Reutlingen University

DBKDA 2018, Nice, 24.05.2018

DBKDA/WEB/GraphSM
Panel discussion on

Getting More from Linked Data

***Moderator**: Fritz Laux, Reutlingen University, Germany*

***Panelists**:*

Jean-Marie Le Goff, CERN- European Organisation for Nuclear Research, Switzerland

Malcolm Crowe, University of the West of Scotland, Scotland

Elio Toppano, Dept. Math & CS, Università di Udine, Italy

Reutlingen University

Scope of Topic

- *„Getting more …“ → of what?*
  - ☞ More semantics, pragmatics, understanding?
  - ☞ More comprehensive, unambiguous, accurate data?

- *„… from Linked Data“ → vs other sources?*
  - ☞ Open access data linked with copyrighted data?
  - ☞ Ontologies linked with XML/RDF data?
  - ☞ Linked with Databases (schema)?
  - ☞ Linked with Web data (schemaless)?

- *Alternatives*
  - ☞ Compared to alternatives (virtual data integration)?

Reutlingen University

Initial Statements/Topics of Panelists

- *Jean-Marie Le Goff: „Linked Data is best in the form of labelled property graphs”*
- *Malcolm Crowe: „I am a skeptic on Data Mining for lots of reason including ethical ones. It is always best to *work with* providers of data“*
- *Elio Toppano: „Exploiting genre meta-models (i.e., ontologies) for extracting and linking data from multimodal artifacts“*
- *Fritz Laux: „We need to introduce quality control to benefit more from Linked Data”*

Reutlingen University

The Basic Concepts (T. Berners Lee)

- *Design rules for Linked Data*[*)]*
  1. Use URIs as names for things
  2. Use HTTP URIs so that people can look up those names.
  3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
  4. Include links to other URIs. so that they can discover more things.

- *Essentially the rules result in a instance based labelled property graph*

- *This allows machines (and humans) to explore data*
  - ☞ Does it help with understanding the data?
  - ☞ Does it improve the quality of data?
  - ☞ How about privacy?

*) *http://www.w3.org/DesignIssues/LinkedData.html*

Reutlingen University

Problem of Data Quality

- *Instance based links are prone to low data quality*

- *Ambiguity problem*
  - No prevention from having multiple URIs for one entity
  - No schema allows non uniform representations and structure → ambiguous description and possible duplicates

- *Inconsistent data*
  - Without schema/ontology data can be inconsistent and even contradicting
  - Redundant data may refer to different ontologies with different semantics

- *Duplicate data*
  - Duplicates are hard to detect without schema/ontology
  - Disambiguation is practically impossible

- *Schema support can help to improve data quality*

Reutlingen University

Problem of Semantics, Pragmatics and Data Privacy

- *The Linked Data concept does not support any privacy mechanism*
  - Sensitive Data should be hidden from disclosing it
  - Removing data is not a good option, because traversal should be still possible
  - Adding protocol support for traversal of disclosed data necessary
- *Linked Data need support for Semantics and Pragmatics through*
  - Unambiguous URIs (identification)
  - Schema definition for semantics
  - Ontology for semantics and pragmatics
  - Enhanced contextual information for pragmatics
- *(Meta)Ontologies/(Meta)Schemata can help to harmonize semantics of different sources*

Linked data is great

- E.g. links scattered through good journalism
  - So we can see what report is being referred to
- Promotes transparency, we can verify data sources
  - Every Excel spreadsheet should justify each value with links
- Links are always legal, though following them may cost
  - All credentials should be linked data
- Every database row should declare provenance
  - Every set of data returned by HTTP should have an Etag
    - So you can check it is still valid

What are computers/databases for?

- Providing reliable calculations/data/information
- The starting data quality is important
- We should be confident in any transformations
- We want the right answer as soon as possible
  - But not sooner, as it will probably be wrong
- Always prioritise correctness, choose data with care
- Cheating will be detected sooner or later
  - Though you can try to blame someone else

DBKDA/WEB/GraphSM
Panel discussion on

Getting More from Linked Data

Exploiting genre meta-models (i.e., ontologies) for extracting and linking data from multimodal artifacts

Panelist:
Elio Toppano, Università di Udine, Italia

Getting More … → of what ?

- My answer:
  - more *contextual knowledge* for interpreting data
  - more *deep knowledge* (i.e., kowledge grounded on expertise, foundamental theories and models)
  - more support for *reflection* viewed as serious thought or consideration aimed at unfolding meanings, values, points of view, assumptions in order to *understand* something better

- **How to do it?**

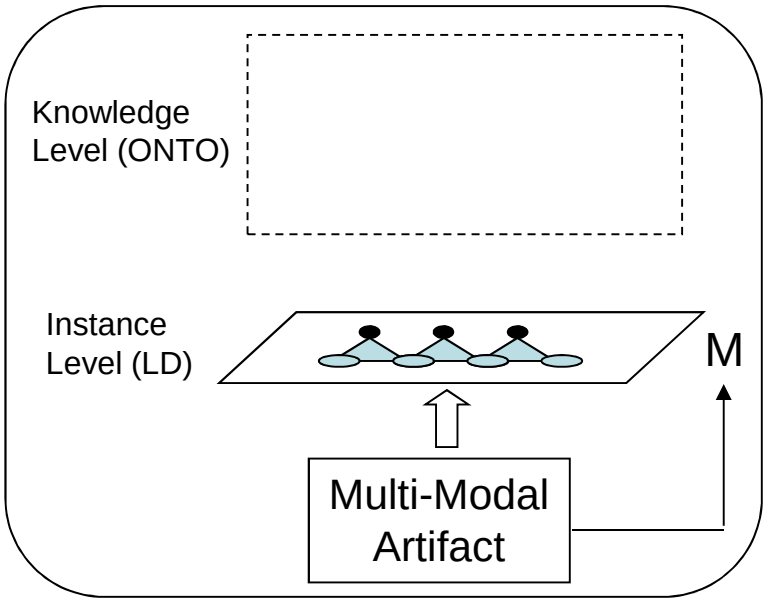
- I will focus on LD generated from description and annotation of Multimodal Artifacts

