Bio2RDF: Convert, Provide And Reuse.

Marc-Alexandre Nolin[1], Jacques Corbeil[1],
Luc Lamontagne[1], Michel Dumontier[1,2]

[1] Laval University, Canada
[2] Carleton University, Canada

manolin@gmail.com

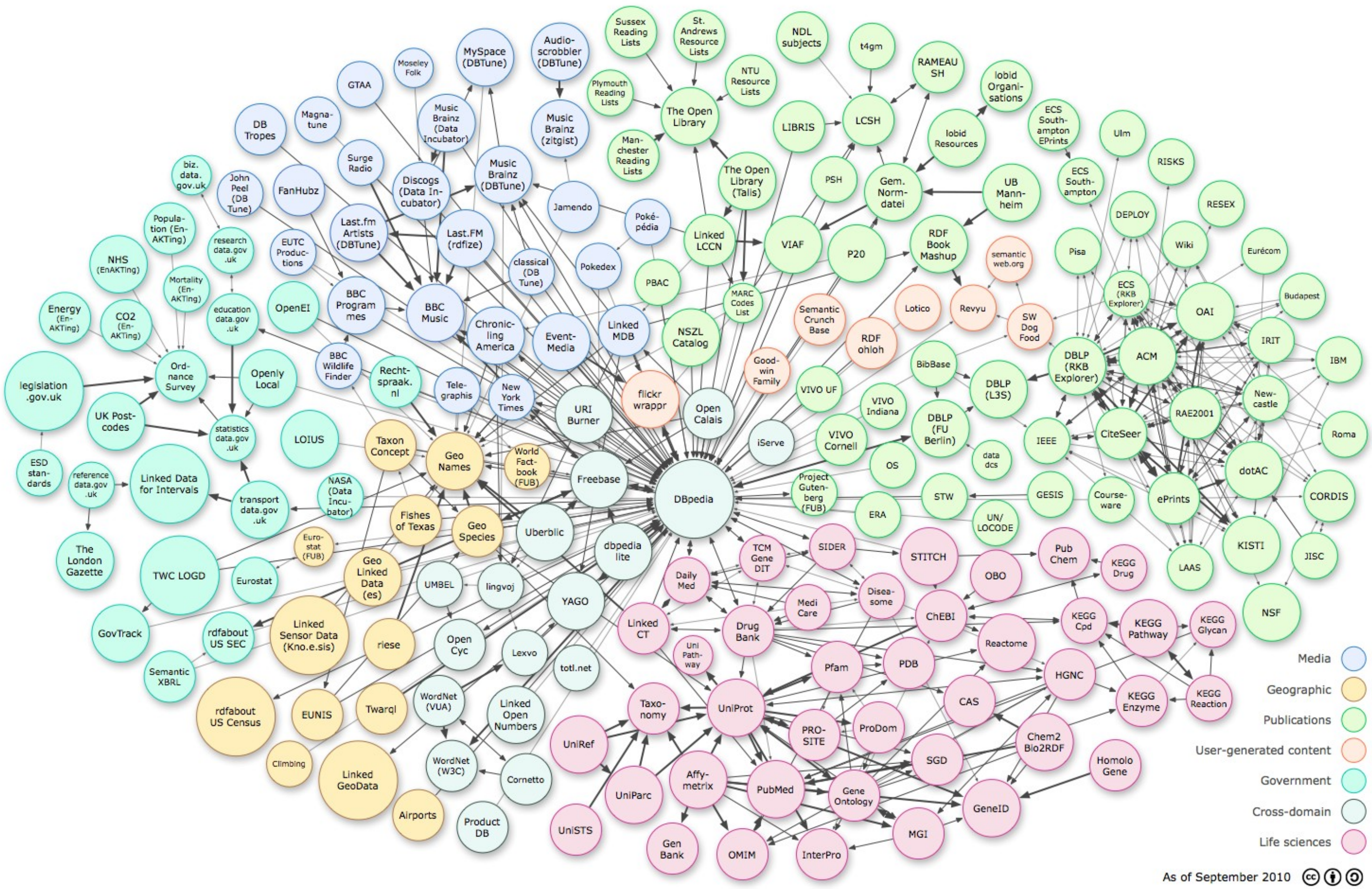
Presentation Plan

- Bio2RDF
  - Methodology
- Data conversion
- Data provision
- Reuse
- Future work

