Maastricht University Leading in Learning!

semantic pipelines to

Molecular Properties

Egon Willighagen (@egonwillighagen)

Dept. of Bioinformatics – BiGCaT, Maastricht University

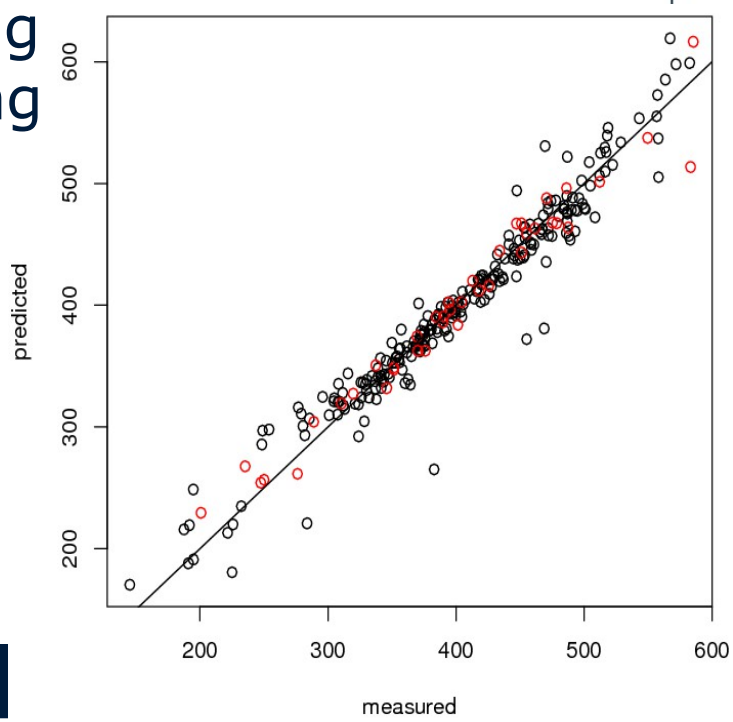
#ACSPhilly 2012

Properties: *Physical, chemical, biological*

- From experiment
  - with **experimental error**: interlab differences, method differences, …
- No experiment? *Predict*
  - with **prediction error**, originating from experimental error, modeling error, incorrect interpretation, …

- ***Where does the prediction error come from?***

Predicted BPs →



Richer Property Prediction

1. Select training data (x,y)
   - experimental data ***with*** errors, detailed description of what the data is
2. Find f(x) = y, minimizing error(y)
   - use ***as much as possible*** information from 1.
3. Validation
   - not just validate (and quantify) the statistics against a *test set*, also compare with ***other*** data