- We can exploit **ontologies** or **meta models** of specific types of artifacts (so called genre meta-models). A genre meta-model is a conceptualization of a kind of artifact aimed at representing prototypical features and characteristics such as its structure, behaviour, function, teleology and intended use

- I envisage several possibilities of using genre metamodels each one corresponding to a specific level of reflection (and understanding)

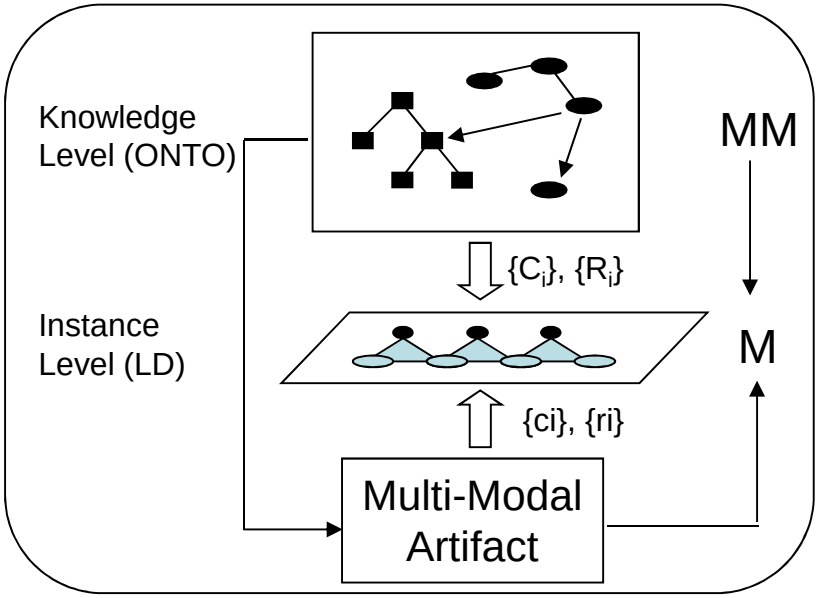
L0: no meta-models

- The model M (a set of statements) represents a collection of (generally) unstructured facts about the MMA without further elaboration or explanation. Recording and revisiting facts. Not reflective.



L1: single meta-model

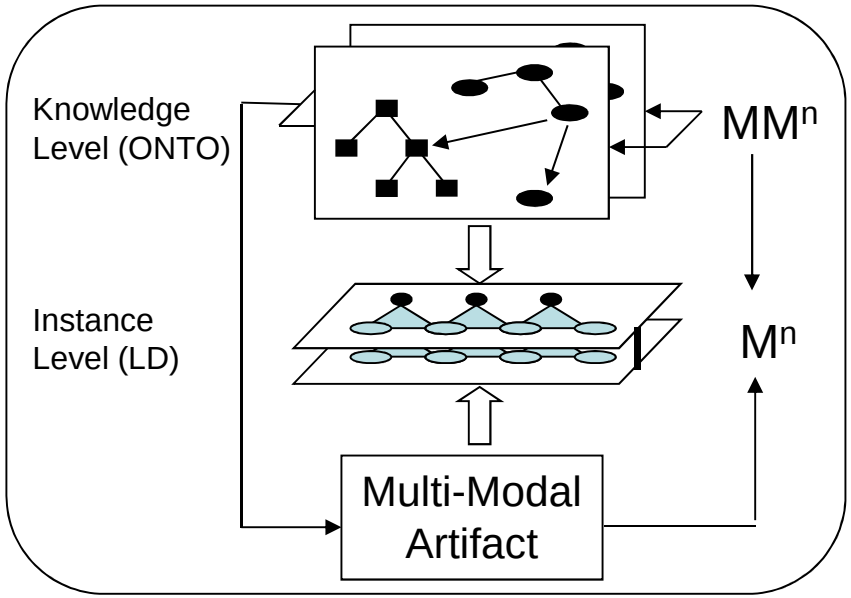
- A genre meta-model MM is used to support the analysis of the MMA. The meta-model provides the (types of) concepts and relations that are used in the description M; specific values (instances) for these elements are provided by the artifact. The model unfolds design knowledge embodied in the artifact. Moderate reflection (e.g., classification, generalization/specialization) and inference. No change of perspective or alternative explanations.



Note: concepts of different epistemological type are mixed together!

L2: single meta-model, multiple perspectives

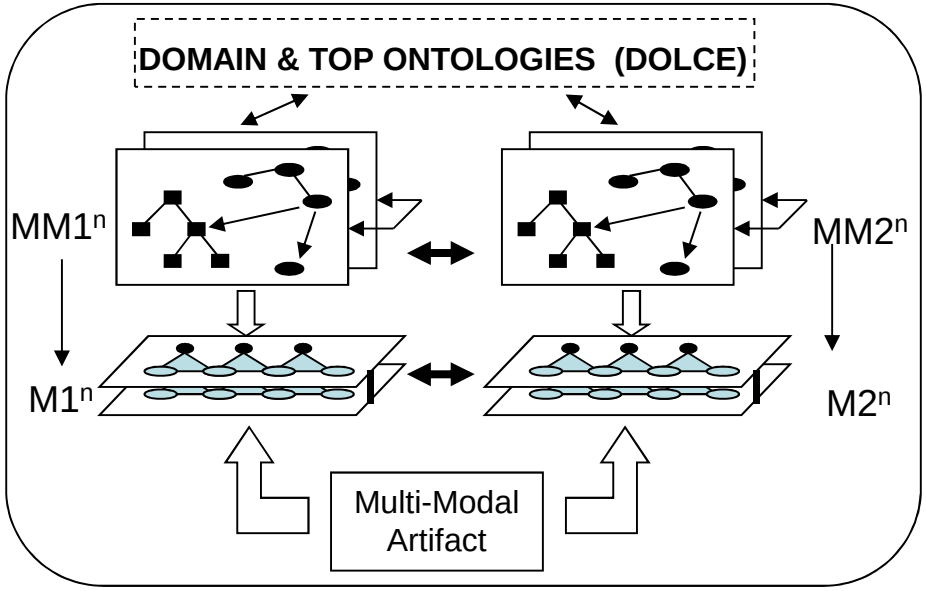
- The metamodel represents the artifact from different (interrelated) homogeneous perspectives. Each perspective focuses on a specific type of knowledge about the artifact (e.g., structural knowledge, aesthetics, narrative content, rhetorical structure, deep values). There is the possibility of means-end explanation, reasoning within a conceptualization and through conceptualizations. Dialogic reflection (seeing things from different PoVs)