Bio2RDF

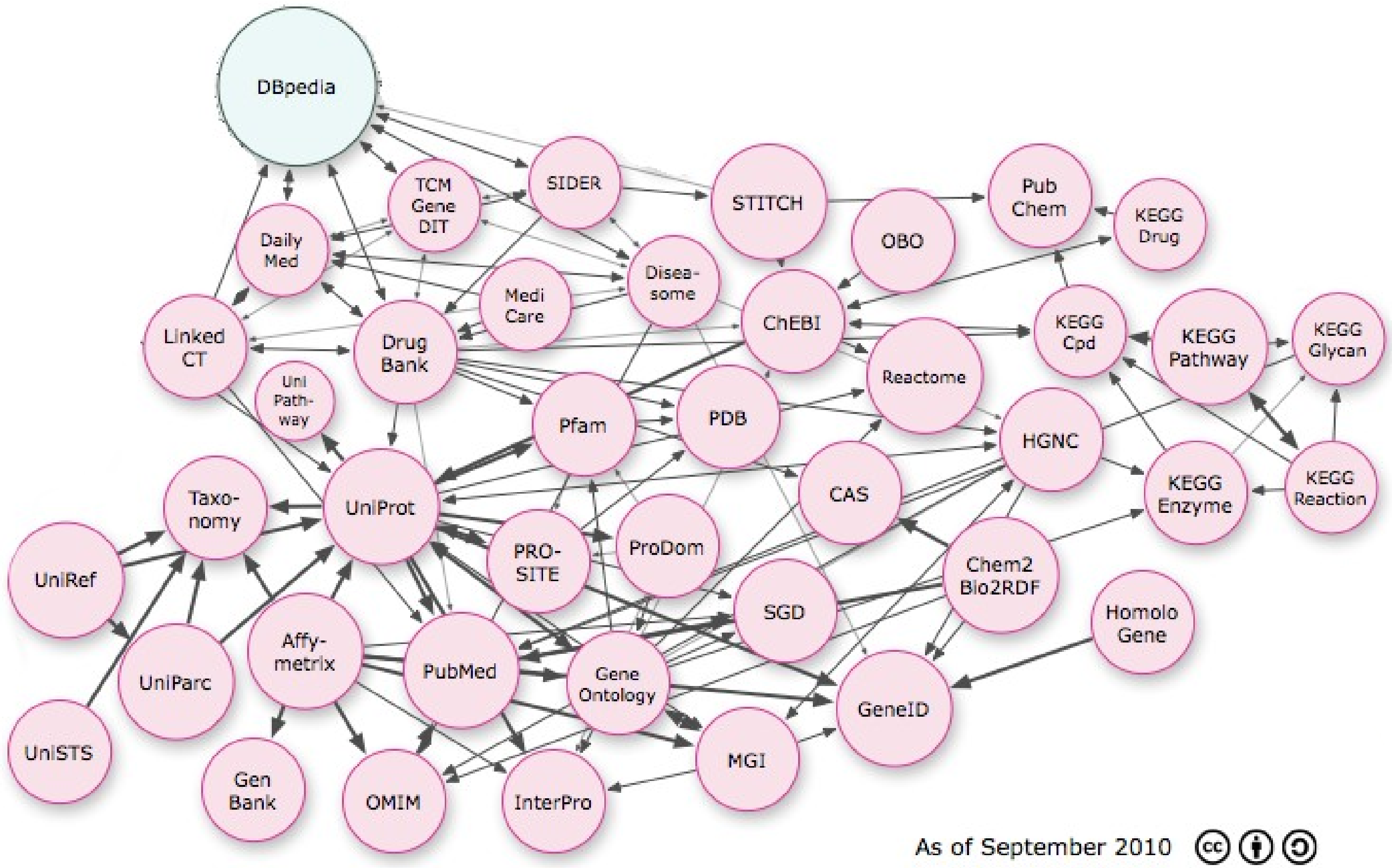
- Bio2RDF uses open-source Semantic Web technologies to provide interlinked life science data to support biological knowledge discovery.
  - Over 40 databases converted
  - Over 30 billion triples
  - Global mirroring
  - – Quebec City, Quebec, Canada
  - – Ottawa, Ontario, Canada
  - – Guelph, Ontario, Canada
  - – Brisbane, Australia
- – part of LOD mashup at http://lod.openlinksw.com

