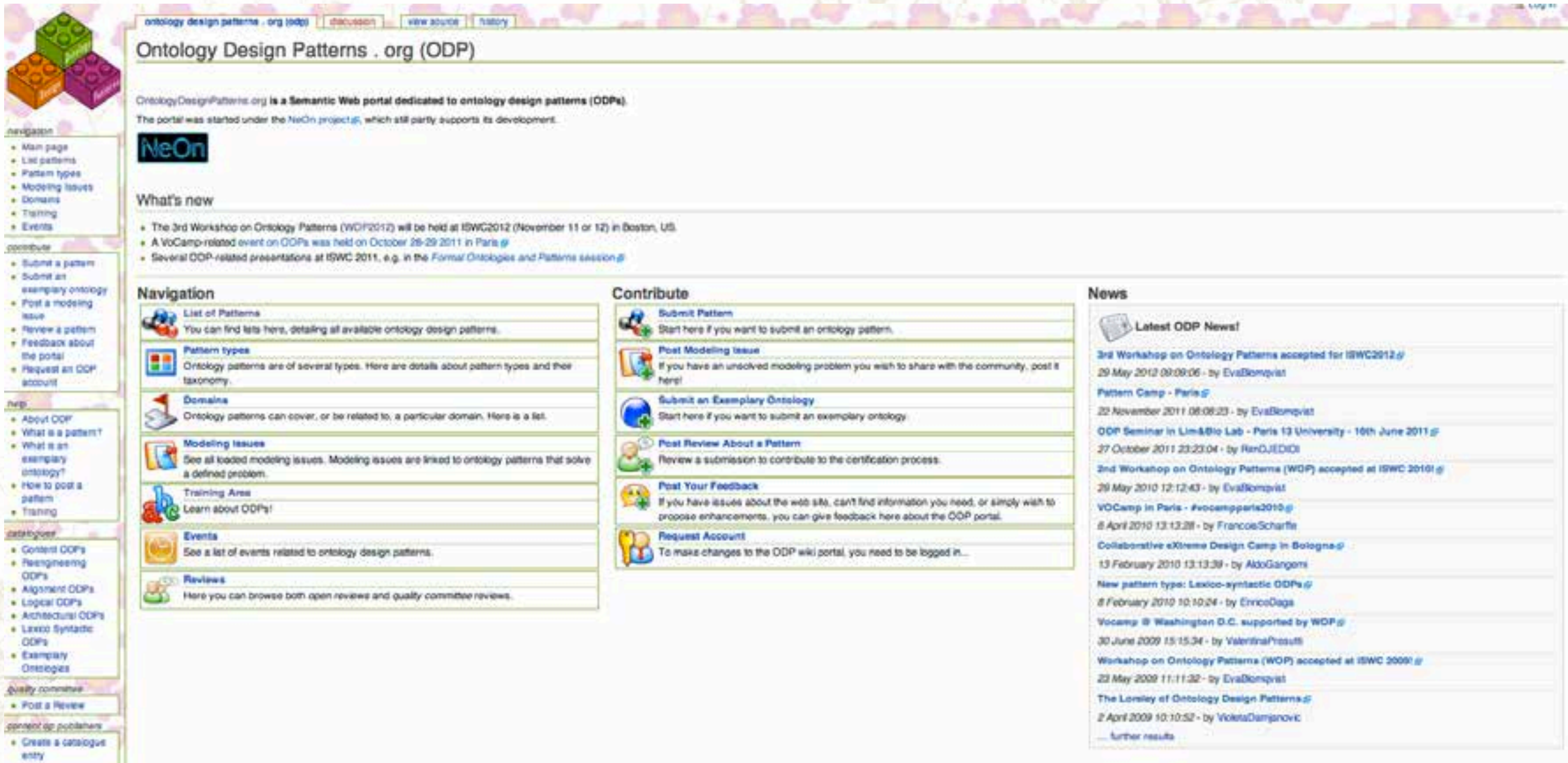
Layered pattern morphisms

- A logical design pattern describes a formal expression that can be exemplified, morphed, instantiated, and expressed in order to solve a domain modelling problem
- owl:Class:_:x rdfs:subClassOf owl:Restriction:_:y
- Inflammation rdfs:subClassOf (localizedIn some BodyPart)
- Colitis rdfs:subClassOf (localizedIn some Colon)
- John’s_colitis isLocalizedIn John’s_colon
- “John’s colon is inflammated”, “John has got colitis”, “Colitis is the inflammation of colon”



ODP

- Collected at http://www.ontologydesignpatterns.org
- Currently: Logical ODPs, Content ODPs, Re-engineering ODPs, and Alignment ODPs



User-study

45 users, 3 sessions, controlled experiments

Task 1 without ODP, Task 2 with ODP

1. Are Content ODPs perceived as useful?
2. Are the ontologies constructed using Content ODPs better, in some modelling quality sense, than the ontologies constructed without patterns?
3. Are the tasks given to the participants solved faster when using Content ODPs?
4. How do participants use the Content ODPs provided, and what support would be beneficial?

Neither methodological nor tool support available in 2009

Eva Blomqvist, Aldo Gangemi, Valentina Presutti:
Experiments on pattern-based ontology design. K-CAP 2009: 41-48

Ontology Evaluation

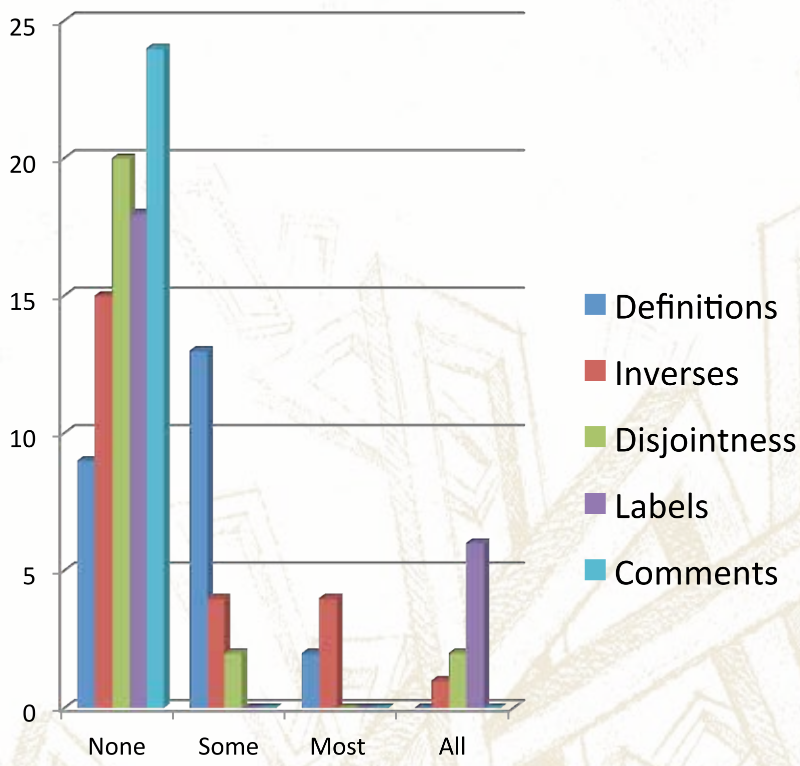
- Terminological Coverage

Task	Sessions	Coverage
Task 1	1,2,3	83.1%
Task 2	1,2,3	69.1%
Task 2	1,3	74.3%

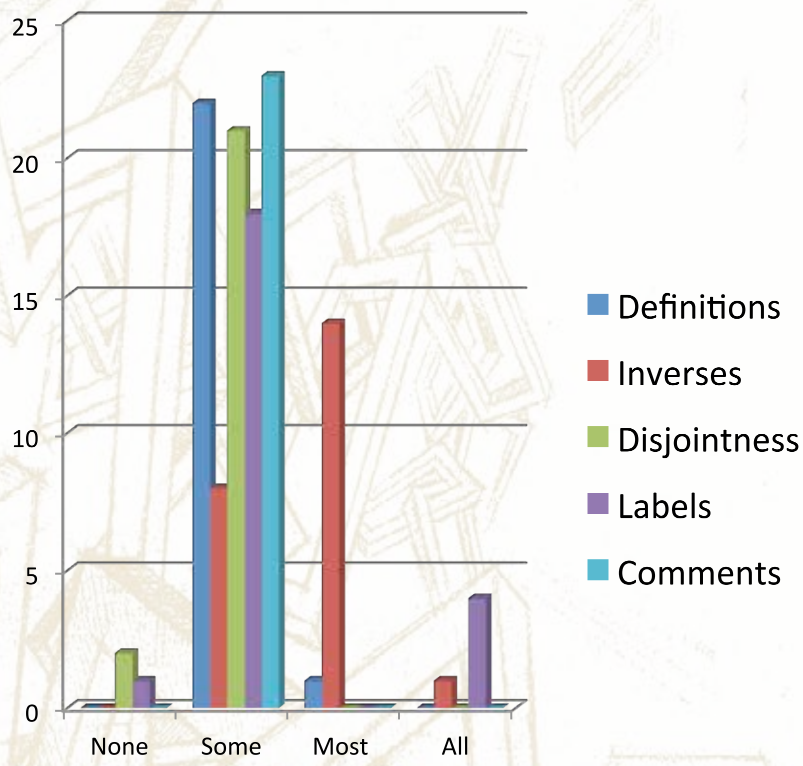
- Task Coverage

Task	Sessions	Measure	Coverage
Task 1	1,2,3	Opt. supported CQs	30.0%
		Supported CQs	65.2%
Task 2	1,2,3	Opt. supported CQs	39.9%
		Supported CQs	73.5%
Task 2	1,3	Opt. supported CQs	48.4%
		Supported CQs	88.3%

Ontology evaluation



Usability features of the ontologies (first task).



Usability features of the ontologies (second task).

Result overview

1. Are Content ODPs perceived as useful?
   – Yes!
2. Are the ontologies constructed using Content ODPs 'better', in some modelling quality sense?
   – Coverage of problem decreased (slower?) but major improvement in usability aspects, and fewer common mistakes
3. Are the tasks given to the participants solved faster when using Content ODPs?
   – Not really, rather slower (too little experience?)
4. How do participants use the Content ODPs provided, and what support would be beneficial?
   – How to find and select ODPs? How to reuse them? Tools?

XD Methodology and Tools (NTK)

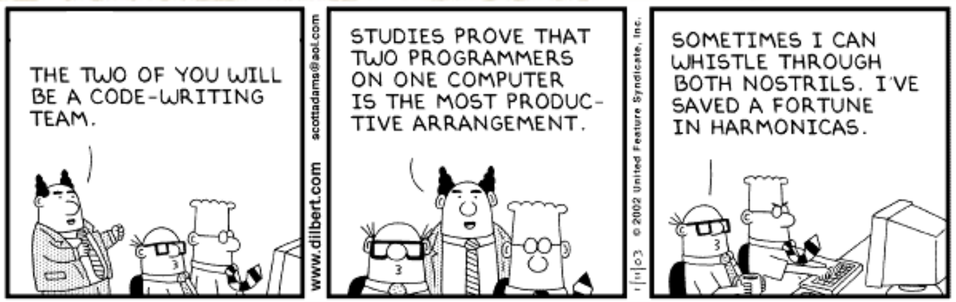
- eXtreme Design (XD)
  - an agile method for developing ontologies with Content Patterns
- XD tool
  - a tool that supports XD method
  - released as both an Eclipse plugin and a NeOn Toolkit plugin

ST LAB
SEMANTIC TECHNOLOGY LABORATORY

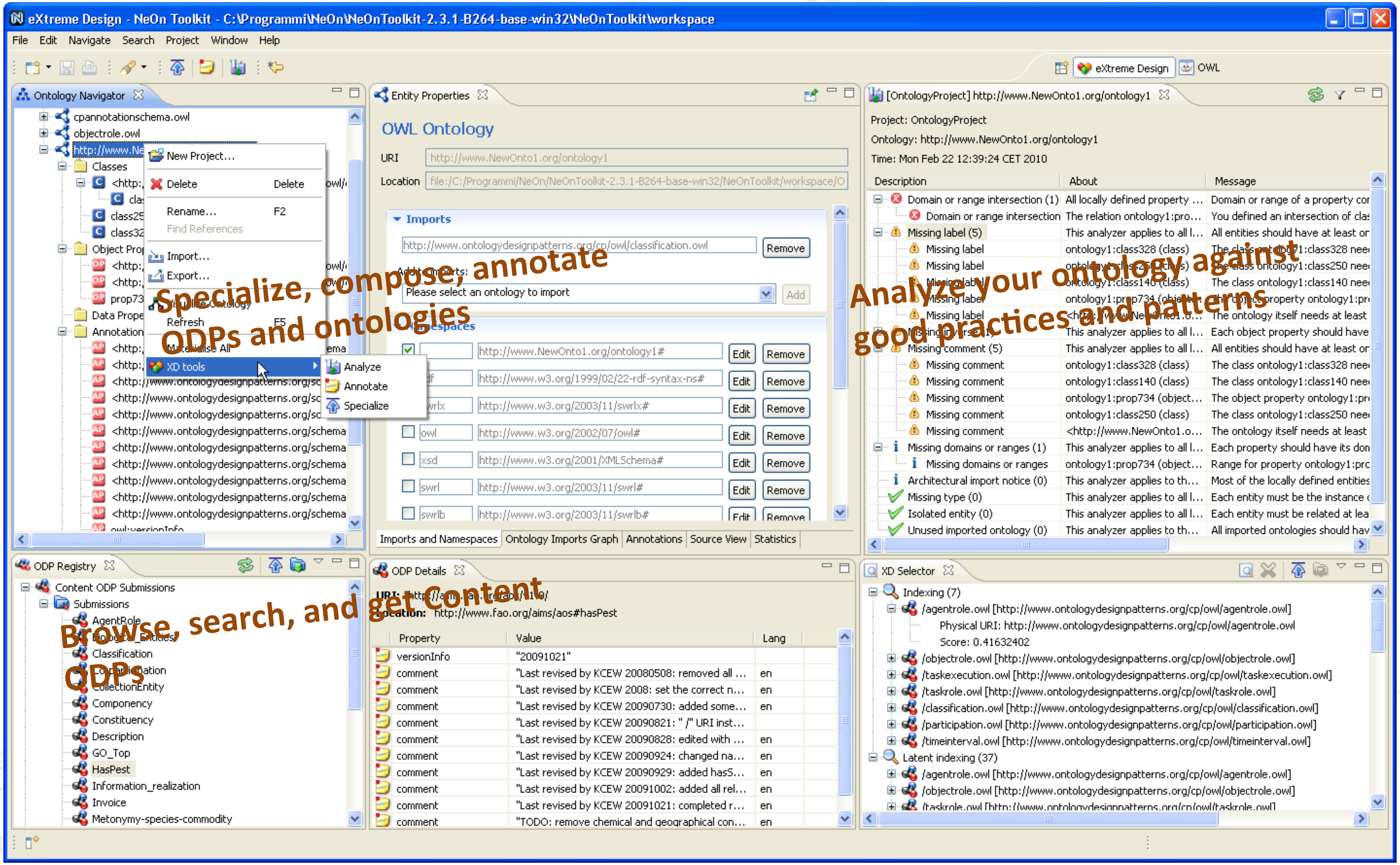
XD principles

- Customer involvement and feedback
- Customer stories, CQs and contextual statements

- CP reuse and modular design (ontology networks)
- Collaboration and integration
- Task-oriented design
- Test-driven design
- Pair design



XD Tools



eXtreme Design - NeOn Toolkit - C:\Programmi\NeOn\NeOnToolkit-2.3.1-B264-base-win32\NeOnToolkit\workspace
Ontology Navigator
Entity Properties
OWL Ontology
Specialize, compose, annotate ODPs and ontologies
XD tools
Analyze
Annotate
Specialize
Analyze your ontology against good practices and patterns
ODP Registry
ODP Details
XD Selector
Browse, search, and get Content ODPs

User-study

35 users, 2 sessions, controlled experiments

Task 1 no ODP, Task 2 ODP+XD Tools, Task 3 + XD methdology

- Summary of research questions:
  1. Can we confirm the results from the previous study (Questions 1-4)? + How is modularity affected?
  2. Does XD Tools support the process of reusing CPs?
  3. Does the XD methodology support the process of reusing CPs, and does it affect any of the aspects from the previous study (e.g., time, quality)?