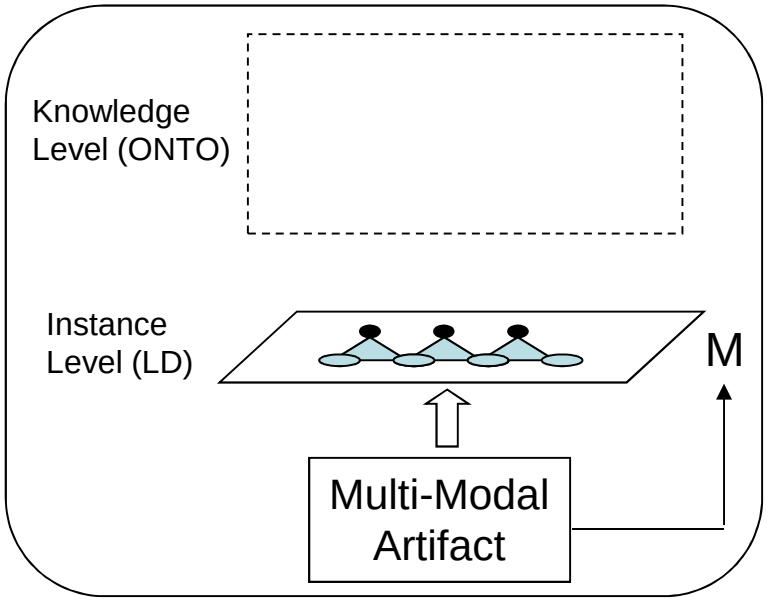


Looking for relationships between pieces of knowledge (related facts) from different perspectives

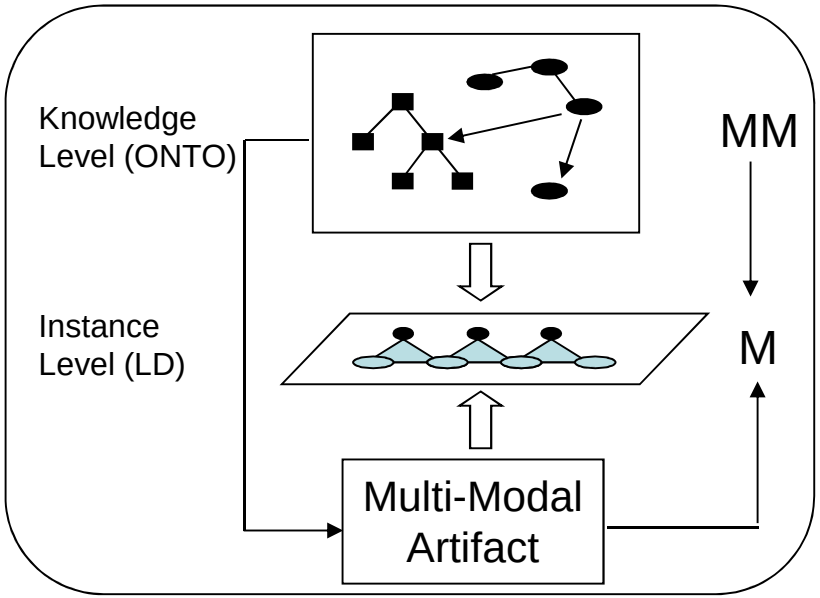
L3: multiple meta-models and perspectives

- Use of alternative meta-models of the MMA. Each MM leads to a reorganization of the knowledge about the artifact. There is the possibility to reflect about the different assumptions, intents, values and social-cultural effects … laying behind each metamodel. Transformative and critical reflection (the original MM is somehow altered in a foundamental way)