2010 Linked Open Data Cloud



"Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/"

Bio2RDF is the major contributor to the Life Sciences LOD



"Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/"

What is Linked Open Data?

- Linked open data is
  - data that is free to use
  - machine understandable (uses RDF/OWL)
  - can be looked up using web protocols
  - has meaningful relations between data items (generated from supplied cross-references, or text-based mappings)

LOD Methodology

Applying Tim Berners-Lee 4 rules

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. **<u>Include links to other URIs, so that they can discover more things</u>**

– http://www.w3.org/DesignIssues/LinkedData

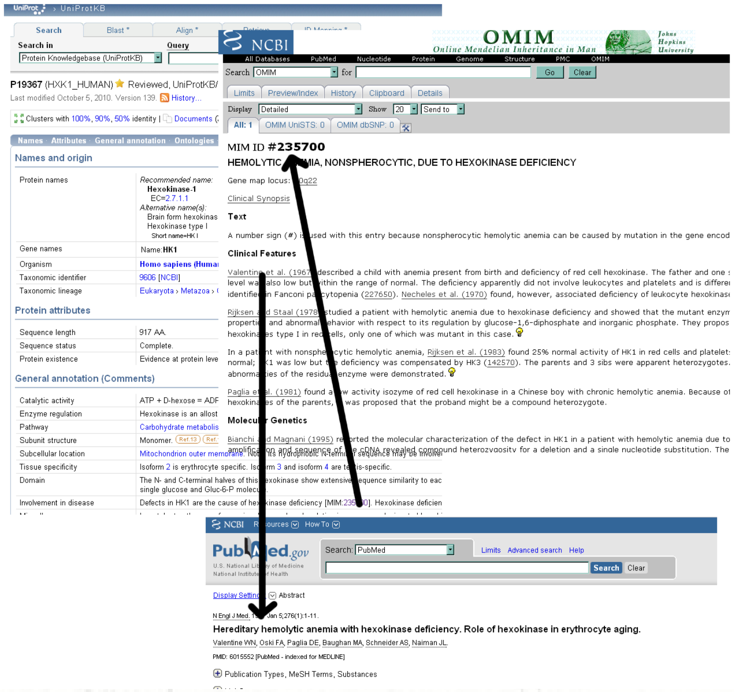
Bio2RDF Methodology

We extend LOD rules by:

1. Consistently naming **all** resources
   http://bio2rdf.org/namespace:identifier
2. Resolving Bio2RDF URIs to a set of statements about the requested resource

Data providers have been linking data for years

- Links are done from one HTML page to another
- This works for human consumption, but doesn't scale with huge amounts of data