Eva Blomqvist, Valentina Presutti, Enrico Daga, Aldo Gangemi: Experimenting with eXtreme Design. EKAW 2010: 120-134

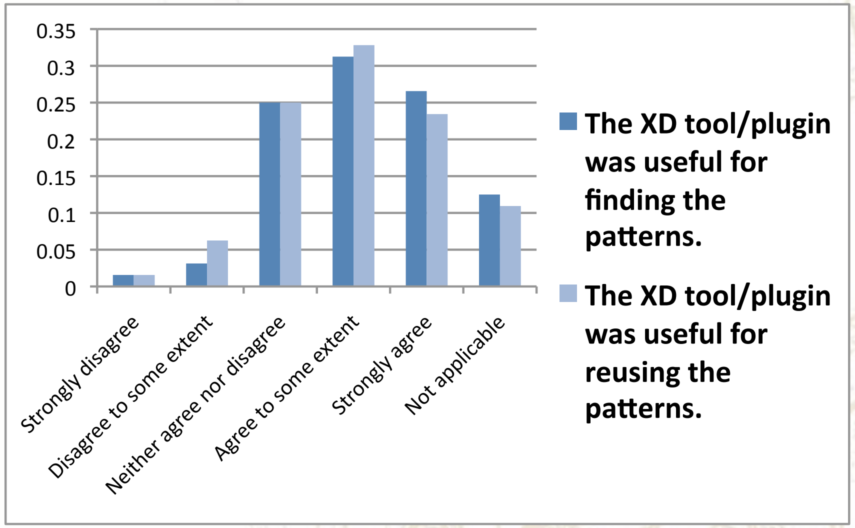
Results – Confirming previous conclusions?

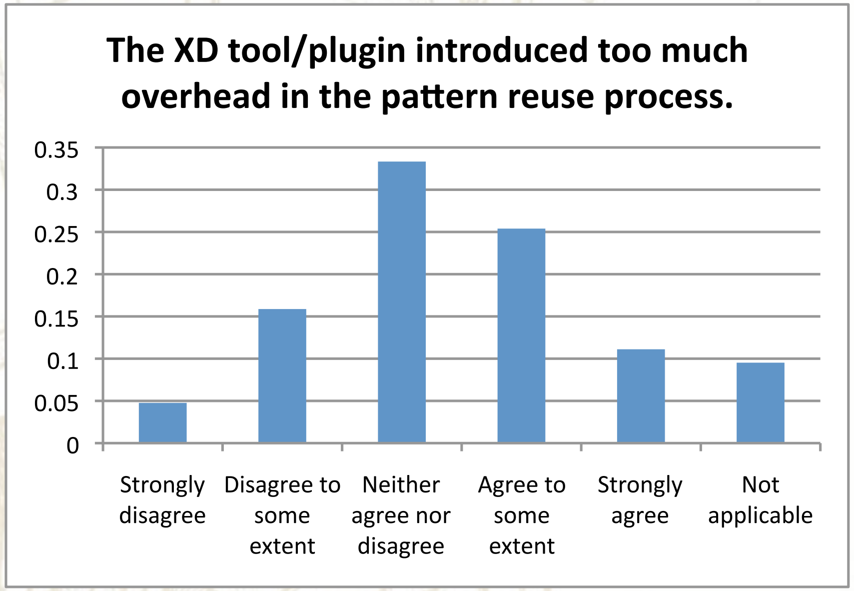
1. Are CPs perceived as useful by the participants?
   – Confirmed – Increase for second session: Due to XD Tools?
2. Are the ontologies constructed using CPs 'better', in some modelling quality sense?
   – Coverage: Reduction of terminological coverage is no longer detected – Due to XD Tools?
   – Usability: Confirmed – Most prominent improvement!
3. Are the tasks solved faster when using CPs?
   – With tool support: no longer slower and less mistakes
4. What common modelling 'mistakes' can be identified, when not using patterns and when using CPs?
   – Decrease in occurrence of most frequent mistakes confirmed (44% average decrease) - Same types of mistakes
   – Two types of errors decrease significantly more than the others:
     - N-ary relations – decrease by 64%
     - Missing datatype properties – decrease by 46%

Results - Modularity

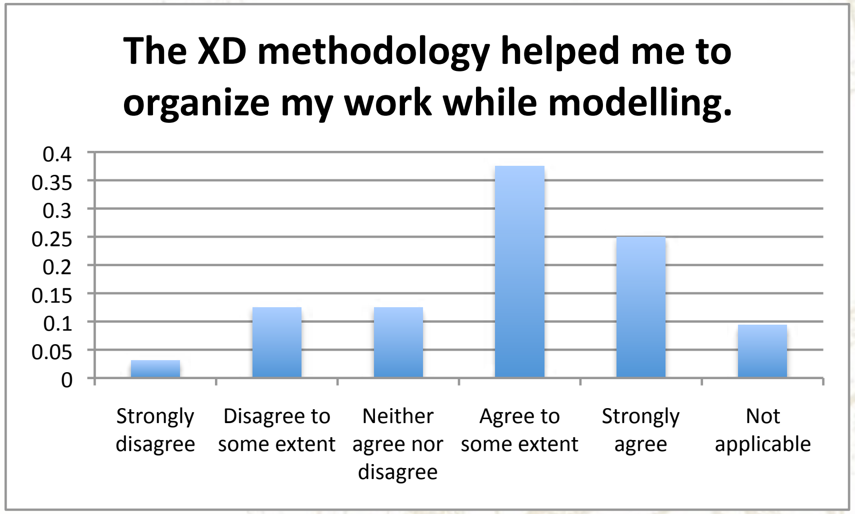
- Do CPs increase the modularity of ontologies?
  - Task 1: no ontologies are modularized
  - Task 2: the ontologies contain on average 7.5 modules
  - Conclusion: Since the participants choose to reuse the CPs as OWL-modules, rather than ideas for solutions, this inherently introduces modularity

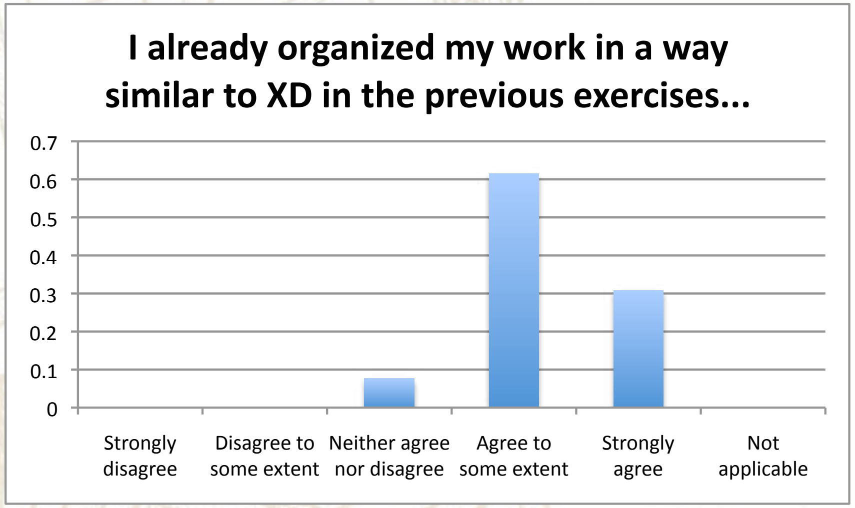
Results: XD Tools





Results XD: methodology



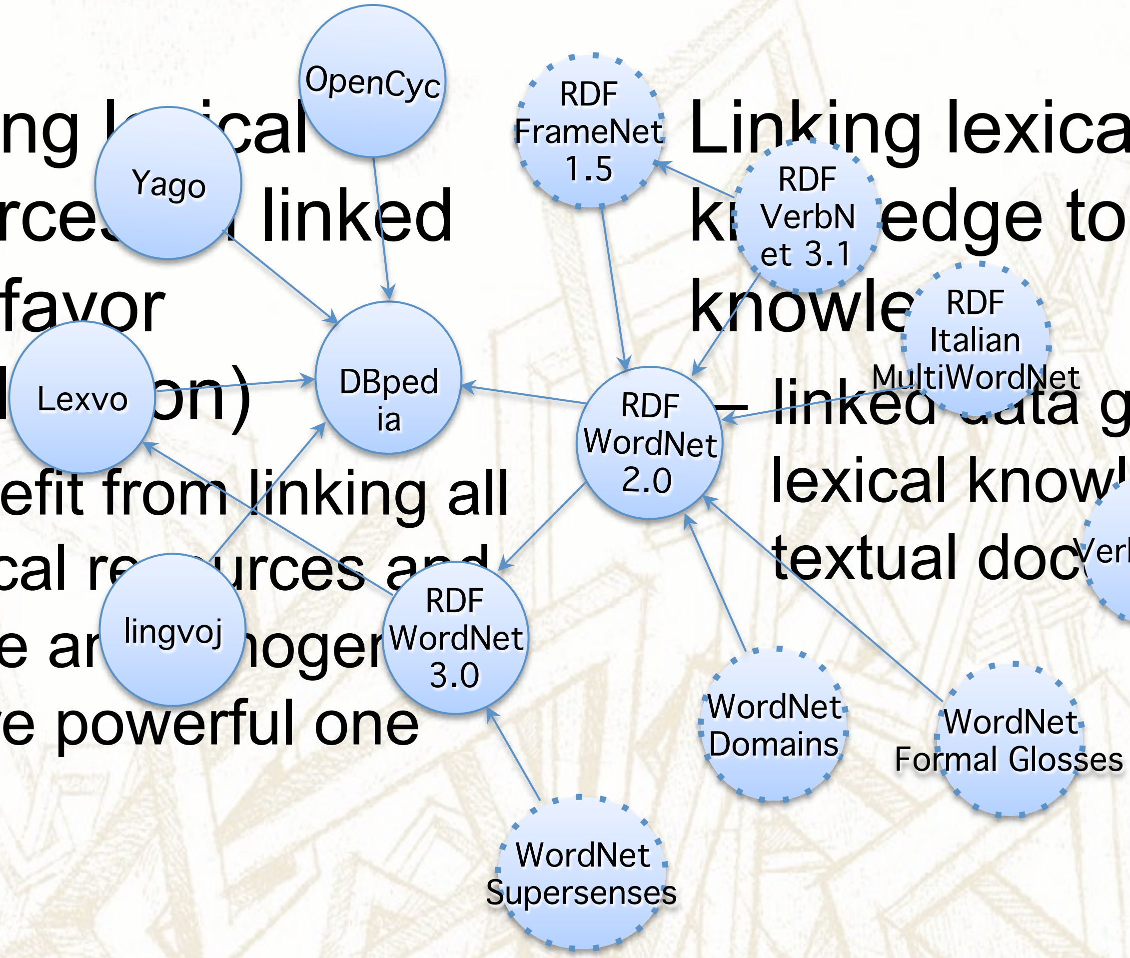


Currently

- Continuing working on ODP-based ontology design
- Testing methodology and XD Tools extension
- Experiments to be conducted

Eva Blomqvist, Azam Seil Sepour, Valentina Presutti: Ontology Testing – Methodology and Tool. EKAW 2012: 216-226

Top-down: FrameNetLOD



- Bringing l...ical resource... linked data (favor hybrid... on)
  - benefit from linking all lexical re... urces a... have ar... hoger... more powerful one
- Linking lexical k... edge to domain knowle...
  - linked ... ground to lexical knowl... ge and textual doc... ts

Andrea Giovanni Nuzzolese, Aldo Gangemi, Valentina Presutti: Gathering lexical linked data and knowledge patterns from FrameNet. K-CAP 2011: 41-48

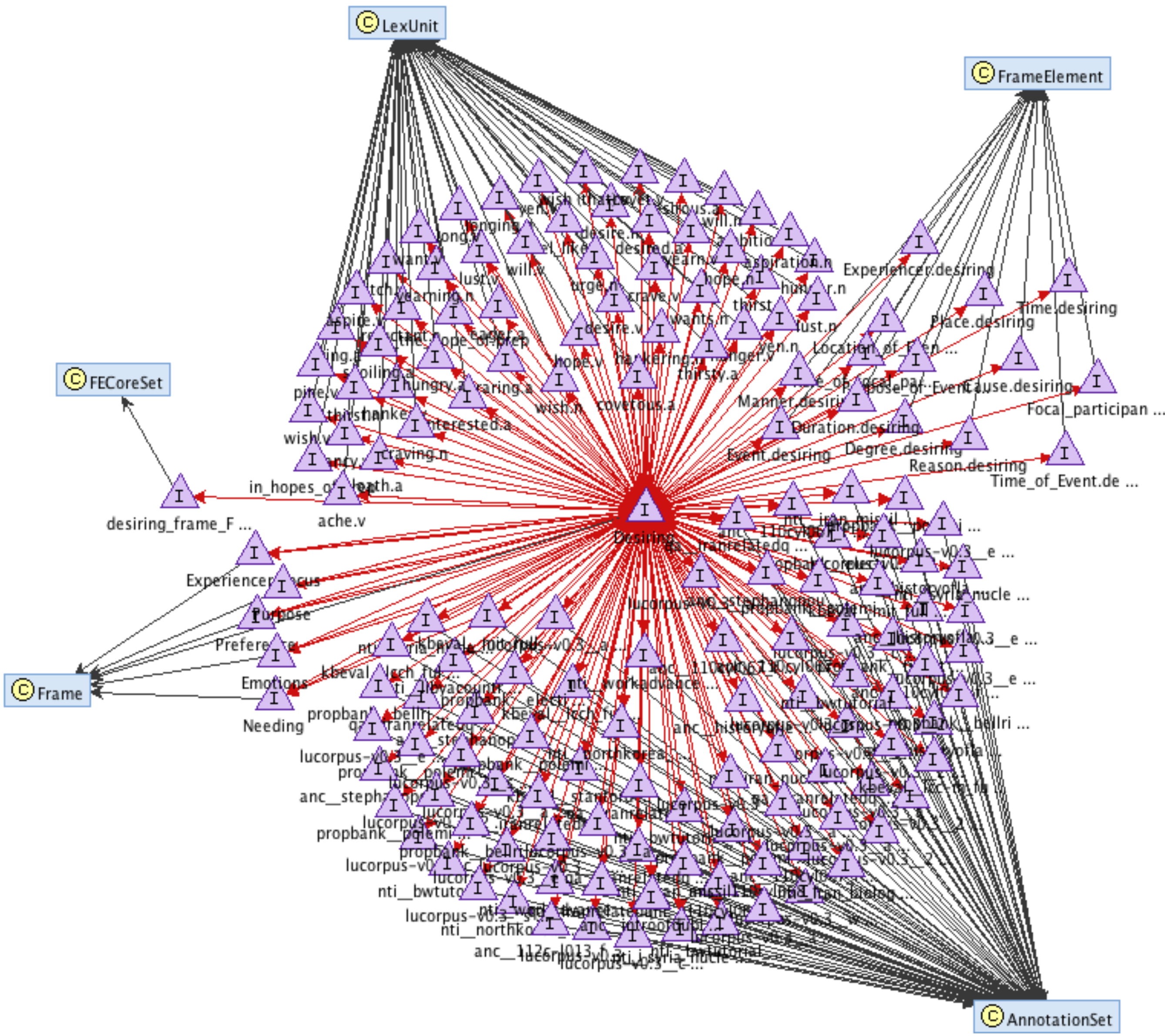
FrameNetLod

- There are many issues related to the conversion of lexical resources
  - more specifically to semantic issues of FrameNet conversion