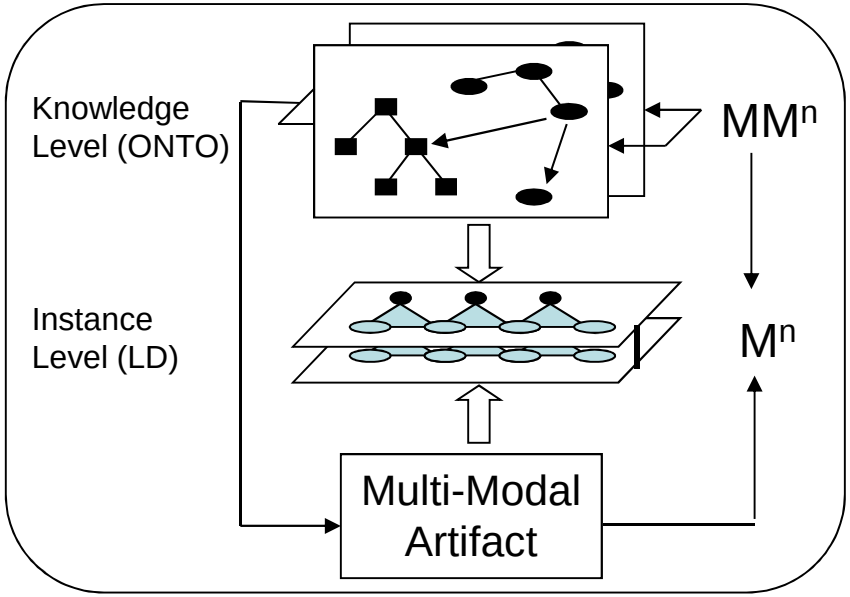




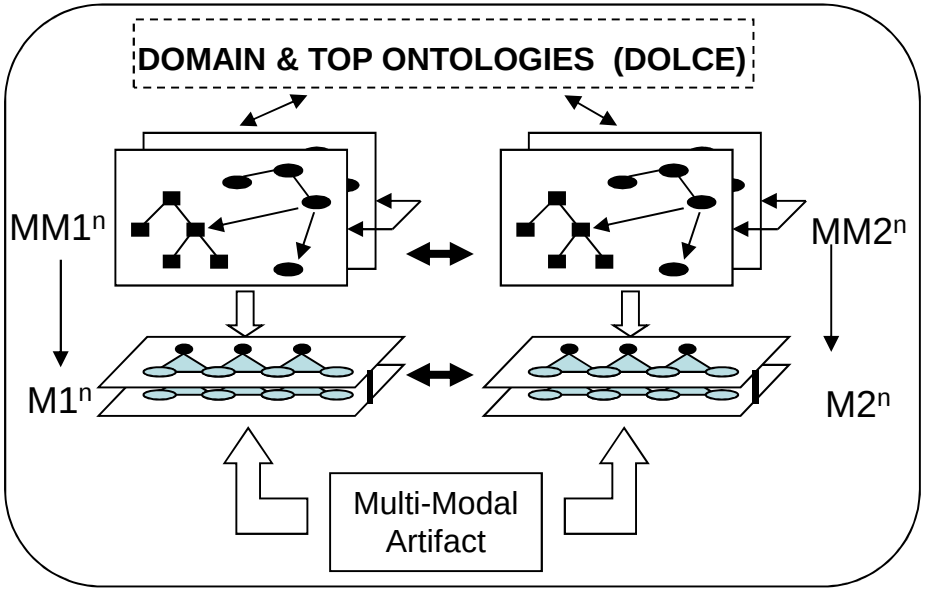
L0: description of facts (no reflection)



L1: … + classification-inference



L2: …. + seeing things from multiple interrelated perspectives



L3: ….+ reorganizing knowledge using alternative MM

Summary

- Linked data provide the “material” that can be used as a basis for reflection.
- Meta models provide structure for data but more importantly for reflection and understanding. Linked data generated by projecting a meta-model on an artifact raise awareness of the fact that there is something to reflect about (a focal point); they have higher quality (e.g., more unity, connectedness and coherence).
- Multiple perspectives enable seening things according to different conceptualizations (e.g., using different epistemological types). This add further structure to linked data and allows shifting from one perspective to another one in a controlled way
- Multiple meta-models support more complex types of reflective thinking such as comparing very different (even contradictory) meta-models, in order to disclose, motivations, intentions, modeling assumptions and deep values implicitly inscribed within the artifact.

Conclusion

- My claim is that we should promote a greater awareness of the mediation effects (e.g., persuasive intents) of our artifacts and technologies even at the cost of greater complexity.

- Therefore, … what we need are technologies for annotation that support opacity (in the sense of a greater visibility) of contextual knowledge, multi-modeling and critical reflection and understanding. I think this approach is more reliable than Data Mining, more accountable, sustainable, and .. an opportunity for new professions and practices (e.g., expert commentators and annotators of multimedia resources)

Thanks for the attention!

Getting more from Linked Data

J.-M. Le Goff

CERN

DBKDA/WEB/GraphSM, Nice, 23 May 2018

Data is distributed and connected

Data is distributed

- **Document** systems with metadata
- Database **tables** with metadata in schema

Connectivity brings value to data

- Connectivity **not materialised** due to the distributed nature of data sources
- Domain expertise brings connectivity