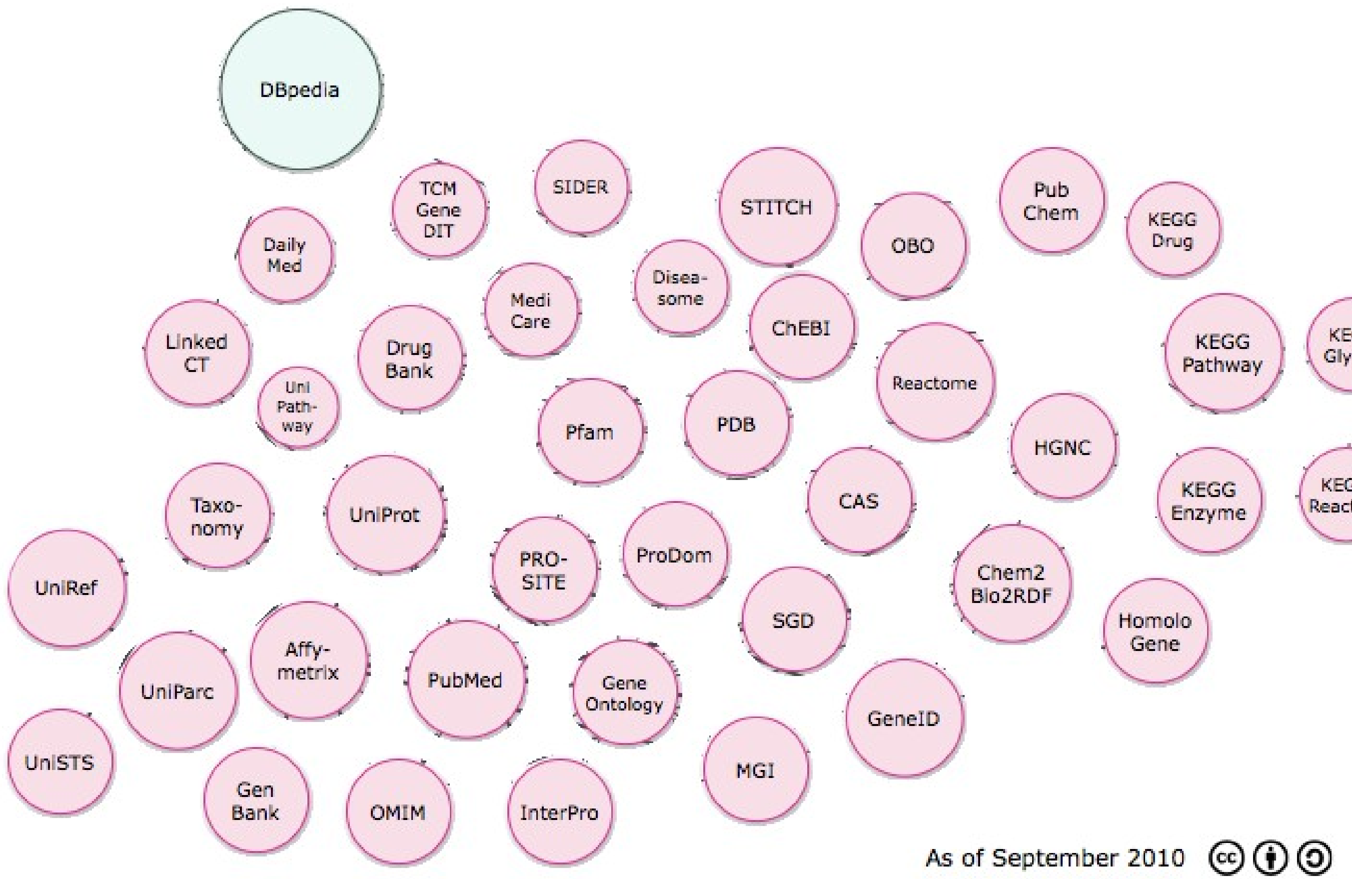


However, most linked open data is created and provided *without* the help of the original data provider

Why be part of the linked data cloud?