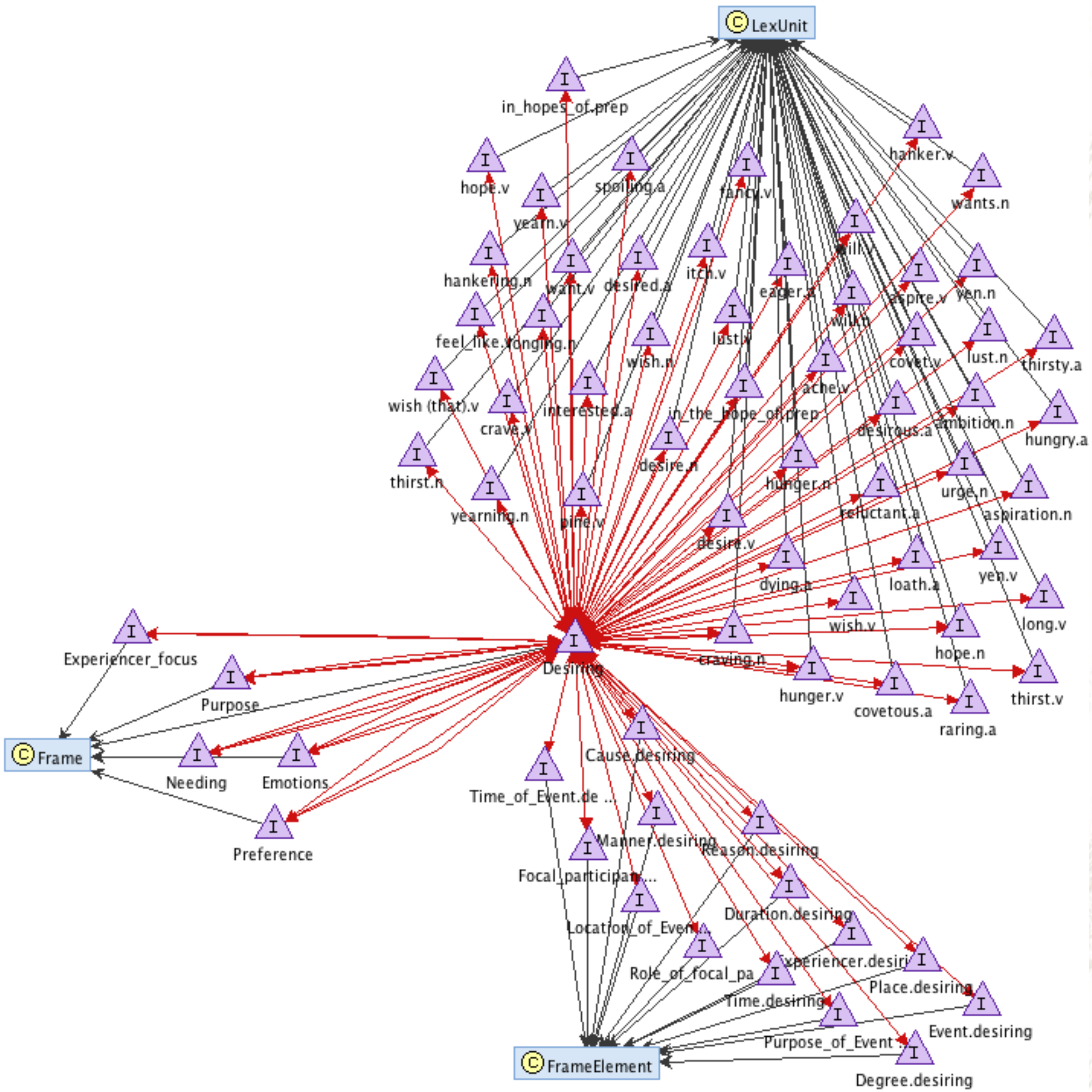
FrameNetLod provides

- A method to solve those issues (supported by a tool)
- A conversion of FrameNet to RDF published as a dataset in the LOD
- A method to convert FrameNet data into knowledge patterns

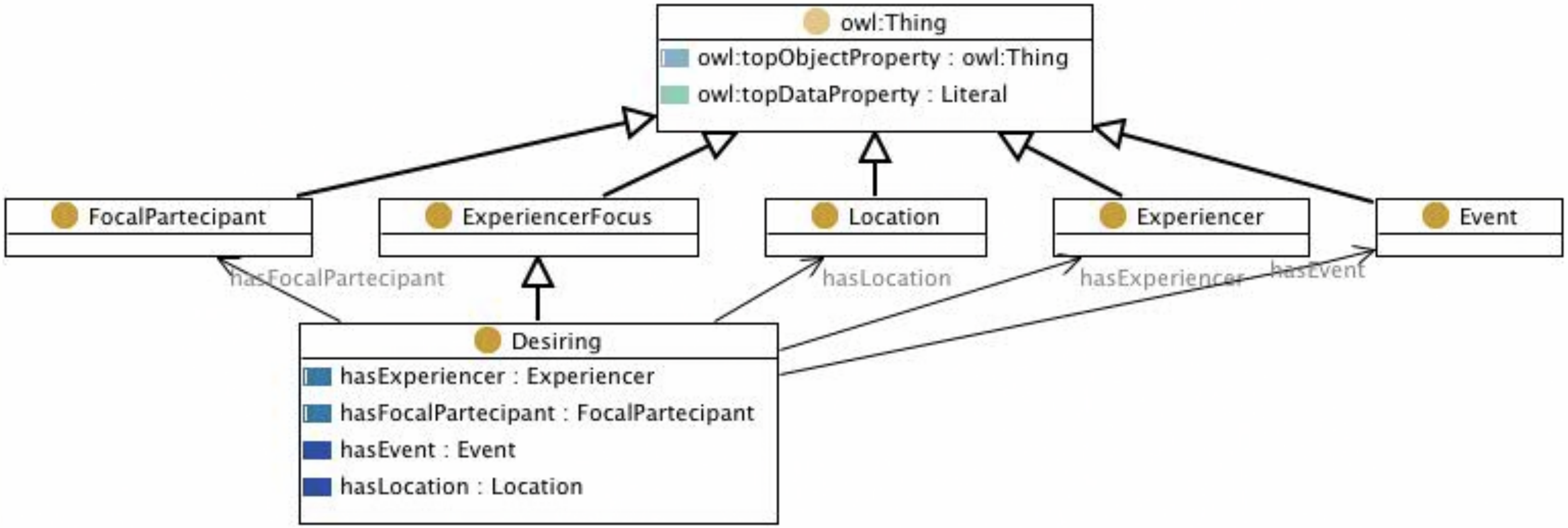
FrameNet as LOD



FrameNet as LOD



FrameNet as ontologies



BOTTOM-UP: ACCESSING TEXTUAL KNOWLEDGE

Robust ontology learning (ROL)

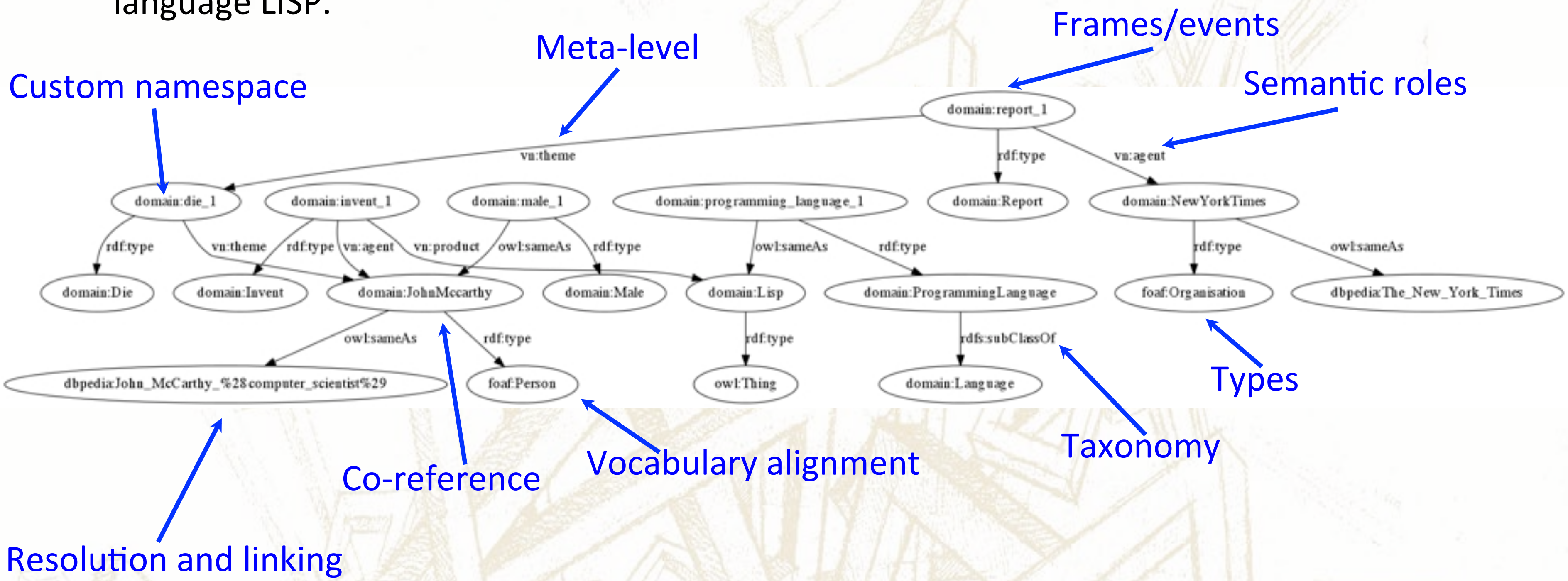
- Fast and accurate NL to RDF/OWL transformation
  - Mid-to-strong variety of machine reading
- Good design quality of the resulting ontologies
  - more than entity + relations: aggregated data (frames, events) vs. sparse data
- Frame-based representation ⟶ Good design quality
  - [Coppola et al., ESWC-2009] shows that n-ary ontology design patterns (ODPs) can be easily derived from frames, and have equivalent conceptual expressivity (and have formal semantics in addition)
  - [Blomqvist, ISWC-2009] provides evidence that OL methods performances improve if the learning cycle is augmented with ODPs

Requirements for ROL on the Web

- Ability to map natural language (Web of documents, still the major part) to RDF/OWL representations (Web of data)
- Ability to capture accurate semantic structures (e.g. complex relations or frames)
- Easy adaptation to the principles of linked data publishing (IRI, links)
- Minimal computing time

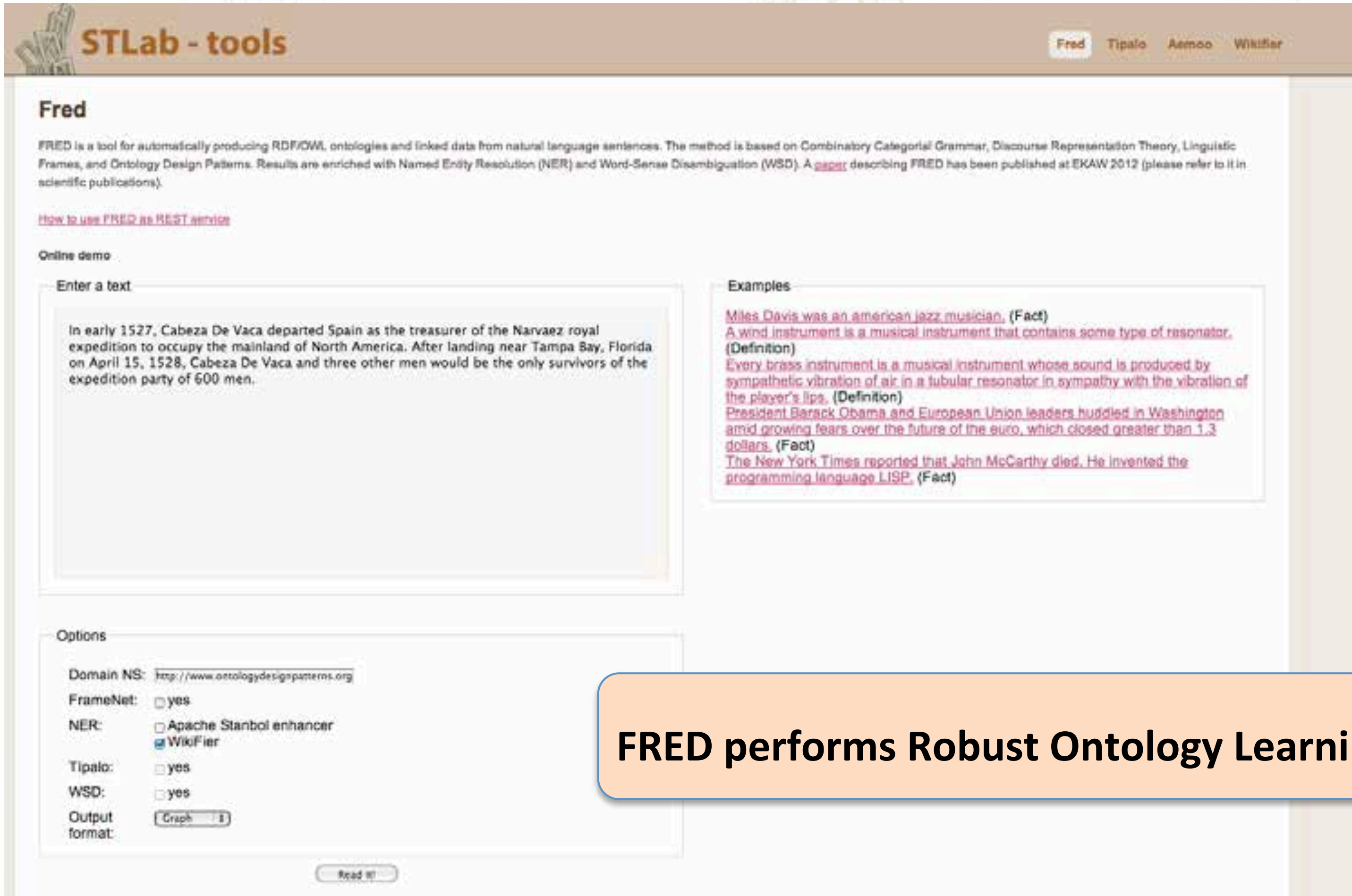
Machine reading with FRED

The New York Times reported that John McCarthy died. He invented the programming language LISP.