Building Data networks

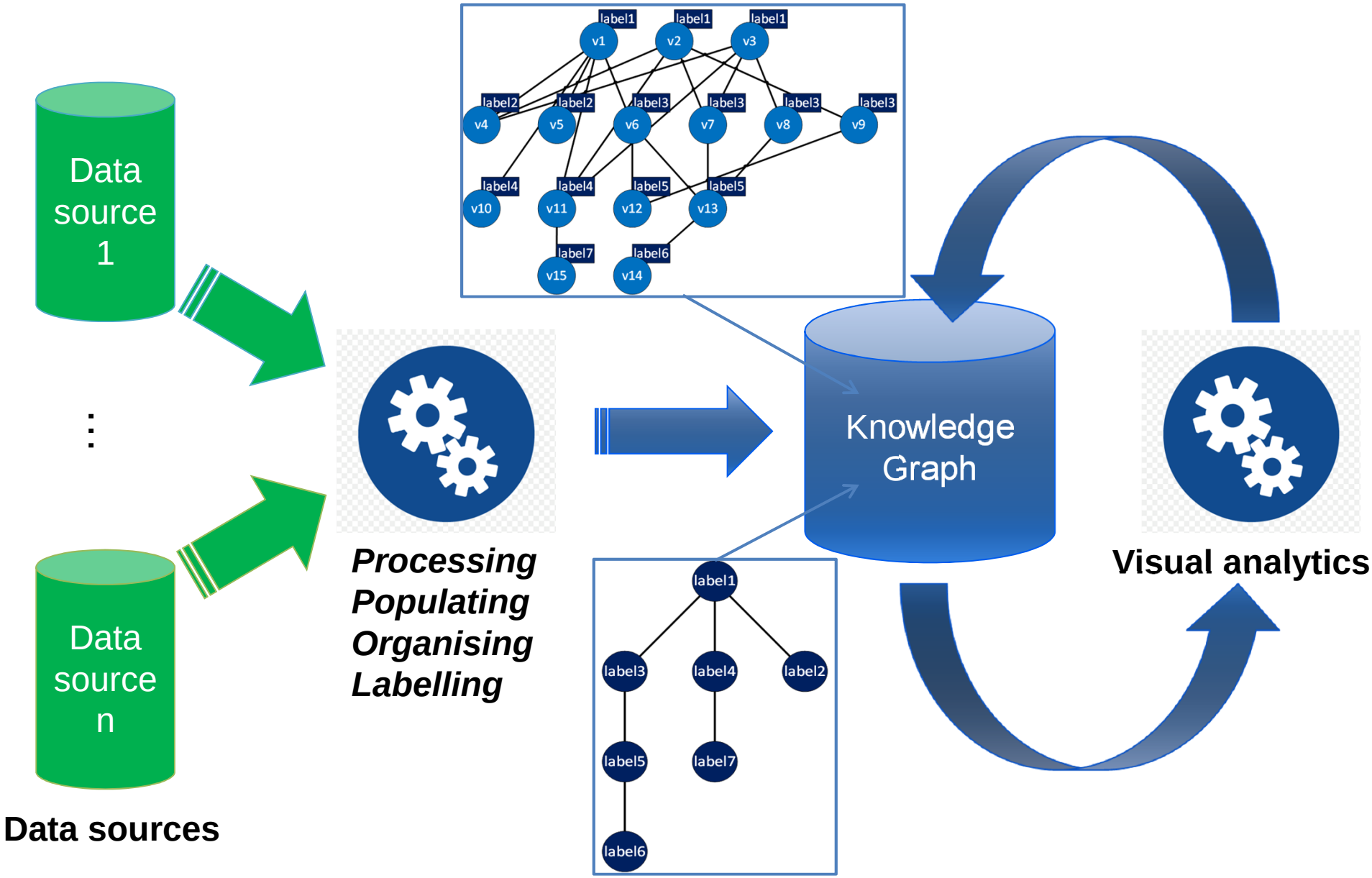
Many to many connections

- Enrich ER model
- Enrich semantics with labelled connections

Scalable and flexible

- Support any information from data sources
- Support any information in data structures
- Support any information on connections