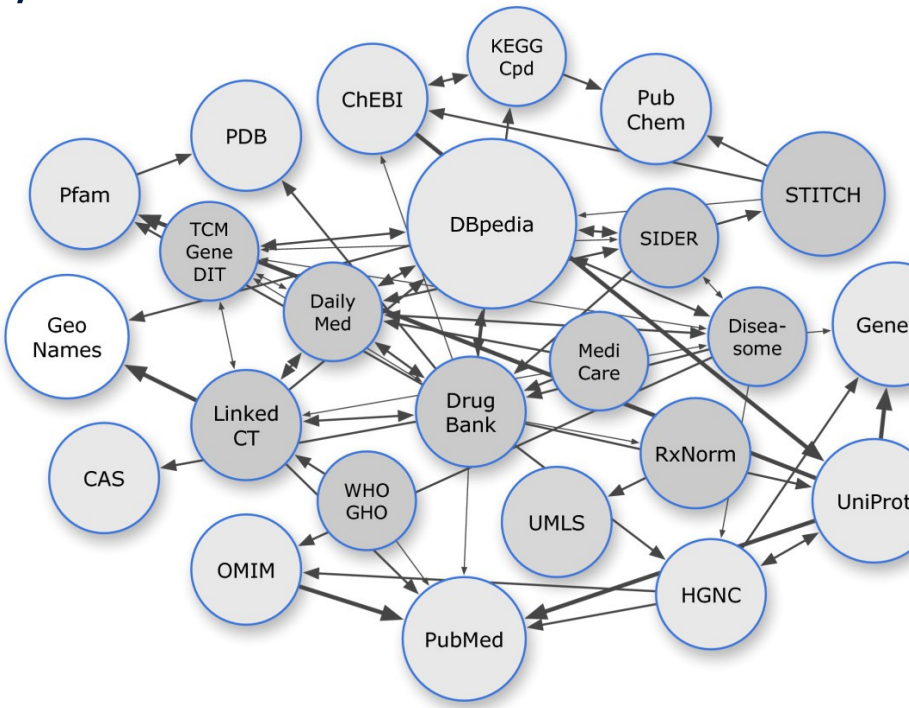
How? → Semantic pipelines

E.L. Willighagen, *et al*. Molecular Chemometrics. Crit. Rev. Anal. Chem. 2006.

Semantic pipelines

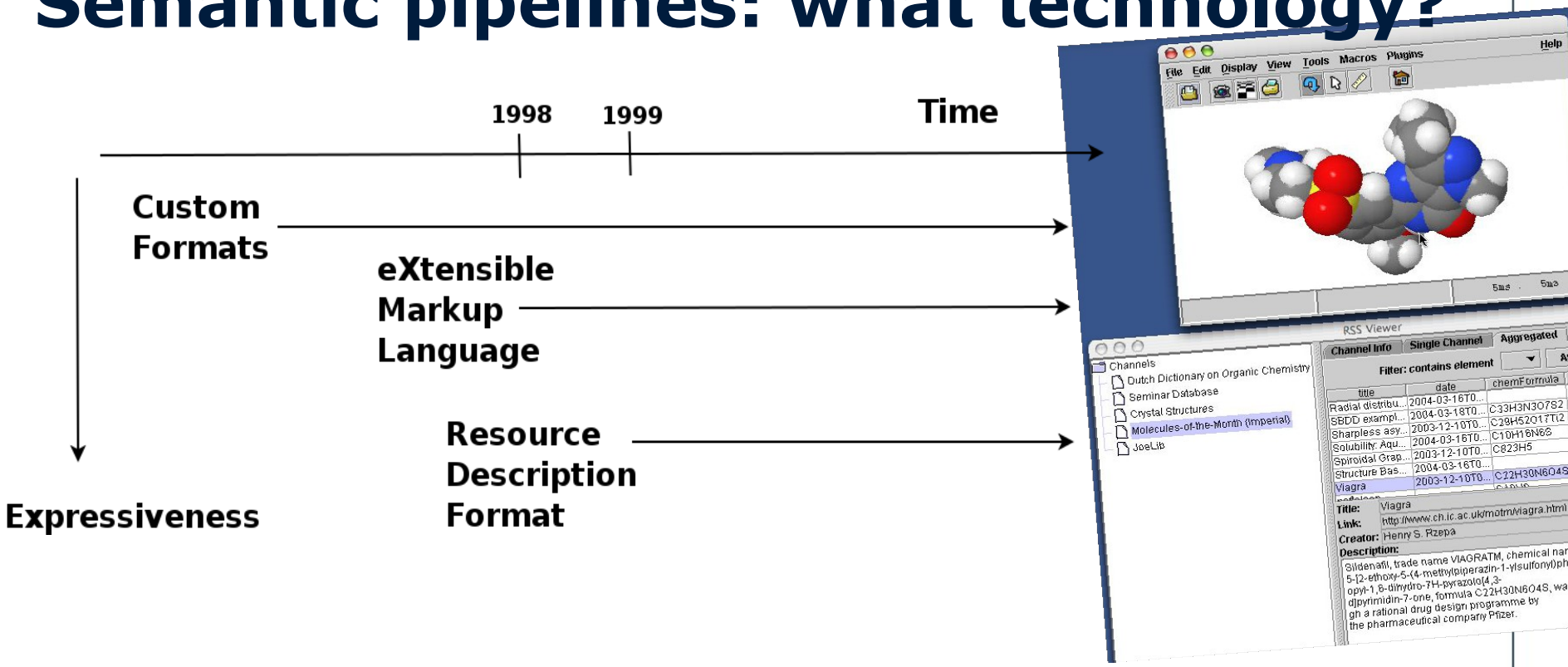
Be clear in what you mean

- Use (open) look-up lists, dictionaries, ontologies
- Be flexible and remove format limitation
- *Link to other data → from other domains*
- Calculation provenance



M. Samwald, *et al.* Linked open drug data for pharmaceutical research and development. J. Cheminf. 2011. 3:19

Semantic pipelines: what technology?