Valentina Presutti, Francesco Draicchio, Aldo Gangemi: Knowledge Extraction Based on Discourse Representation Theory and Linguistic Frames. EKAW 2012: 114-129

FRED on STLab tools[1]



FRED performs Robust Ontology Learning

(1) http://wit.istc.cnr.it/stlab-tools/fred/

FRED

- Based on Discourse Representation Theory and heuristics
- Produces OWL/RDF TBox and Abox
- Resolves entities on Linked Data
- Performs frame detection with a rule-based approach, with good performances
  - i.e. no training phase needed
- Is much faster as compared to other existing tools
- Quality of resulting ontologies?
  - Still to be rigorously evaluated but…

- We have demonstrated that this approach leads to promising results in large scale knowledge extraction

Tìpalo

- A FRED application
- Automatic typing of Wikipedia entities based on FRED
- Results are very good
- An indirect evaluation of FRED performances

Andrea Giovanni Nuzzolese, Aldo Gangemi, Valentina Presutti, Francesco Draicchio, Alberto Musetti, and Paolo Ciancarini: Automatic Typing of Dbpedia entities. International Semantic Web Conference (1) 2012 (To appear)

STLAB
SEMANTIC TECHNOLOGY LABORATORY

Tìpalo on STLab tool

STLab - tools

Fred · Tìpalo · Aemoo · Wikifier

Tìpalo

Tìpalo uses FRED and automatically assigns types to Wikipedia entities. Given a Wikipedia page URI, the tool returns an RDF graph composed of rdf:type, rdfs:subClassOf, owl:sameAs, and owl:equivalentTo statements providing typing information about the entity referred by the Wikipedia page.

How to use Tìpalo as REST service

Online demo

Enter a Wikipedia page URI:
e.g., http://en.wikipedia.org/wiki/Wind_instrument

Get types

Examples

Wind instrument (http://en.wikipedia.org/wiki/Wind_instrument)
Pakito (http://en.wikipedia.org/wiki/Pakito)
Neutron star (http://en.wikipedia.org/wiki/Neutron_star)
Alter ego (http://en.wikipedia.org/wiki/Alter_ego)
Lupercal (http://en.wikipedia.org/wiki/Lupercal)
Chaise longue (http://en.wikipedia.org/wiki/Chaise_longue)
French Revolution (http://en.wikipedia.org/wiki/French_Revolution)

Powered by:
C&C · Boxer · Apache Stanbol · Graphviz · Python · Pydot

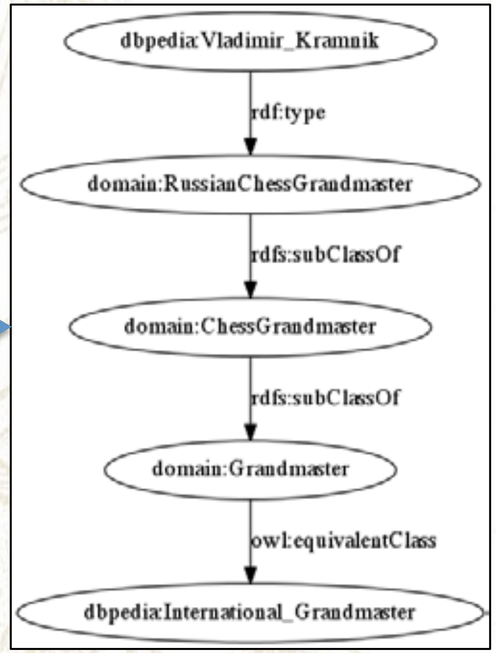
What does Tìpalo do?

- Goal: to guess the type of entities referred by Wikipedia, given their definition as provided by their Wikipedia page abstract

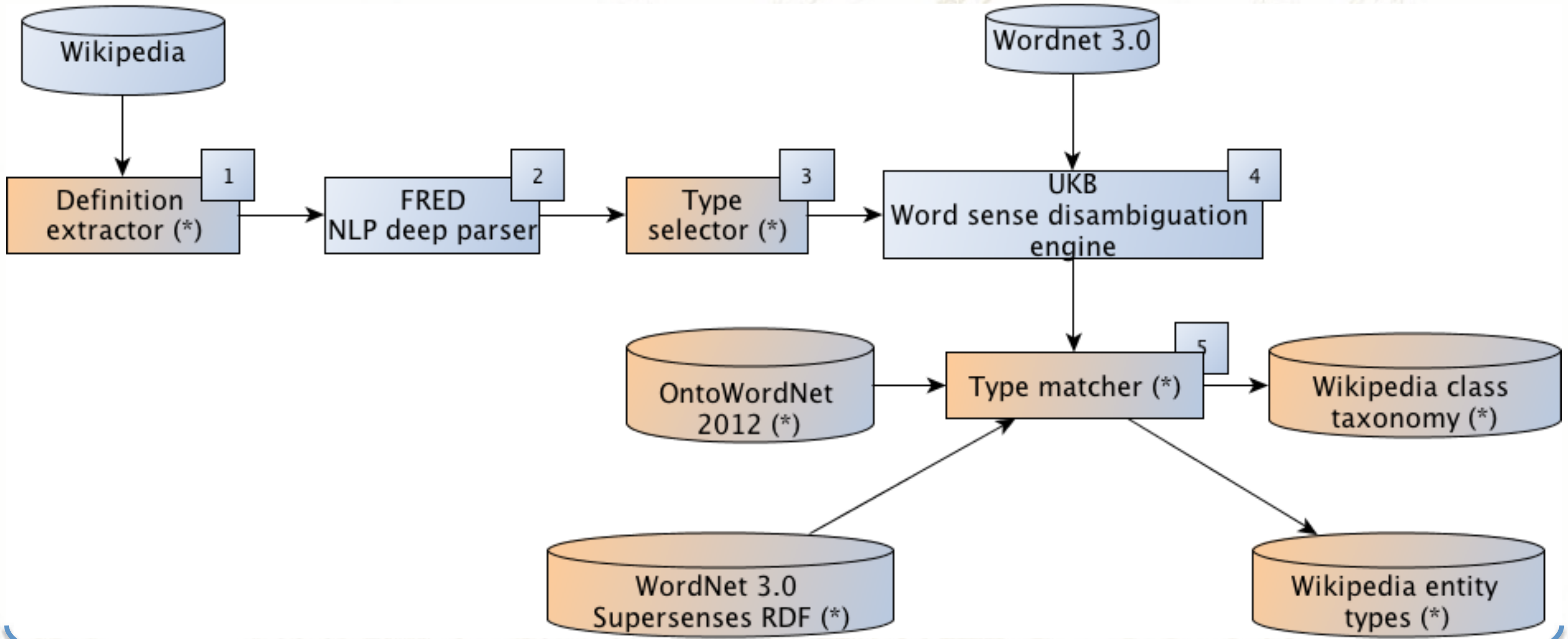
Vladimir Kramnik

From Wikipedia, the free encyclopedia

Vladimir Borisovich Kramnik (Russian: Владимир Борисович Крамник; born 25 June 1975) is a Russian chess grandmaster. He was the Classical World Chess Champion from 2000 to 2006, and the undisputed World Chess Champion from 2006 to 2007. He has also won the two strongest tournaments (by rating strength) in chess history: the 2009 Mikhail Tal Memorial and the 2010 Grand Slam Masters Final. He has won three team gold medals and three individual medals at Chess Olympiads.[2]



How does it do it?



Tìpalo

Performance of the individual components

Component	Precision	Recall	F-measure (F1)
Type selector	0.93	0.9	0.92
WSD (UKB)	0.86	0.82	0.84
WSD (most frequent sense)	0.77	0.73	0.75
Type matcher (Supersense)	0.73	0.73	0.73
Type matcher (DUL+/D0)	0.8	0.8	0.8

Performance of the overall process

Typing process	Precision	Recall	F-measure (F1)
WordNet types	0.76	0.74	0.75
Supersenses	0.62	0.6	0.61
Dul+/D0	0.68	0.66	0.67

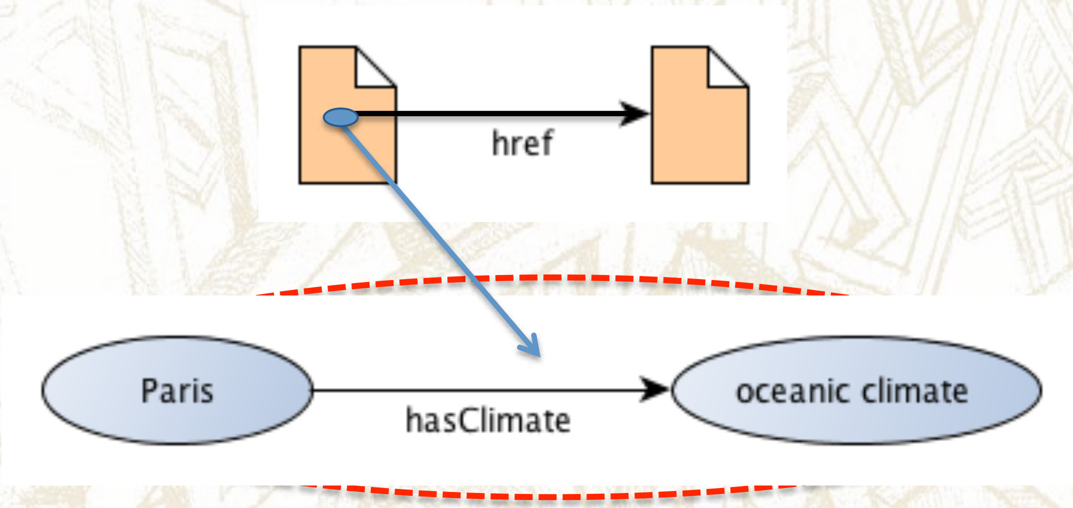
User-based evaluation

Task	Type extraction	Taxonomy induction	WSD
Correctness	0.84	0.96	0.81

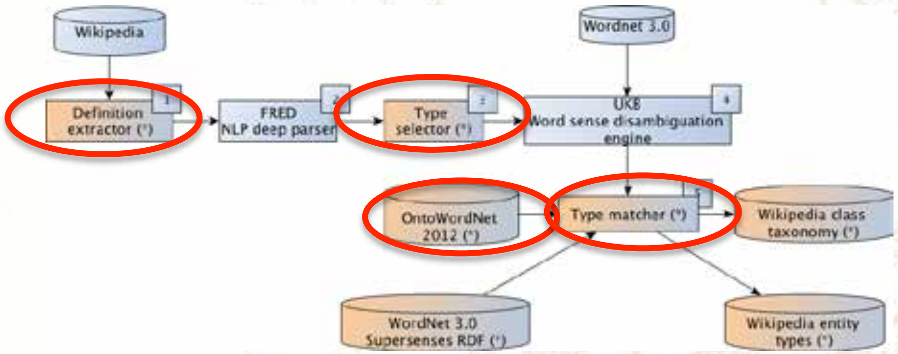
ONGOING…

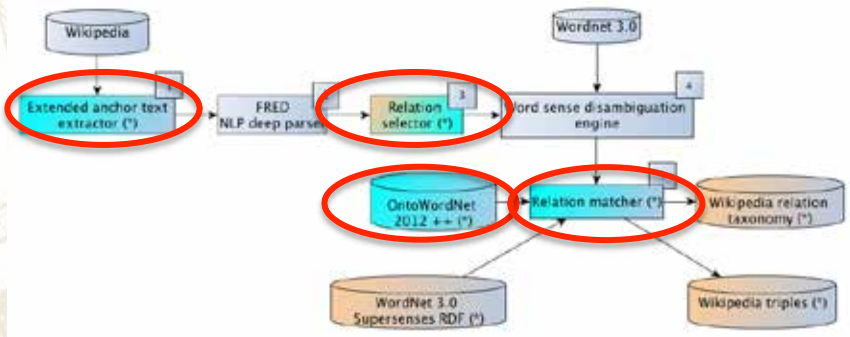
Lègalo

- To guess the meaning hidden by hypertextual links, given the text surrounding anchors (href)



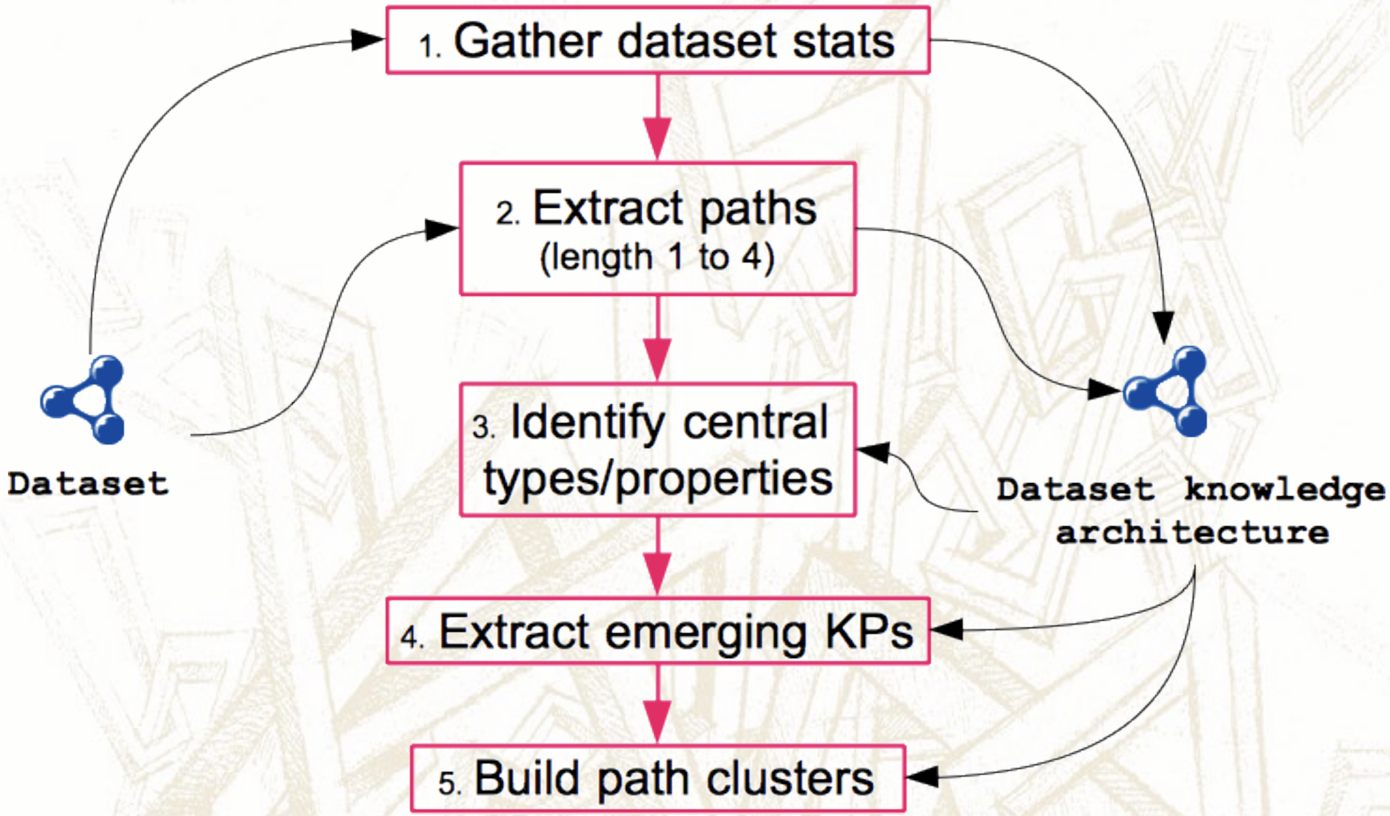
Adapting Tìpalo process for guessing semantics of links