A Domain independent platform

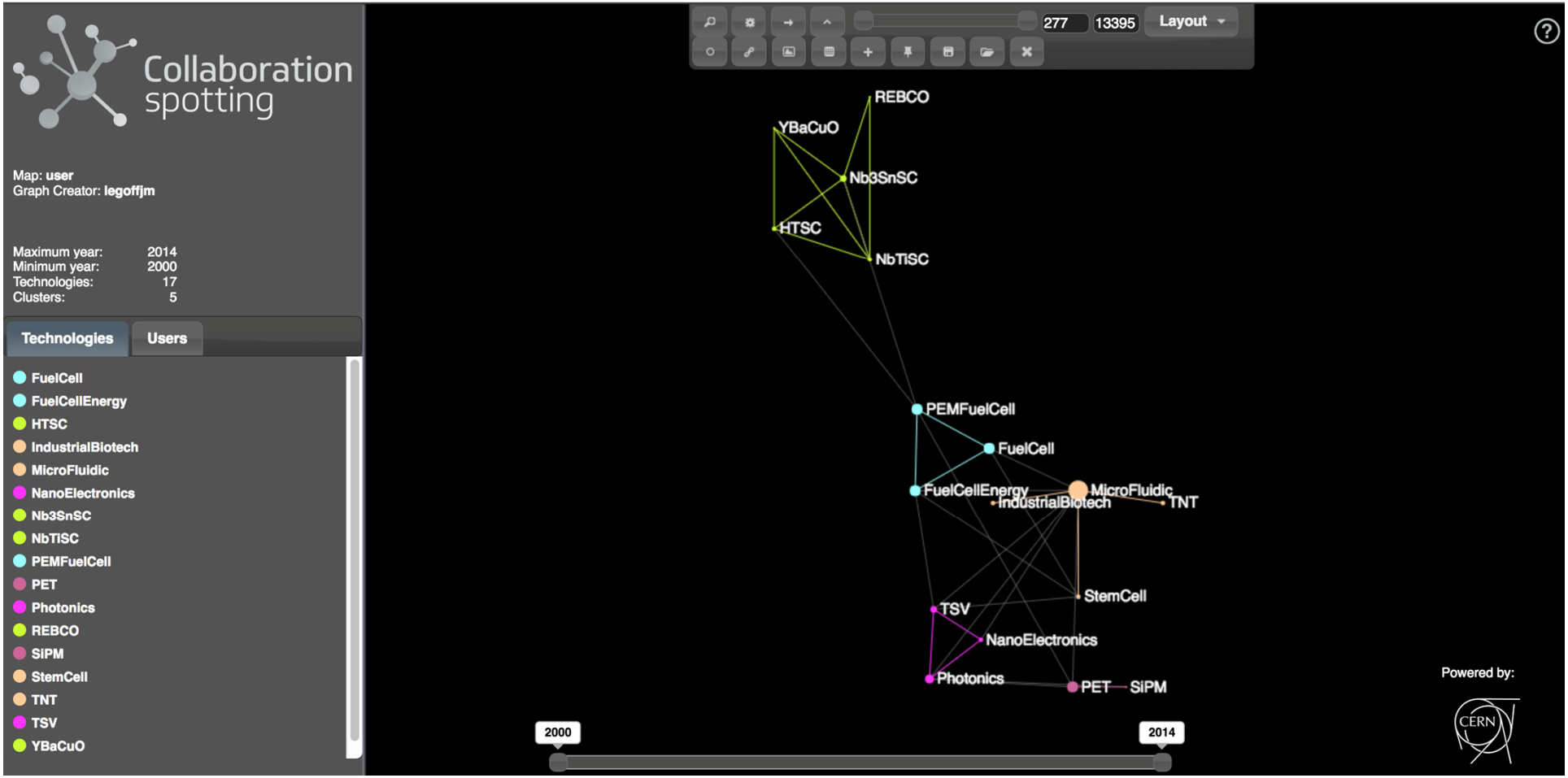


Visual analytics is performed on the Knowledge graph using Its data model

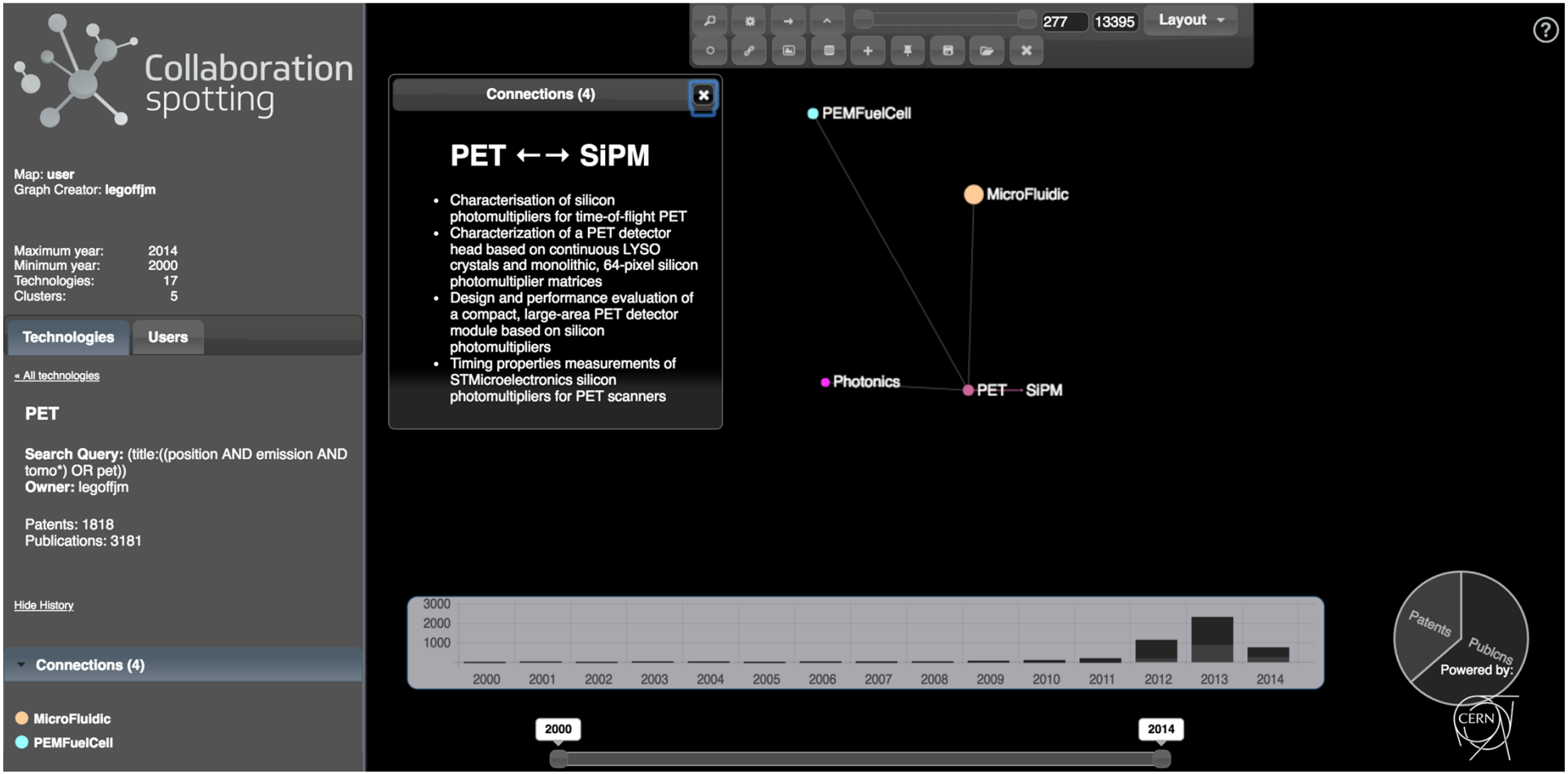
Getting more from linked data

- **Beneficiaries**
  - Analytics
  - Visual Analytics
- **Contributors**
  - Domain experts → Links via common data types
  - Machine learning → Associations, concepts
- Preserving Information processed out of data
  ➔ Enriching and Linking data.

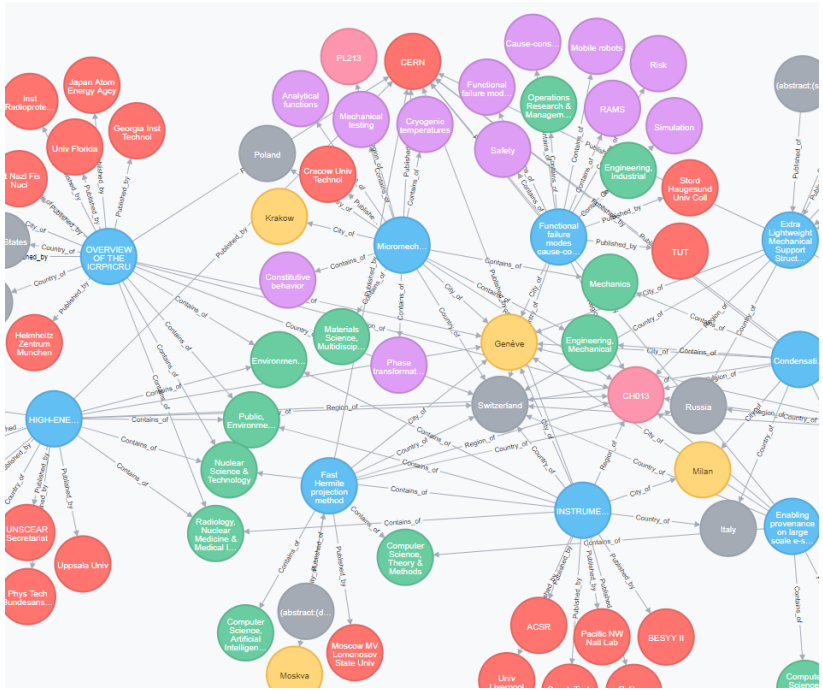
Technology Landscape
Links via common data types