- Enable queries that span over more than 1 database.
  - Example: filtering a PubMed search by a microarray level of expression filter
- Reduce the size of a database by only referencing data instead of including it in a database record (e.g. citations)

But something is missing !



derived from Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/”

Presentation Plan

- Bio2RDF
- <u>**Data conversion**</u>
- Data provision
- Reuse
- Future work

Convert

- Take a step back and think about what your data represents.
- Forget about the way it is currently represented in your relational database

What is RDFizing?

- **RDFizing** converts legacy data into one or more RDF statements or triples:

  <Subject> <Predicate> <Object>

- Triples have correspondence for a standard database
  - Subject → row identifier
  - Predicate → column name
  - Object → value

Convert

- Converting to RDF just for the sake of providing another format will not add much to your offering if the result is your old relational database format … in RDF
- Get rid of relational database artifacts while **rdfizing**

1. Uses simple and stable identifiers to name resources
2. Create ***types*** for the entities that your data refers to with in your data and specify the nature of the ***relations*** that hold between them

Presentation Plan

- Bio2RDF
- Data conversion
- <u>**Data provision**</u>
- Reuse
- Future work

Provide – RDF Documents

- URIs can be created with a REST-like look
- Example of stable URIs

  http://purl.uniprot.org/uniprot/P19367

  http://bio2rdf.org/uniprot:P19367
- Documents containing statements should be accessible using web technology (HTTP protocol)
- Provide data dumps
- But in the LOD world, we want to query online databases!

Provide - SPARQL

- SPARQL is the query language for RDF/OWL that uses web technology (HTTP)
- SPARQL endpoints make it possible to query databases using SPARQL
- Distributed SPARQL will carve up the query and determine which endpoints need to be queried

Provide

- Publish the scheme you will use for your URIs so that other providers may use it
- Provide access to documents with resolvable URIs (can be looked up using a web browser)
  http://geneprovider.com/gene:identifier
- Now other data providers can use this identifer instead of copying the data into their own!

Presentation Plan

- Bio2RDF
- Data conversion
- Data provision
- **Reuse**
- Future work

Reuse

- RDF version of your documents without resolvable external links is just another file format
- One of the most problematic issues is that the RDF generated by some providers are only inward looking -> they don’t reuse published URIs (Polite URI)

Reuse

- Example with Uniprot RDF
- Uniprot is one of the first data providers to offers stable and resolvable URI for its documents. However, we can't use directly the RDF they provide. Look at this extract of Human HK1 in RDF

```
<rdf:Description rdf:about="http://purl.uniprot.org/uniprot/P19367">
    <rdf:type rdf:resource="http://purl.uniprot.org/core/Protein" />
    <rdfs:seeAlso rdf:resource="http://purl.uniprot.org/refseq/NP_277035.2"/>
</rdf:Description>
```

- The problem is that `http://purl.uniprot.org/refseq/NP_277035.2` resolves to the NCBI HTML page of NP_277035.2
- Since NCBI does not provide RDF, it's a dead end

Reuse

- Uniprot did the same things we have done at Bio2RDF. They create a URI in their namespace. From that URI which they control, they decided to redirect to the original HTML document of the specified ID
- The difference with Bio2RDF is that we also resolve the other URI to an RDF document