BOTTOM-UP: EXTRACTING KP FROM LINKED DATA

Schema extraction

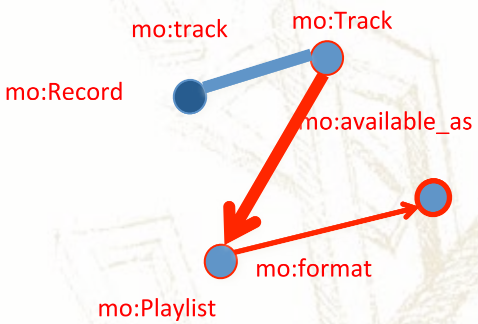


Valentina Presutti, Lora Aroyo, Alessandro Adamou, Balthasar Schopman, Aldo Gangemi, Guus Schreiber: Extracting Core Knowledge from Linked Data. COLD2011, CEUR-WS.org Vol-782.

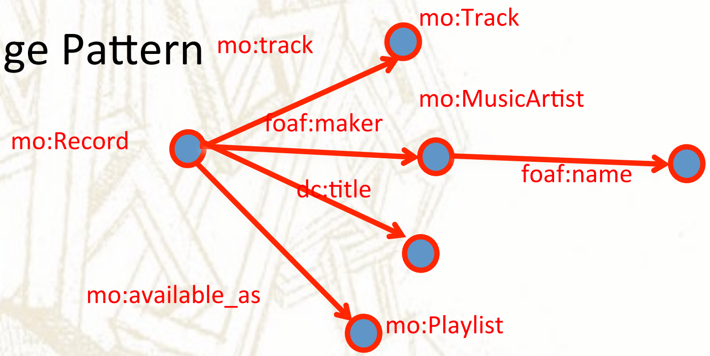
Results

- A method for extracting the main knowledge patterns of a LD dataset

Path



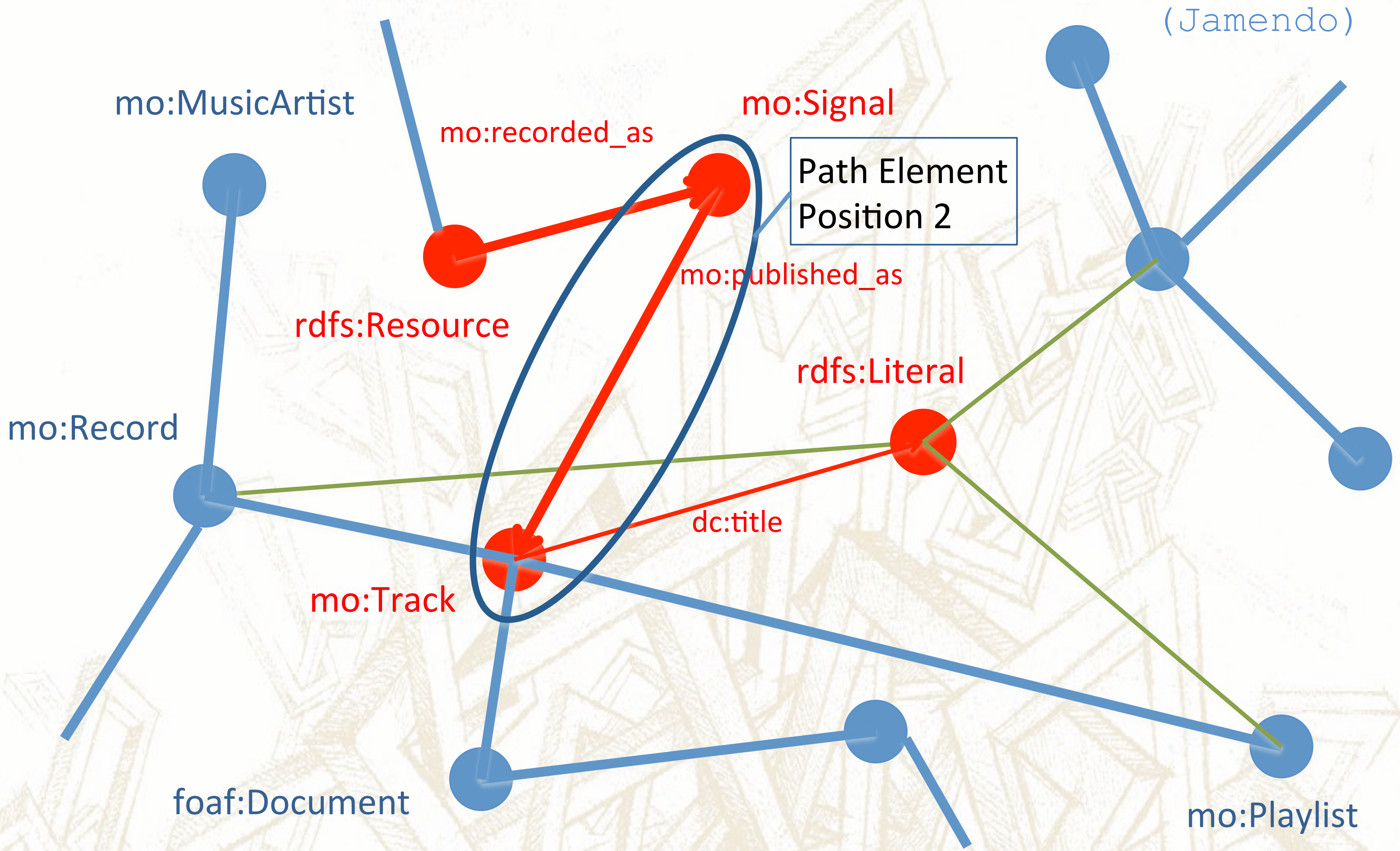
Knowledge Pattern



Cluster of paths

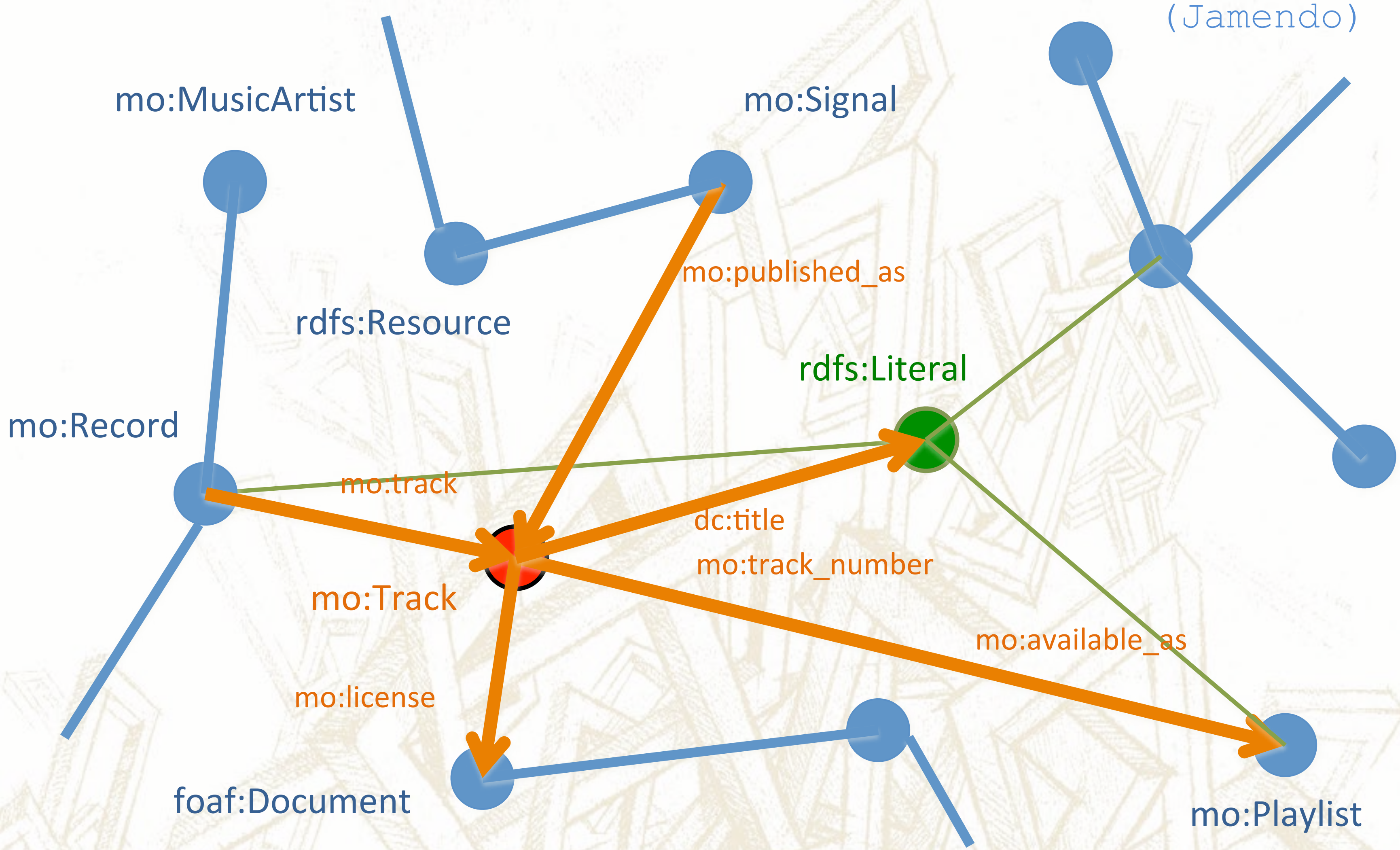
mo:Record	mo:track	mo:Track	mo:available_as	mo:Playlist	dc:format	rdfs:Literal
mo:Signal	mo:published_as	mo:Track	mo:available_as	mo:Playlist	dc:format	rdfs:Literal
mo:MusicArtist	foaf:made	mo:Record	mo:available_as	mo:Torrent	dc:format	rdfs:Literal
mo:MusicArtist	foaf:made	mo:Record	mo:available_as	mo:ED2K	dc:format	rdfs:Literal
mo:MusicArtist	foaf:made	mo:Record	mo:available_as	mo:Playlist	dc:format	rdfs:Literal

Path identification (length 3)



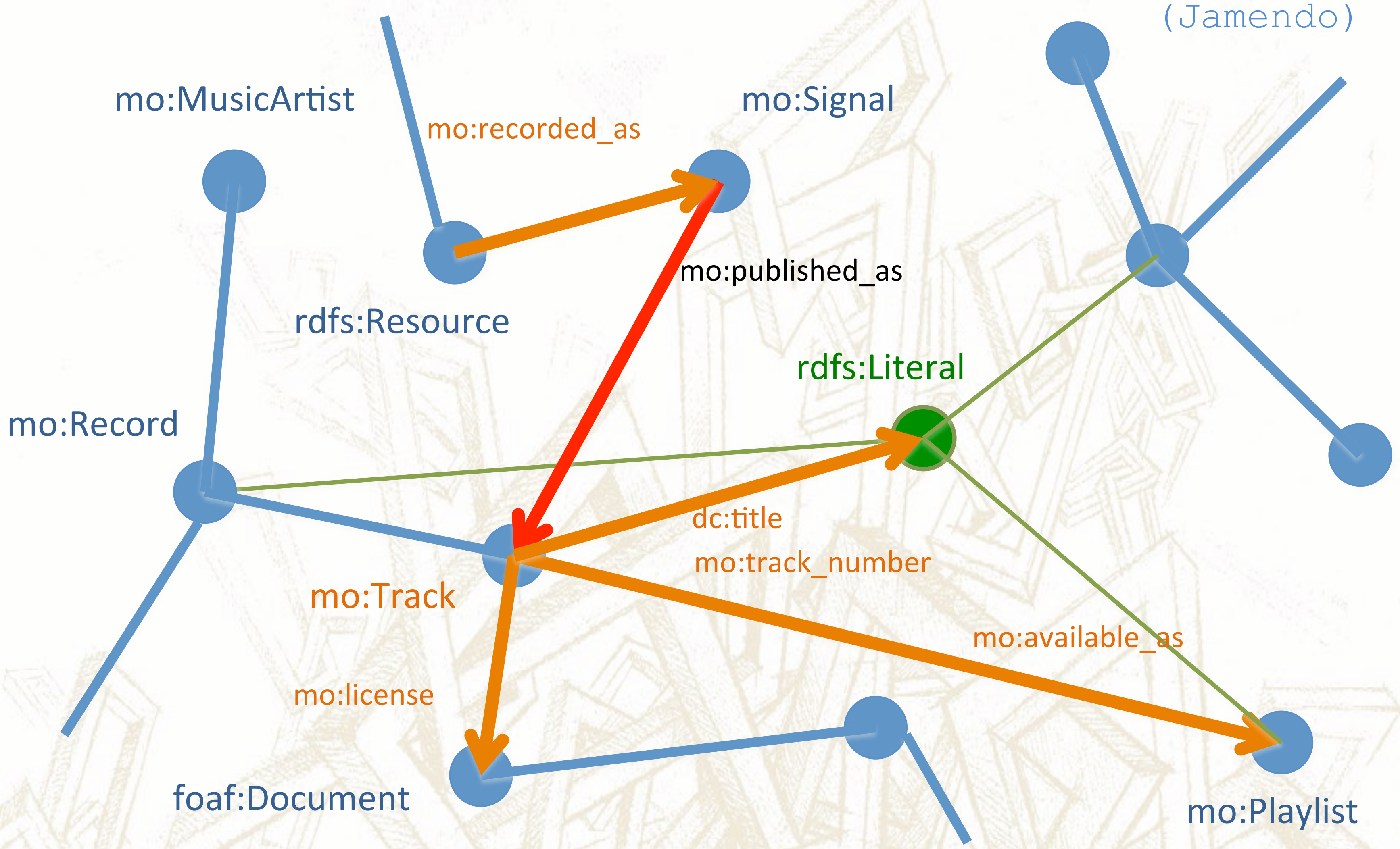
PREFIX mo : http://purl.org/ontology/mo/MusicArtist