CML: *semantic, flexible, embeddable in HTML, RSS, … but only in XML*

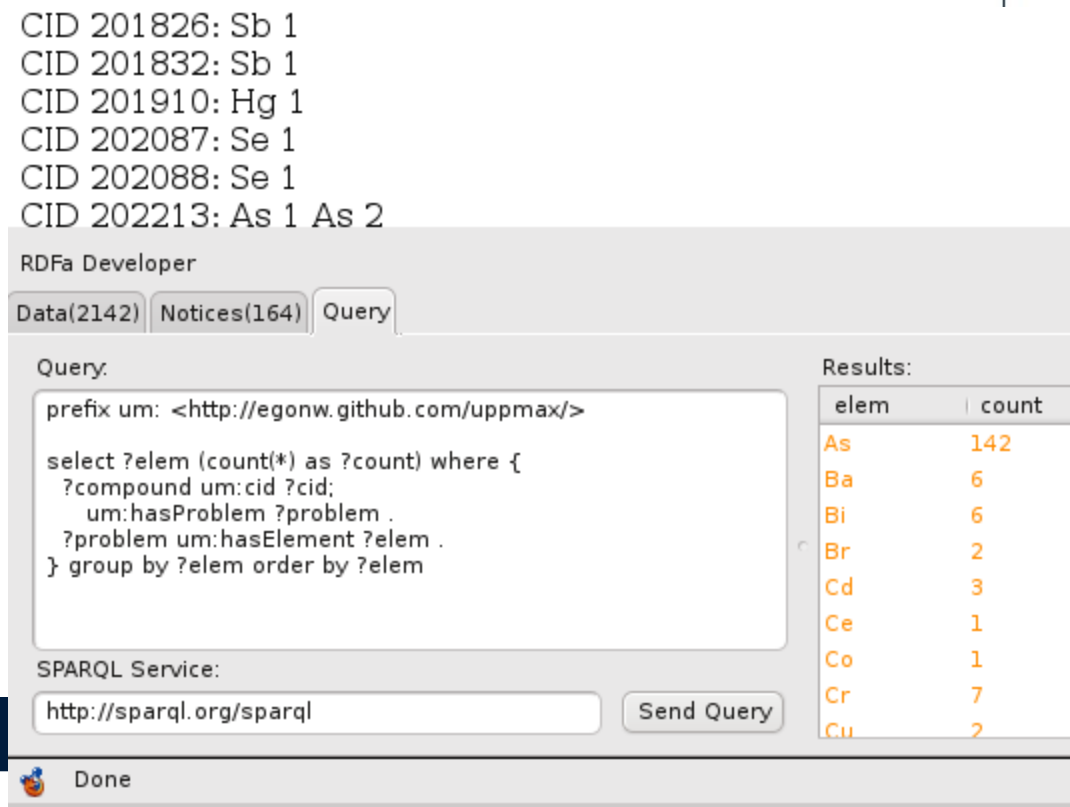
RDF: *~10yrs before adopted! But JSON, XML, Turtle, … Also: linked data.*

RDF and friends ...