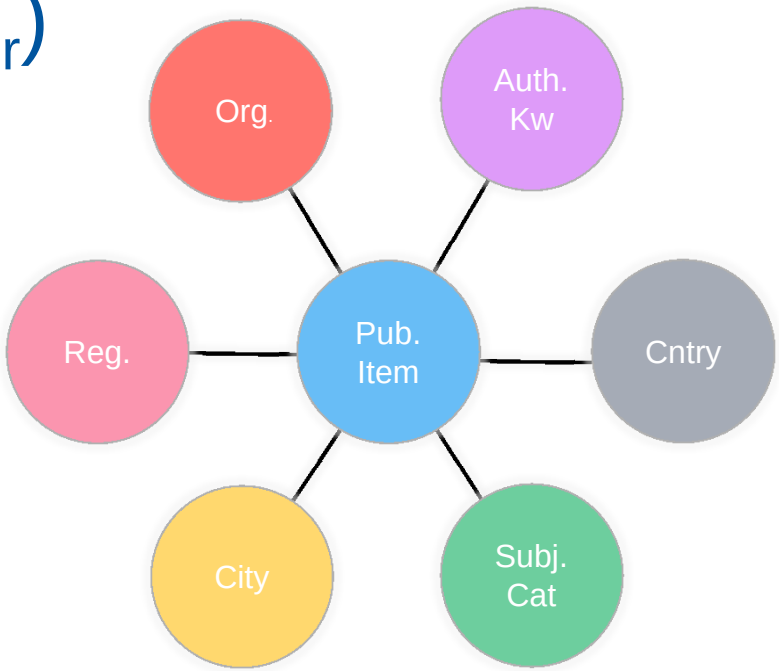


Technology Landscape



Knowledge graph
(α: homomorphism $G \rightarrow G_r$)





- Reachability graph $G_r = (L, E_r)$

Linking Data for Drug Discovery

GO (Gene Ontology)

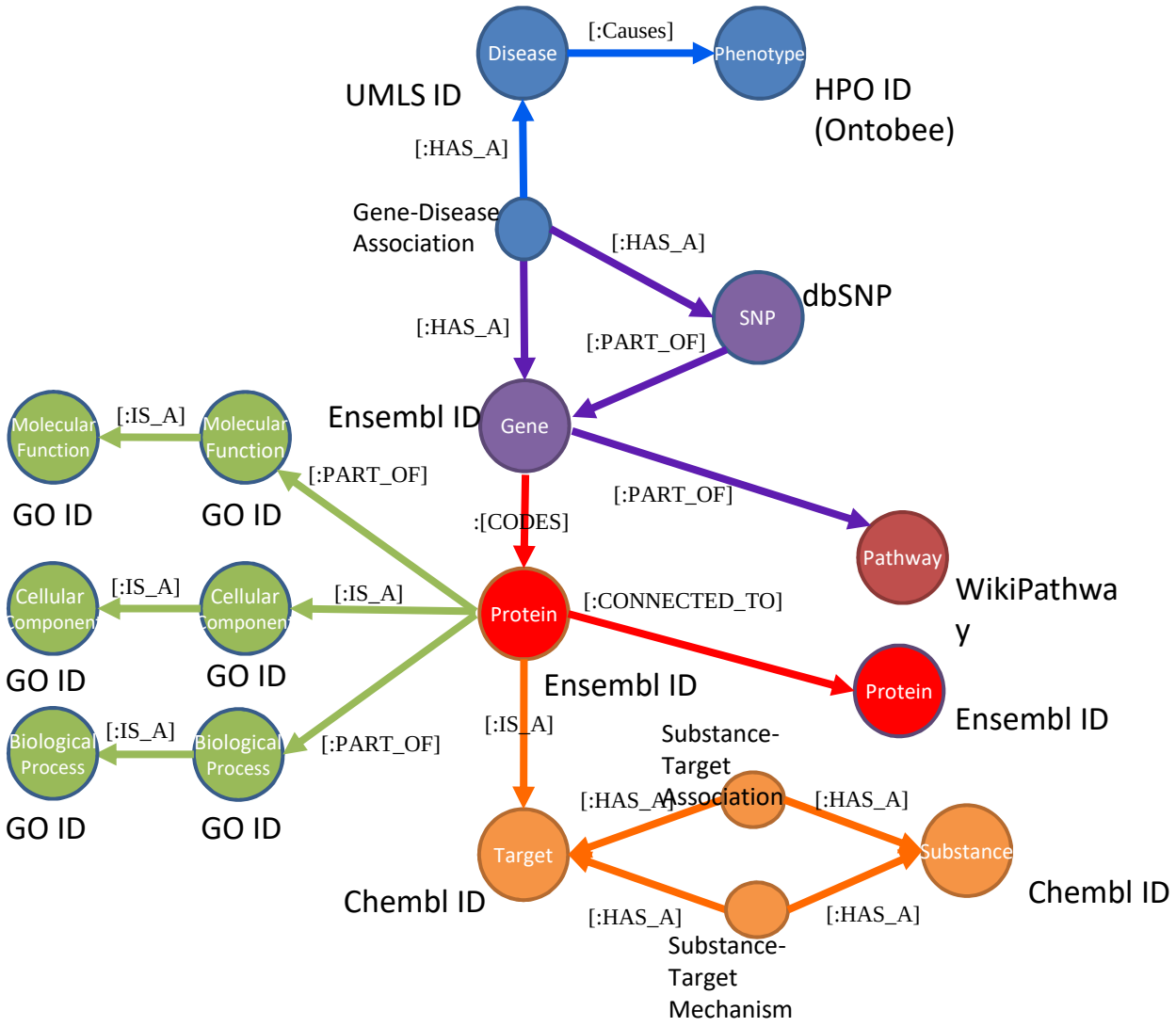
DisGenet (+ phenotypes)

Ensembl (dbSNP)