Centrality (types)



PREFIX mo : http://purl.org/ontology/mo/MusicArtist

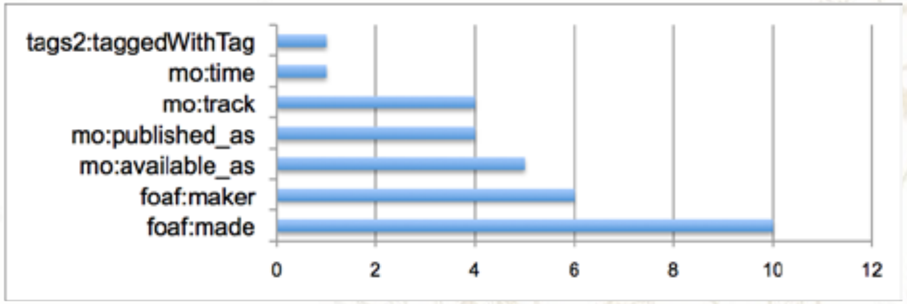
Centrality (properties)

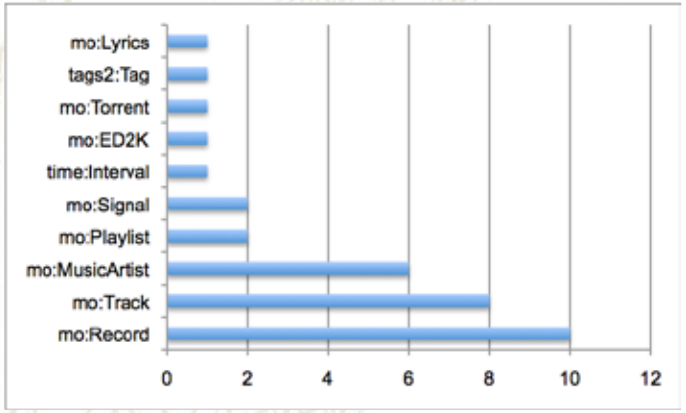


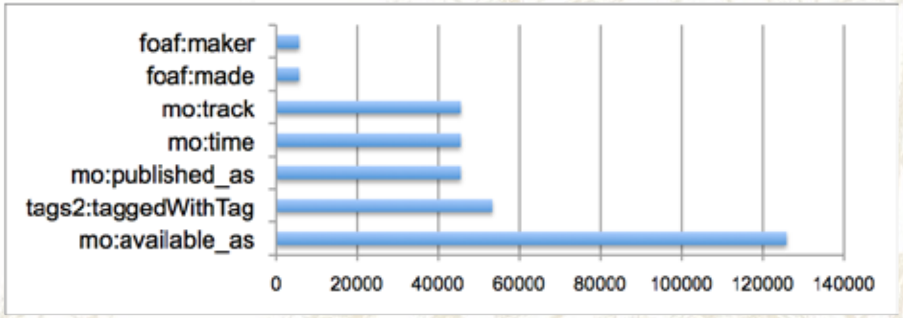
PREFIX mo : http://purl.org/ontology/mo/MusicArtist

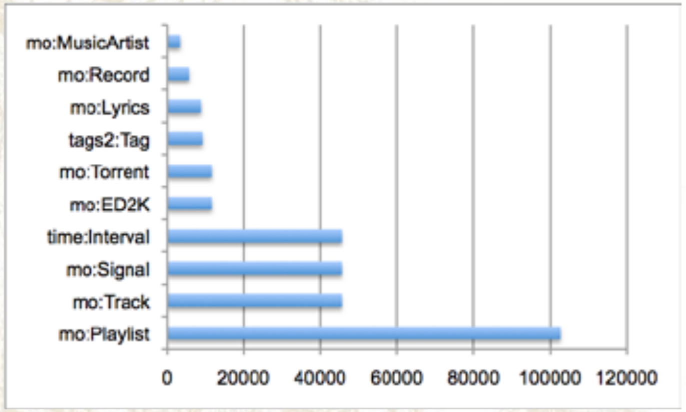
Centrality in Jamendo

betweenness



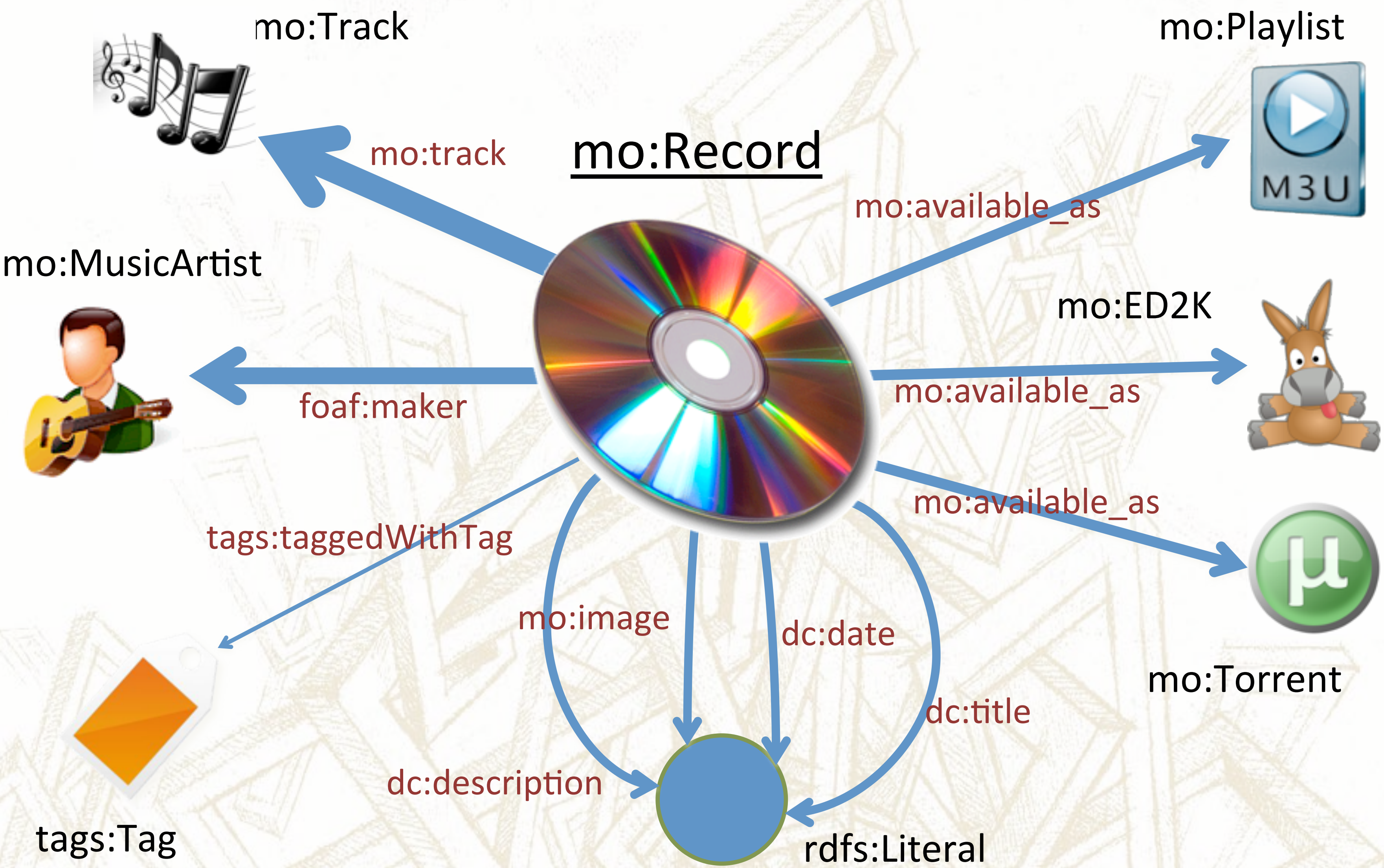






frequency

Emerging Knowledge Pattern



Bottom-up: Encyclopedic Knowledge Patterns (EKP)

- Improving knowledge exploration and summarization by:
  - Empirically discovering invariances in conceptual organization of knowledge – encyclopedic knowledge patterns – from Wikipedia crowd-sourced page links
  - Understanding the most intuitive way of selecting relevant entities used to describe a given entity
  - Identifying the typical / atypical types of things that people use for describing other things
  - Enabling serendipitous search

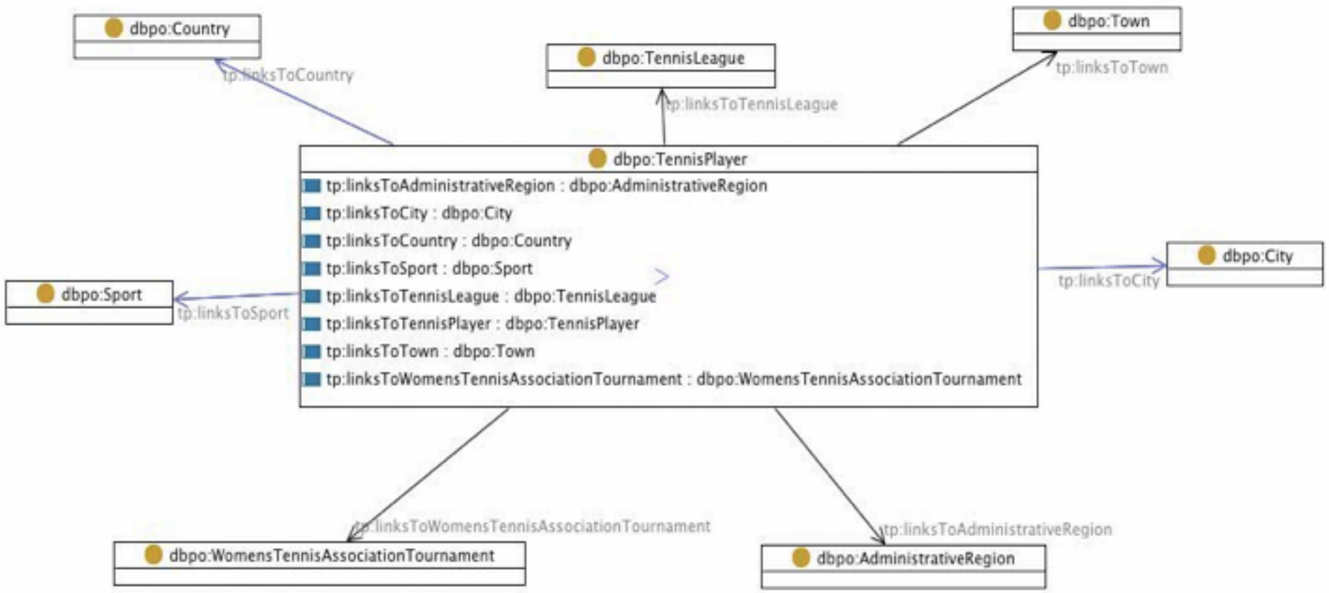
Andrea Giovanni Nuzzolese, Aldo Gangemi, Valentina Presutti, Paolo Ciancarini: Encyclopedic Knowledge Patterns from Wikipedia Links. International Semantic Web Conference (1) 2011: 520-536

Input data

- Wikipedia page links generate 107.9M triples
- Infobox-based triples are 13.6M, including data value triples (9.4M)
- “Unmapped” object value triples are only 7% of page links
- Paths are used to discover *Encyclopedic Knowledge Patterns*
  - Such patterns should make it emerge the most typical types of things that the Wikipedia crowd uses to describe a resource of a given type

Encyclopedic Knowledge Patterns

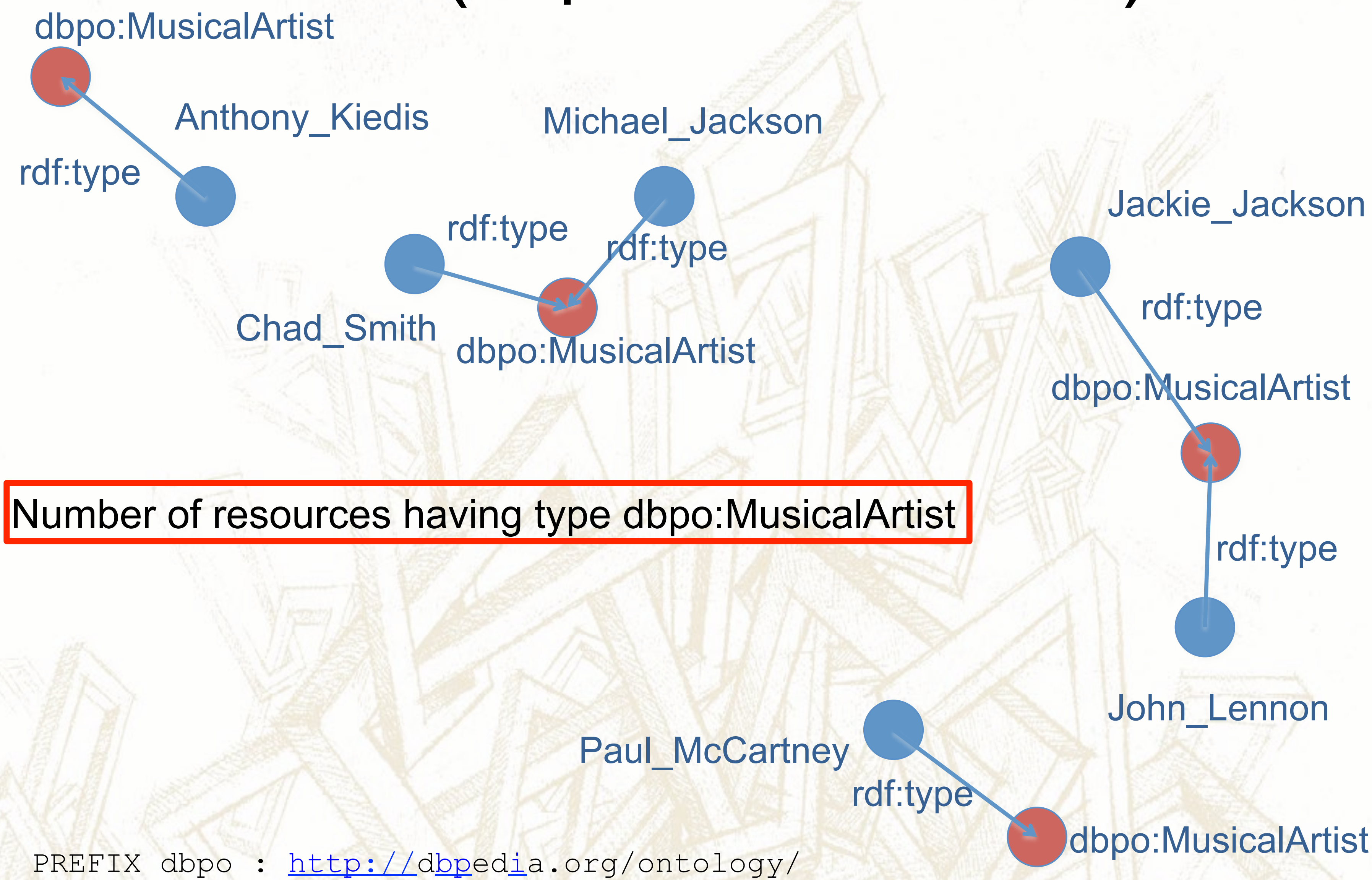
- An Encyclopedic Knowledge Pattern (EKP) is discovered from the paths emerging from Wikipedia page link invariances
- They are represented as OWL2 ontologies