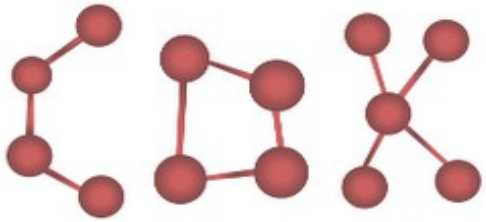
- format independent (JSON, XML, ...)
- db technology independent (RDB2RDF)
- embeddable in HTML (*e.g.* RDFa)
- Open Standard → widely supported

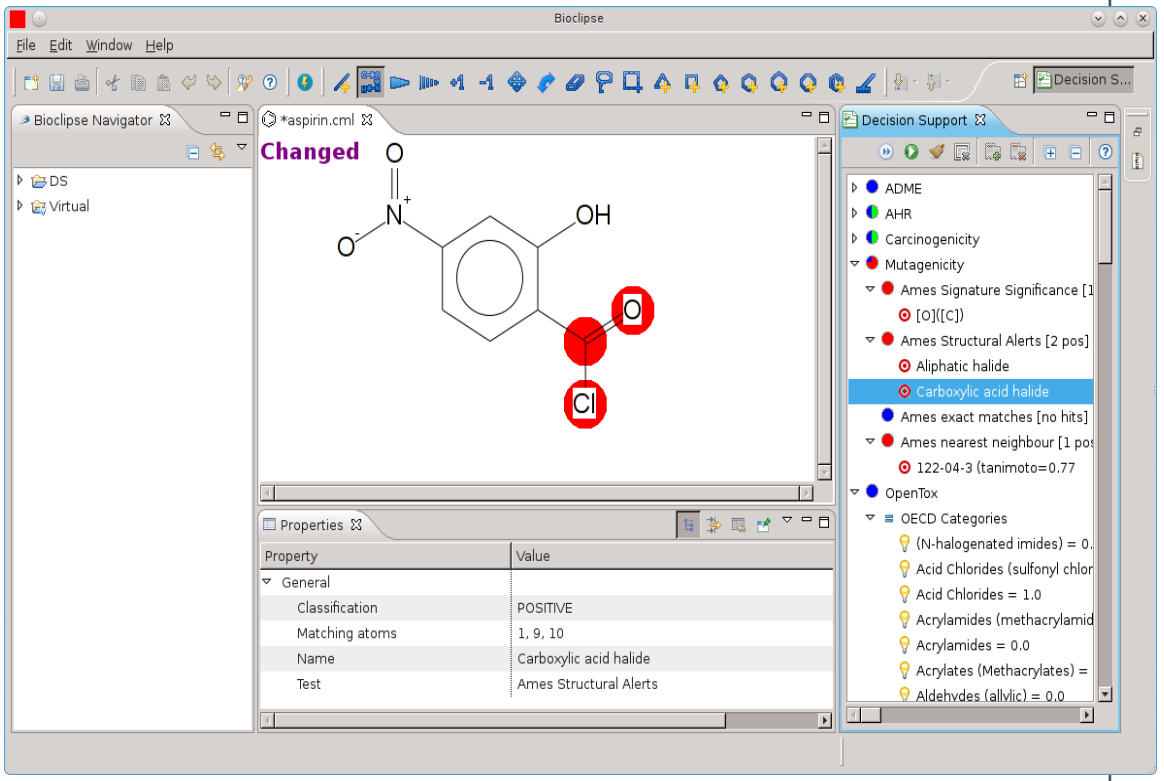
Querying

- SPARQL: like SQL access →
- ***Federated*** SPARQL link multiple SPARQL endpoints and other RDF sources