STRING

Wikipathway

Chembl

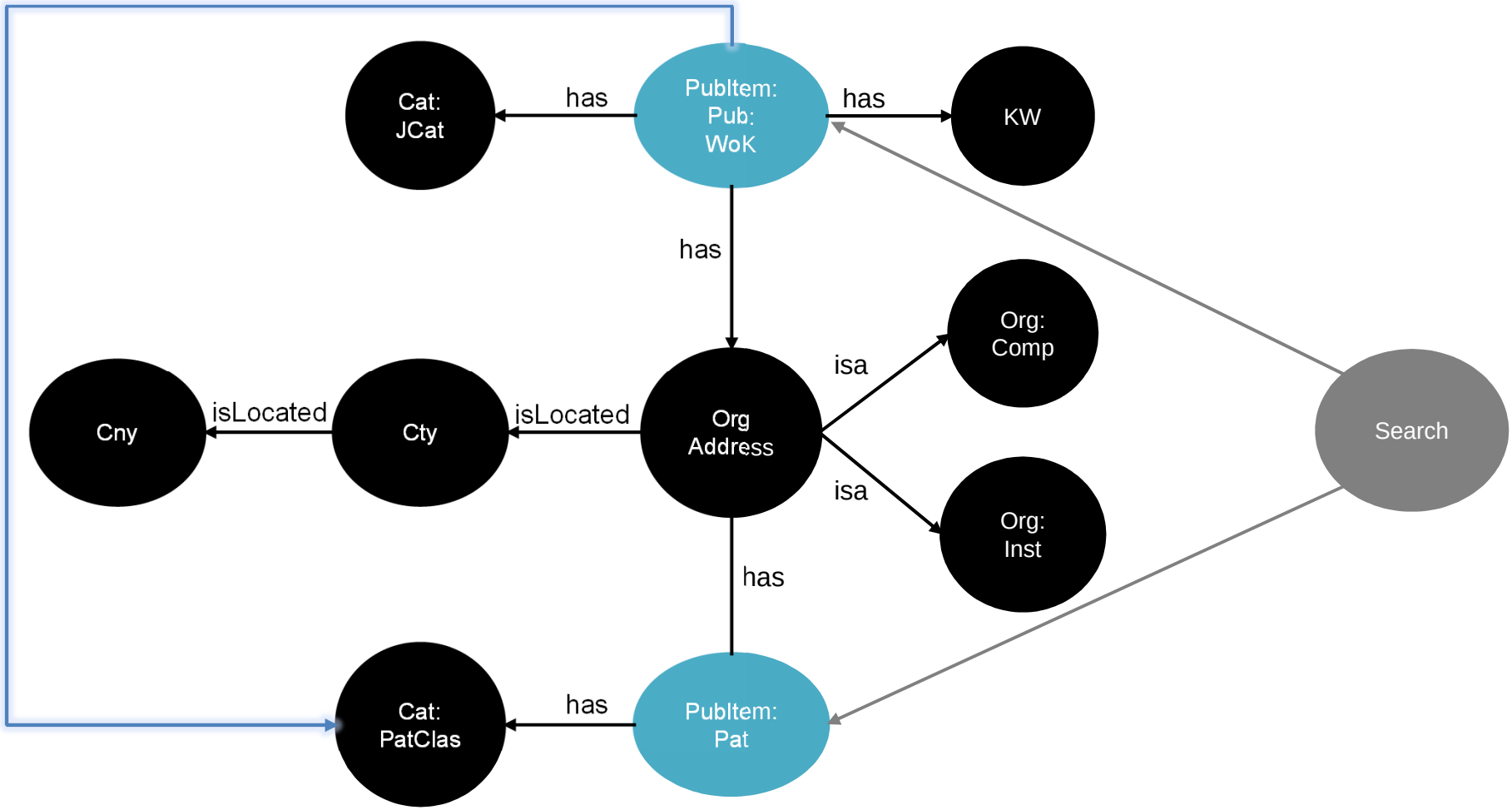


Links via derived relationships

Linking patent with publications

- Patent classes ⬅➡ Journal categories
  - Have to use citations
  - 8% - 12% of patents citing publications and even less publications citing patents
  - Citations may not be specific to the search
- Use the document embedding techniques to position publications w.r.t. patents and assign a patent class
- Links via associations

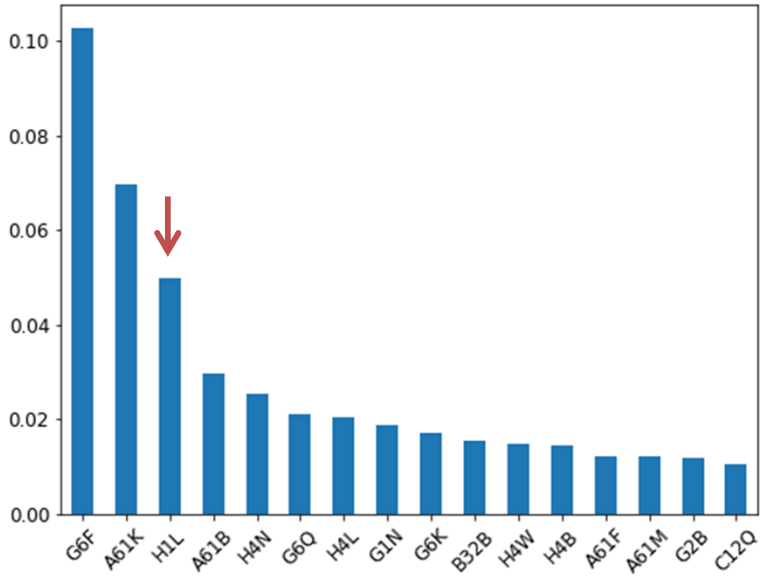
Linking patents & publications
Derive new relationships?



Search extended to publications

- Publications: Titles + Abstracts
- Publications offer a different viewpoint
- Publications are classified according to the n-closest patent classes

- G6F: Electric Digital Data Processing (Through Silicon Via (Antenna))
- A61K: Medical or Veterinary Science (Taura-Syndrom-Virus, tachycardia beat, Tellerspülvermögens)



Thank you for your attention!