Paths and indicators

- Emerging paths are stored in RDF according to the “Knowledge Architecture” vocabulary
  - Cf. our COLD2011 paper “Extracting core knowledge from linked data”
- Paths and types are associated with a set of indicators

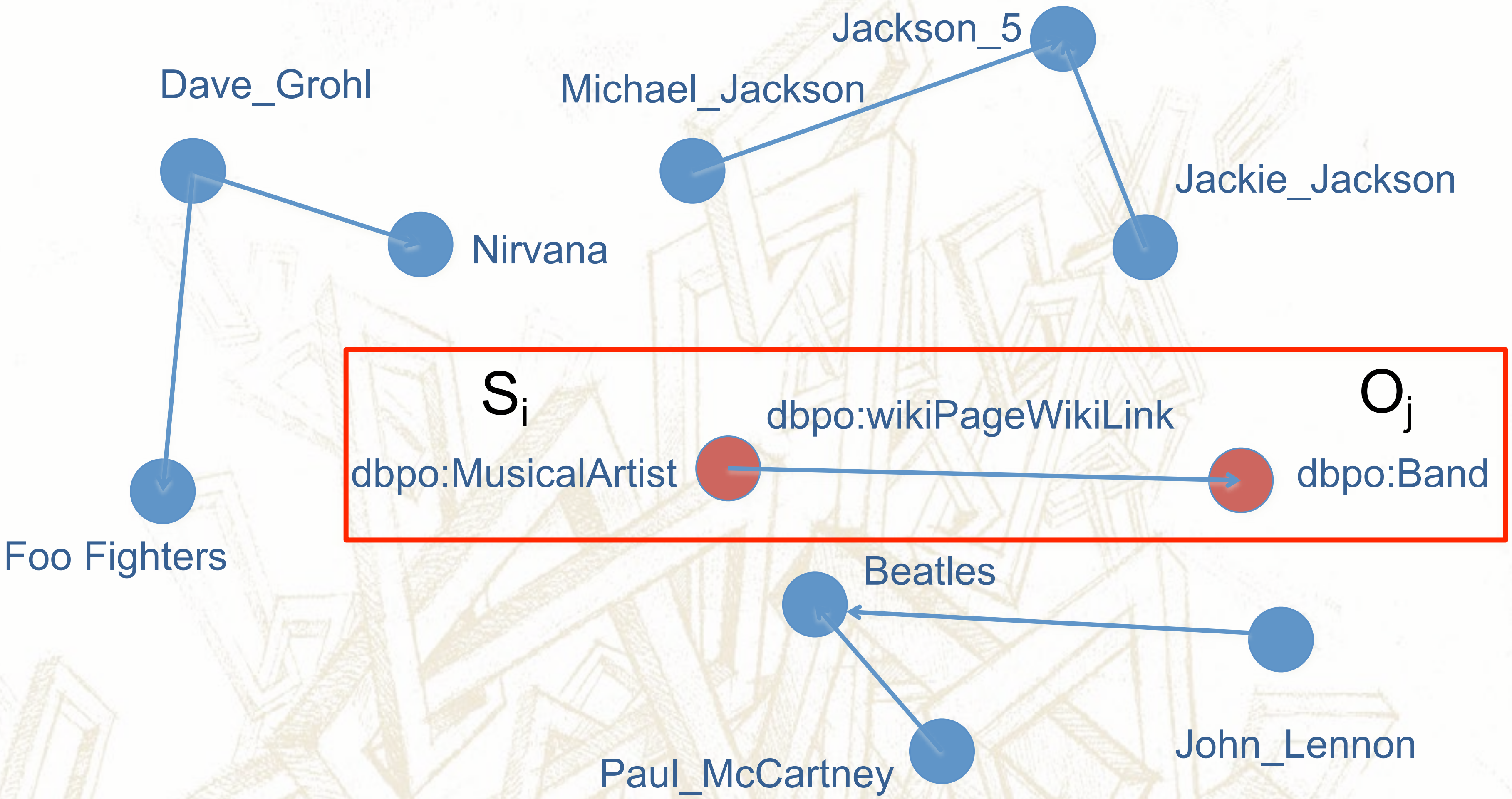
nRes(dbpo:MusicalArtist)



Number of resources having type dbpo:MusicalArtist

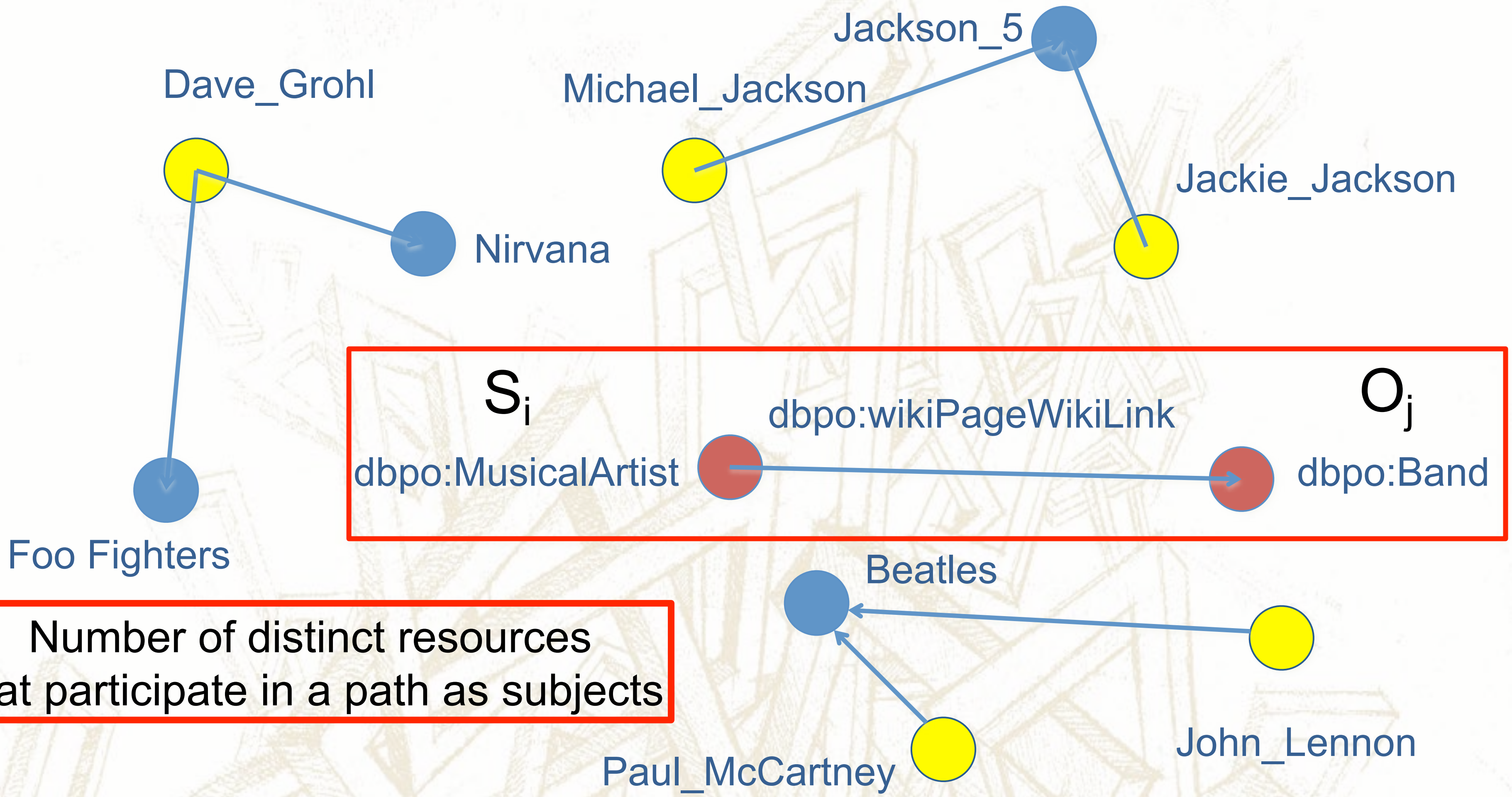
PREFIX dbpo : http://dbpedia.org/ontology/

nSubjectRes($P_{i,j}$)



PREFIX dbpo : http://dbpedia.org/ontology/

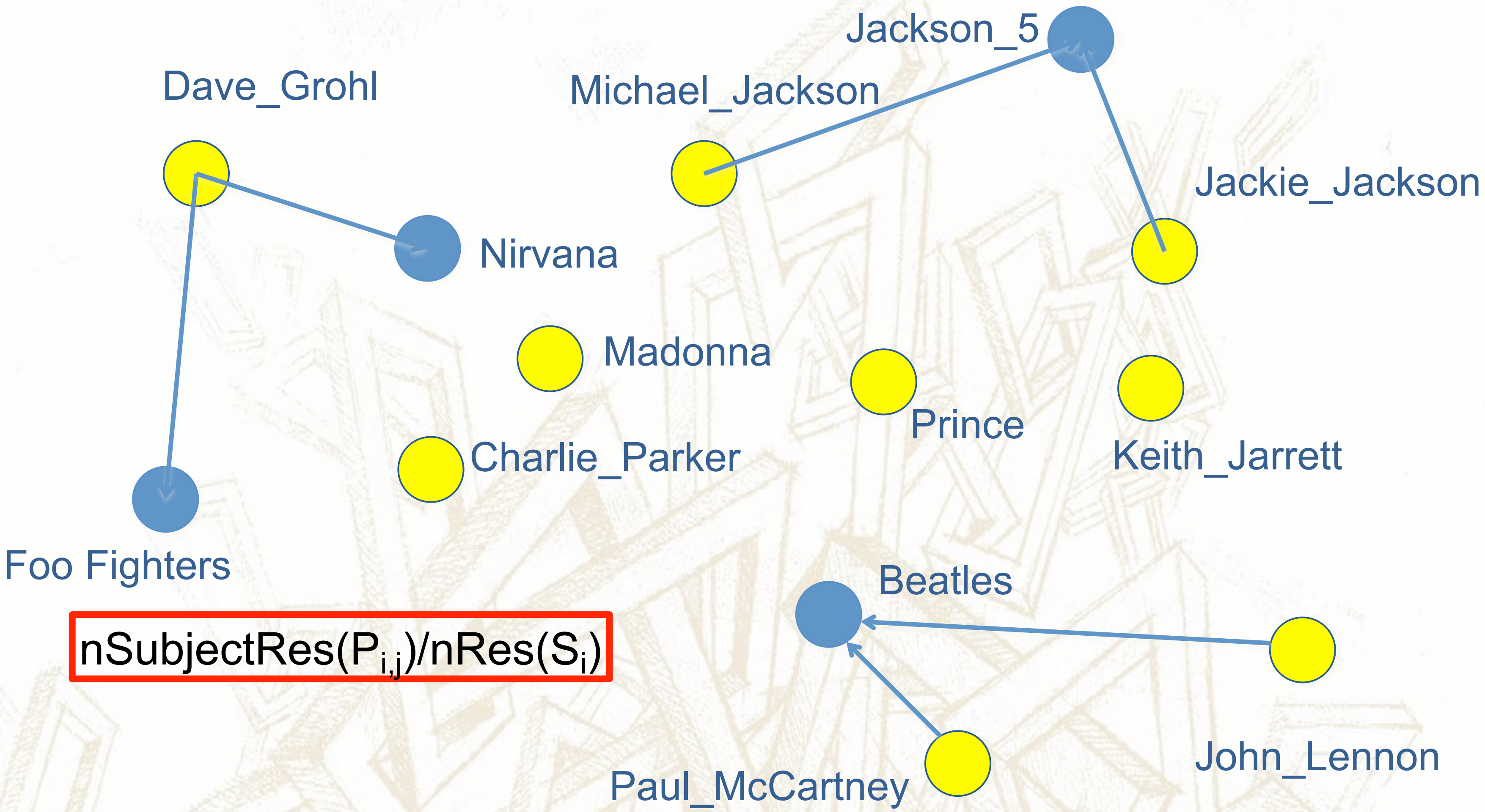
nSubjectRes($P_{i,j}$)



Number of distinct resources
that participate in a path as subjects

PREFIX dbpo : http://dbpedia.org/ontology/

pathPopularity($P_{i,j}$,S_i)



Path Popularity distribution example

Path	pathPopularity
[MusicFestival,Band]	82.33
[MusicFestival,MusicalArtist]	74.17
[MusicFestival,Country]	74.02
[MusicFestival,MusicGenre]	72.81
[MusicFestival,City]	38.37
[MusicFestival,AdministrativeRegion]	32.78
[MusicFestival,MusicFestival]	23.26
[MusicFestival,Album]	18.13
[MusicFestival,Film]	12.39
[MusicFestival,Stadium]	9.52
[MusicFestival,RadioStation]	9.52
[MusicFestival,Single]	8.76
[MusicFestival,Town]	8.61
[MusicFestival,Magazine]	8.46
[MusicFestival,Broadcast]	7.55
[MusicFestival,Newspaper]	6.95
[MusicFestival,TelevisionShow]	6.04
[MusicFestival,University]	5.74
[MusicFestival,Continent]	5.59
[MusicFestival,Comedian]	5.29
[MusicFestival,OfficeHolder]	4.98
[MusicFestival,Island]	4.98

Boundaries of EKPs

- An $EKP(S_i)$ is a set of paths, such that

- $P_{i,j} \in EKP(S_i) \Leftrightarrow pathPopularity(P_{i,j}, S_i) \geq t$

- t is a threshold, under which a path is not included in an EKP
- How to get a good value for t?