App #1: Bioclipse speaks RDF and OpenTox





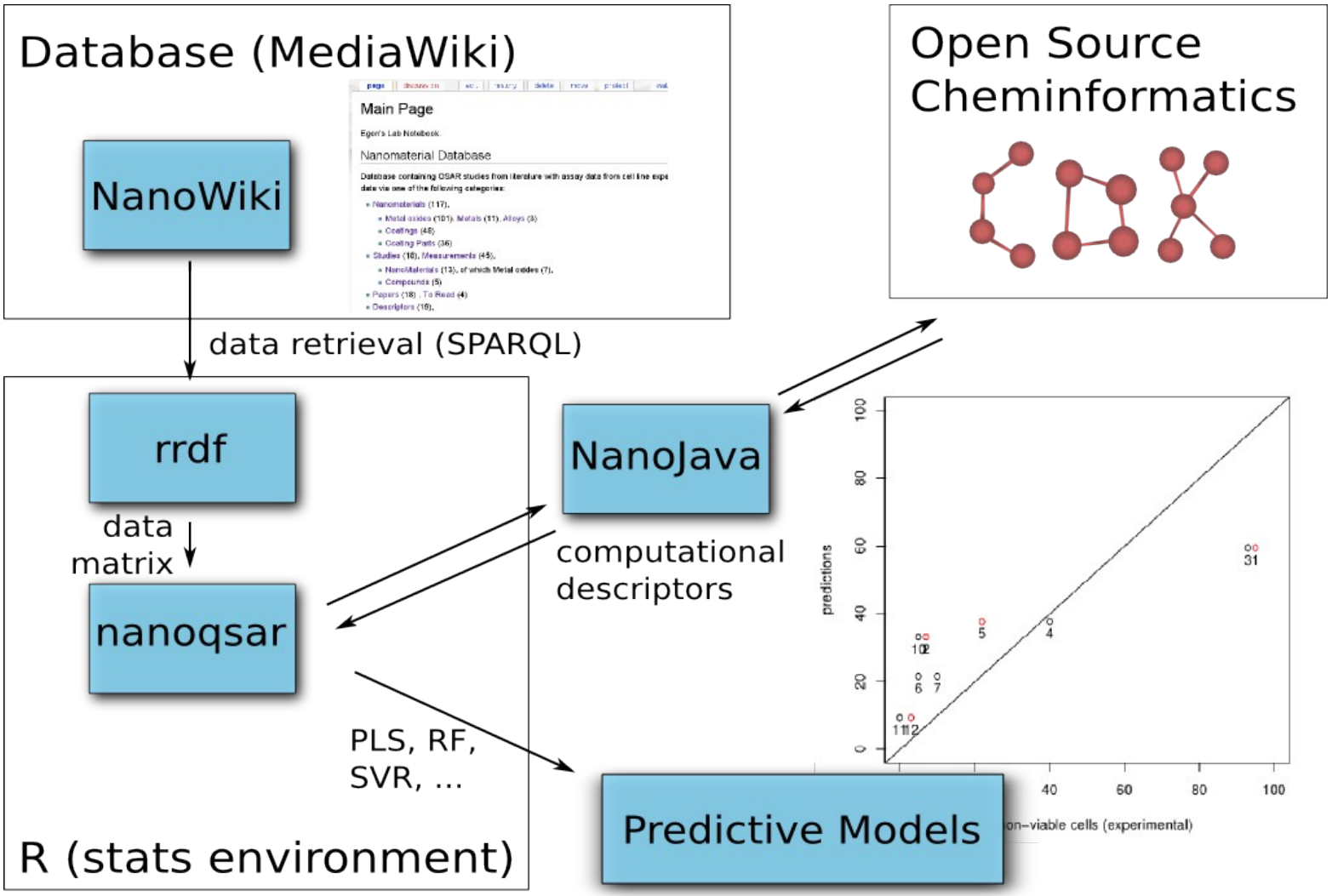
Willighagen *et al. BMC Research Notes* 2011, **4**:487
http://www.biomedcentral.com/1756-0500/4/487

SHORT REPORT **Open Access**

Computational toxicology using the OpenTox application programming interface and Bioclipse

Egon L Willighagen[1,2*], Nina Jeliazkova[3], Barry Hardy[4], Roland C Grafström[2,5] and Ola Spjuth[1]

App #2: Nanotoxicity

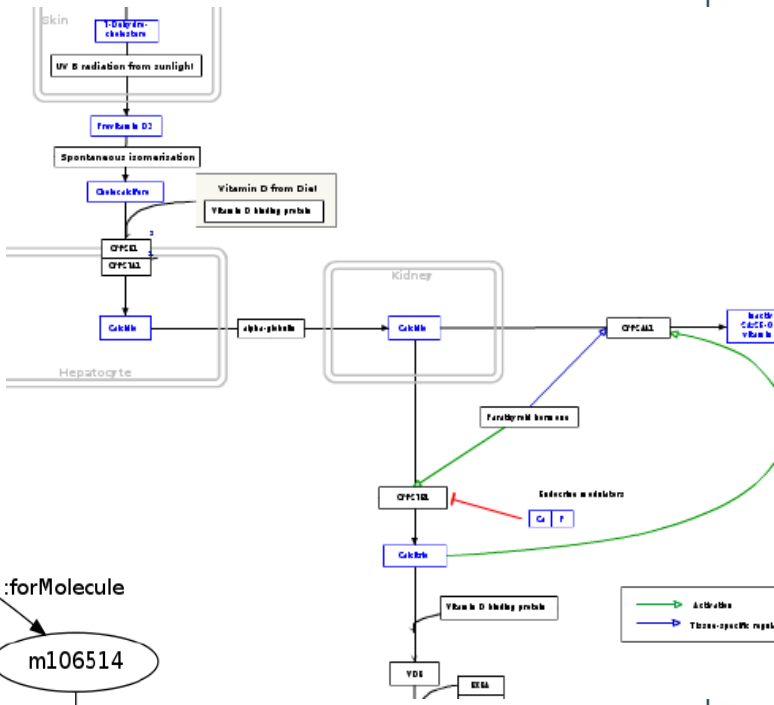


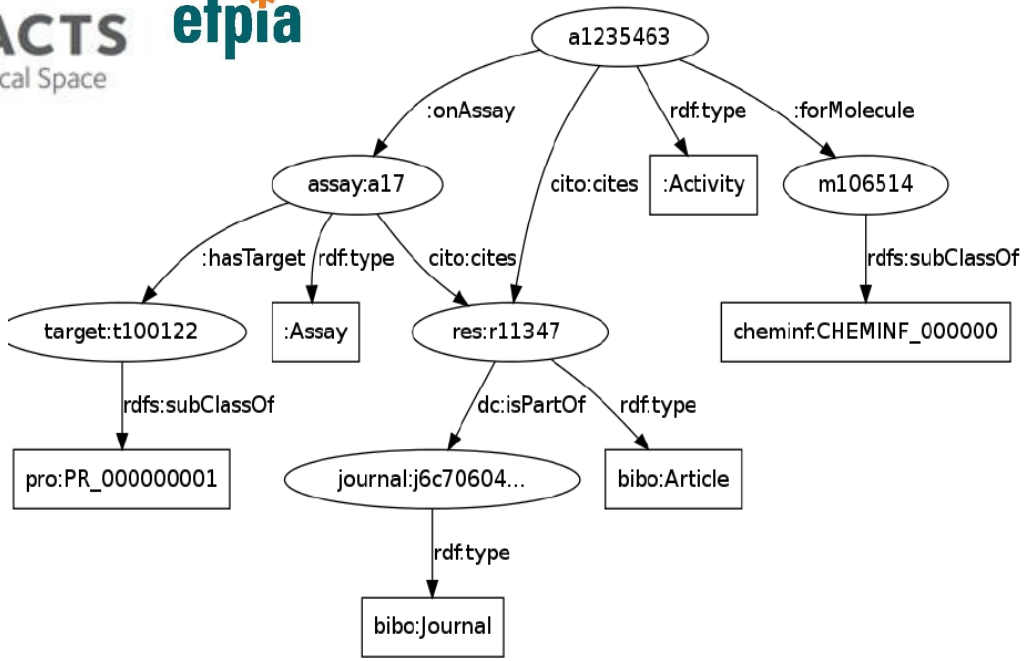
What's next?

Innovative Medicines Initiative

Open PHACTS
Open Pharmacological Space

efpia





Collaborations / Thanx

Uppsala University

Ola Spjuth et al.

Karolinska Institutet

Roland Grafström, Bengt Fadeel, Hanna Karlsson

W3C Heath Care and Life Science interest group

CDK community

Christoph Steinbeck, Rajarshi Guha, many more

OpenTox community

Nina Jeliazkova, Barry Hardy

CHEMINF community

Nico Adams, Michel Dumontier, Janna Hastings

CML community (you know :)

Take home tweet

Improve your property prediction training and validation by adopting semantic pipelines!

Blue Obelisk

Further examples:

blog: chem-bla-ics

twitter: @egonwillighagen

CiteULike: egonw

Mendeley: egon-willighagen

GScholar: u8SjMZ0AAAAJ