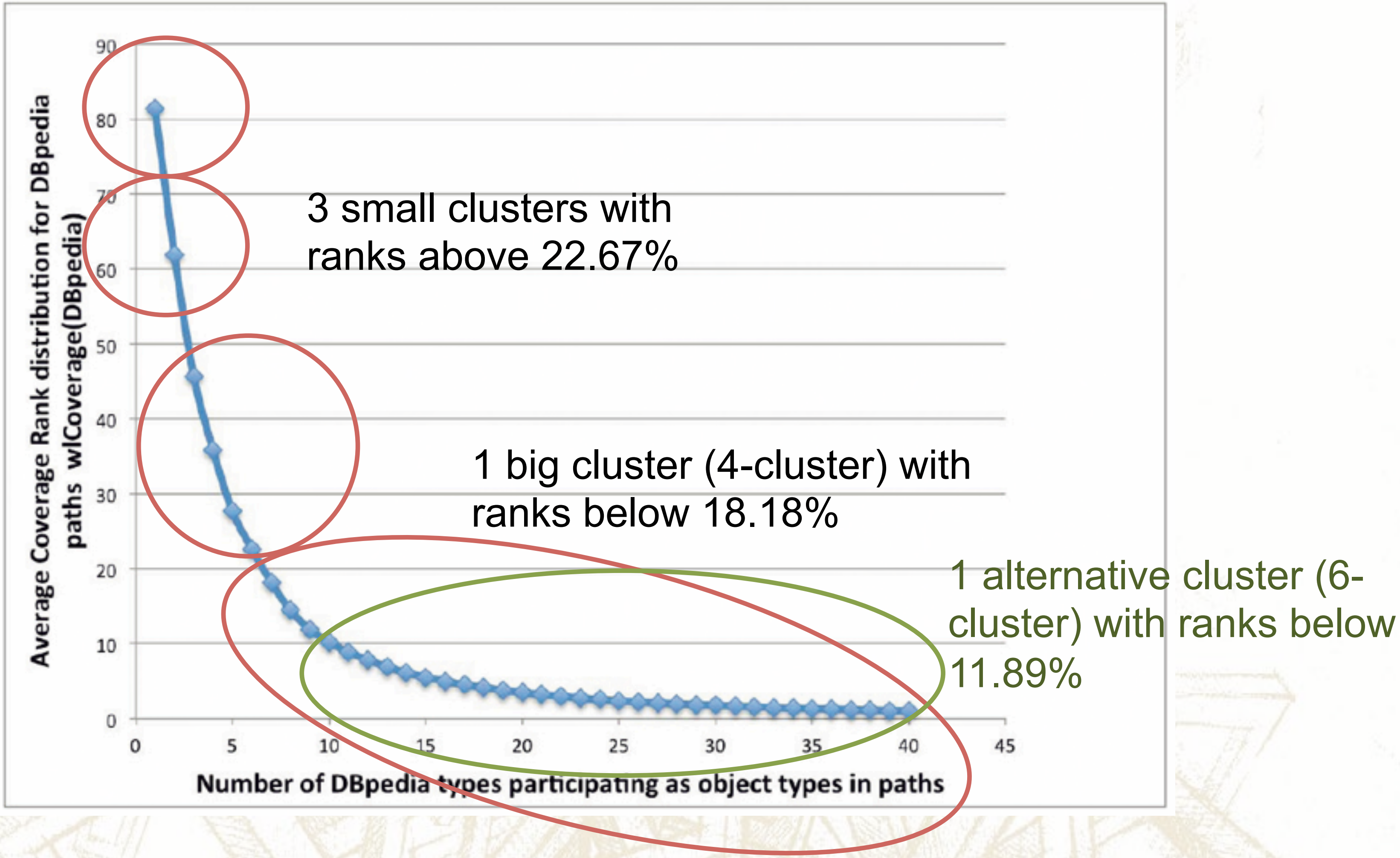
Boundary induction

Step	Description
1.	For each path, calculate the path popularity ✓
2.	For each subject type, get the 40 top-ranked path popularity values* ✓
3.	Apply multiple correlation (Pearson ρ) between the paths of all subject types by rank, and check for homogeneity of ranks across subject types ✓ .9
4.	For each of the 40 path popularity ranks, calculate its mean across all subject types ✓
5.	Apply k-means clustering on the 40 ranks ✓
6.	Decide threshold(s) based on k-means as well as other indicators (e.g. FrameNet roles distribution) ✓

* 40 covers most “core” path popularity values, as well as many of the unusual ones.

k-means clustering on Path Popularity

Sample distribution of pathPopularity for DBpedia paths. The x-axis indicates how many paths (on average) are above a certain value t for pathPopularity



What is the "agreement" between DBpedia and our sample users?

Average multiple correlation (Spearman ρ) between users' assigned scores, and $pathPopularity_{DBpedia}$ based scores

Spearman ρ range: [-1,1]

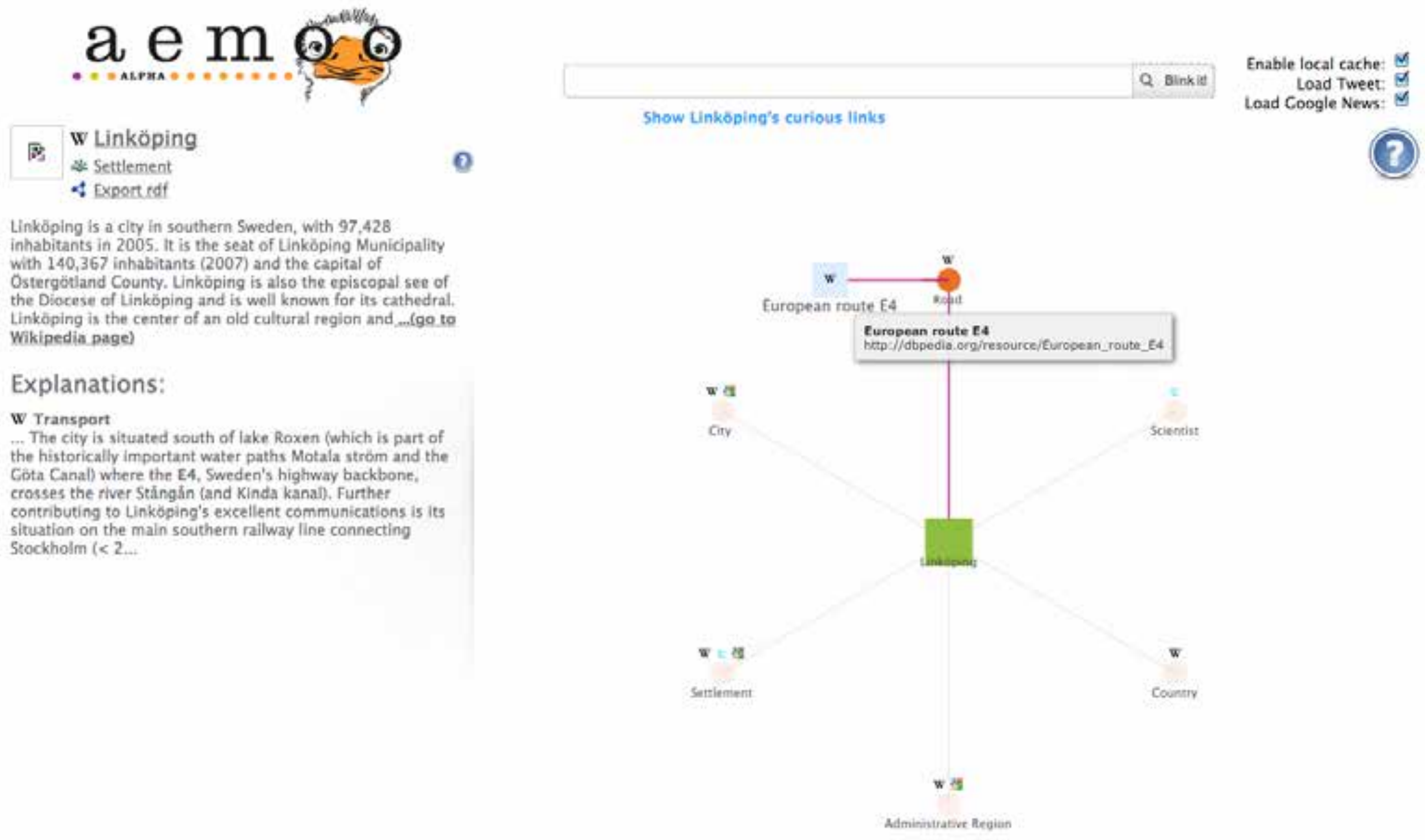
$$P_{i,j} \in EKP(S_i) \Leftrightarrow pathPopularity(P_{i,j}, S_i) \geq 11\%$$

Satisfactory precision

	users / DBpedia		users / DBpedia
Language	0.893	Philosopher	0.661
Writer	0.748	Ambassador	0.655
Legislature	0.716	Album	0.871
Radio Station	0.772	Administrative Region	0.874
Country	0.665	Insect	0.624
Disease	0.824	Aircraft	0.664

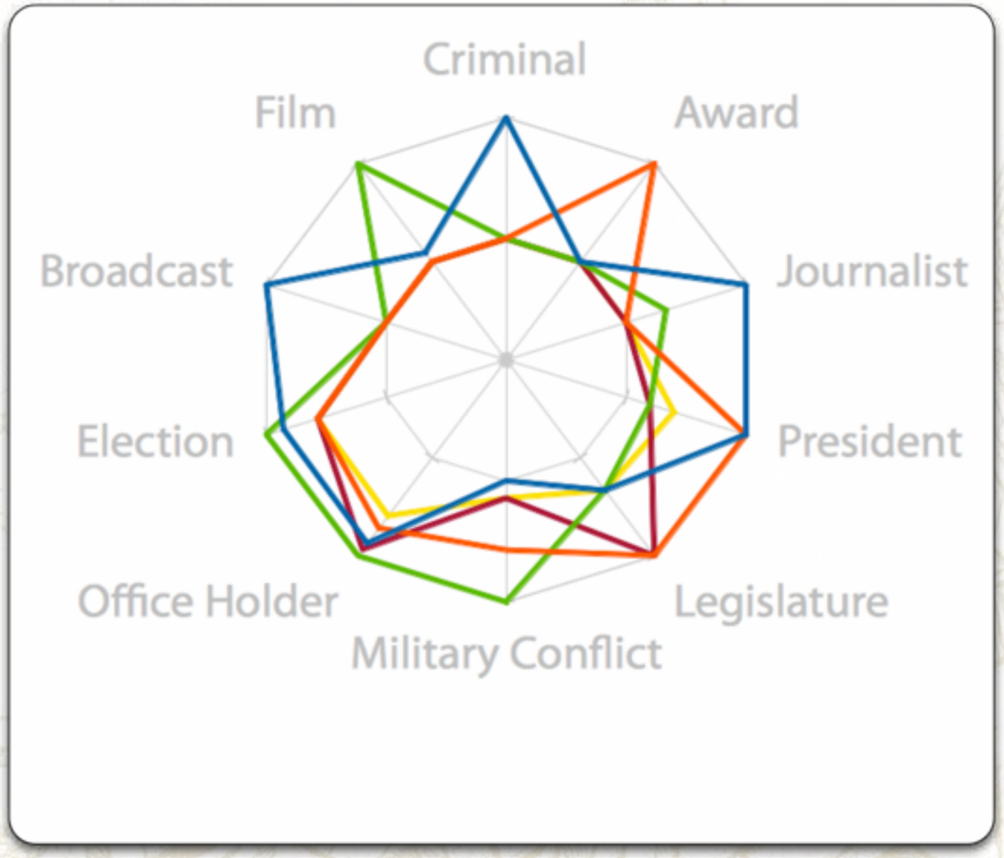
Aemoo

- http://aemoo.org exploratory search application based on EKP



Semantic Web Challenge @ISWC 2011 – Short listed, 4th place

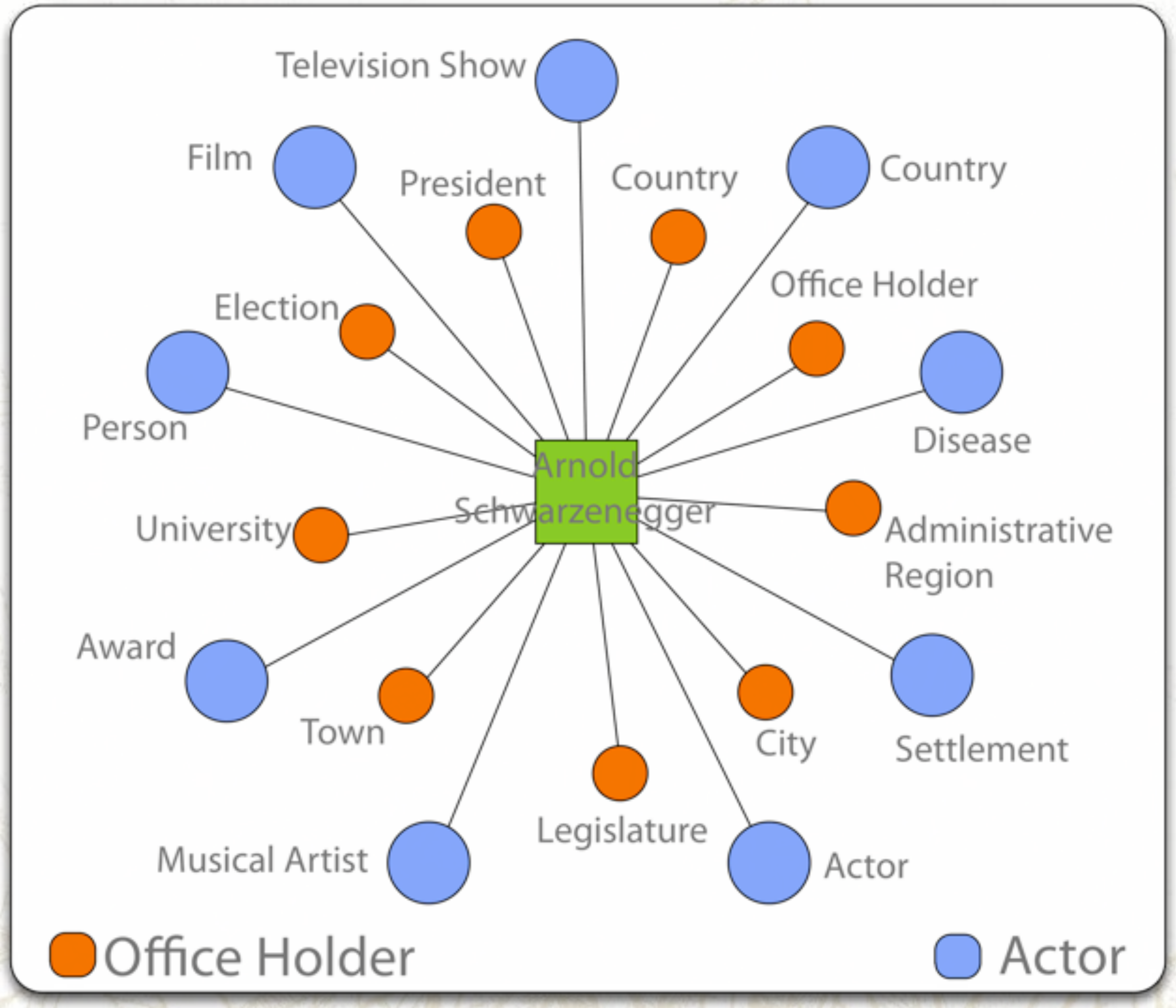
Comparing entities of the same type



Nicolas Sarkozy Ronald Regan Angela Merkel

Barack Obama Silvio Berlusconi

Applying topic-sensitive EKP as lenses



Conclusions

- Cognitive science is quite explicit on what meaning is for humans: an activity of "framing reality for a purpose" (with broad sense of reality and purpose)
- Frames can be keys to that relational meaning
- The Semantic Web is now growing a lot of data: our tools are trying to make sense of them by using appropriate keys
- Empirical extraction and discovery of frames/ knowledge patterns is feasible
- KP can be used for improving HCI e.g. in exploratory search