Reuse

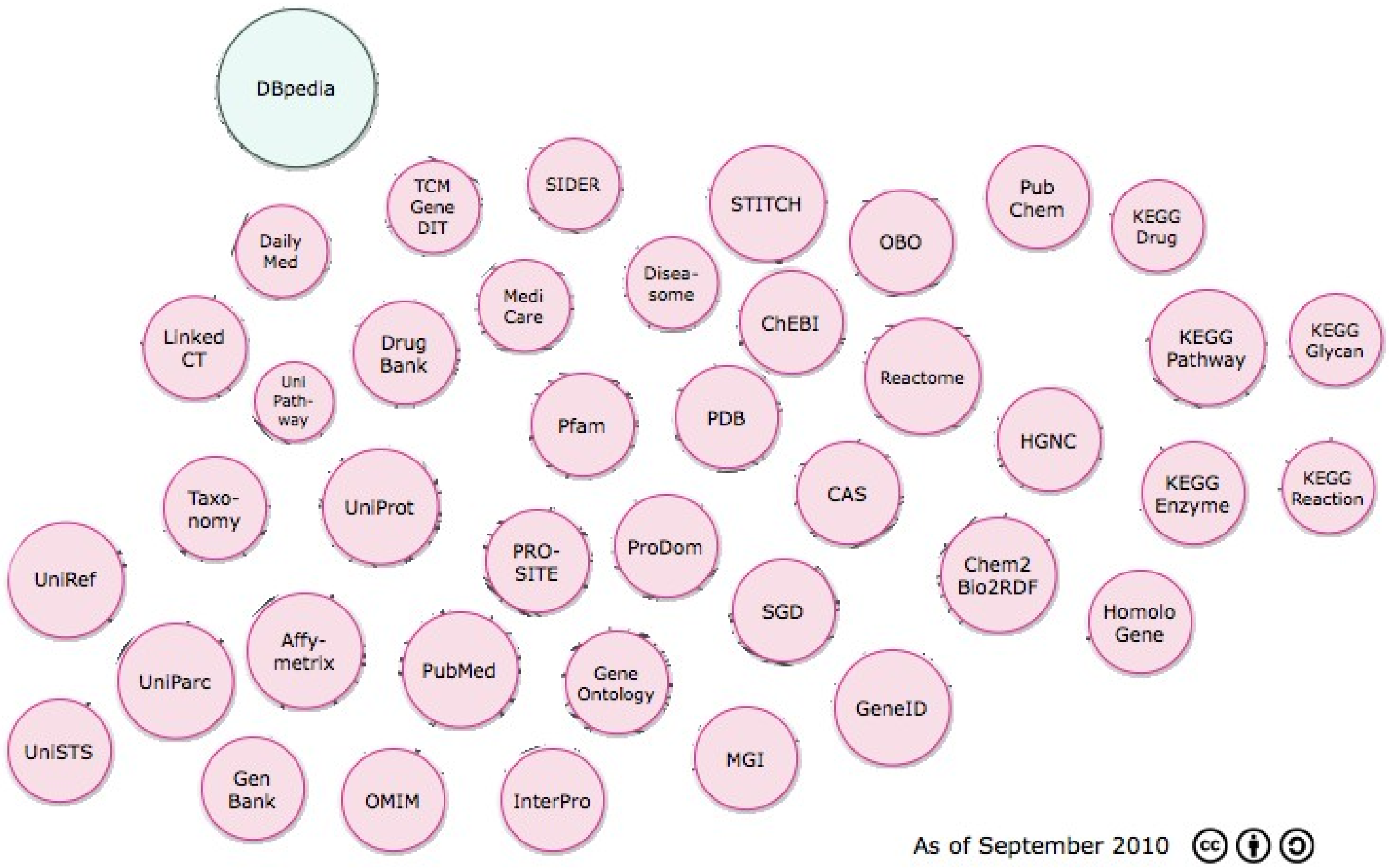
- Relational database artifacts : copying of data from one provider to another
- Uniprot Citation entry IN a protein document

```
<rdf:Description rdf:about="http://purl.uniprot.org/citations/10686099">
<rdf:type rdf:resource="http://purl.uniprot.org/core/Journal_Citation" />
<title>Crystal structures of mutant monomeric hexokinase I reveal multiple
    ...</title>
<author>Aleshin A.E.</author>
<author>Kirby C.</author>
<skos:exactMatch rdf:resource="http://purl.uniprot.org/medline/20223513" />
<skos:exactMatch rdf:resource="http://purl.uniprot.org/pubmed/10686099" />
</>
```

- In a Linked Data world, you only need to have a link to the PubMed URI at NCBI. The up to date information is there.

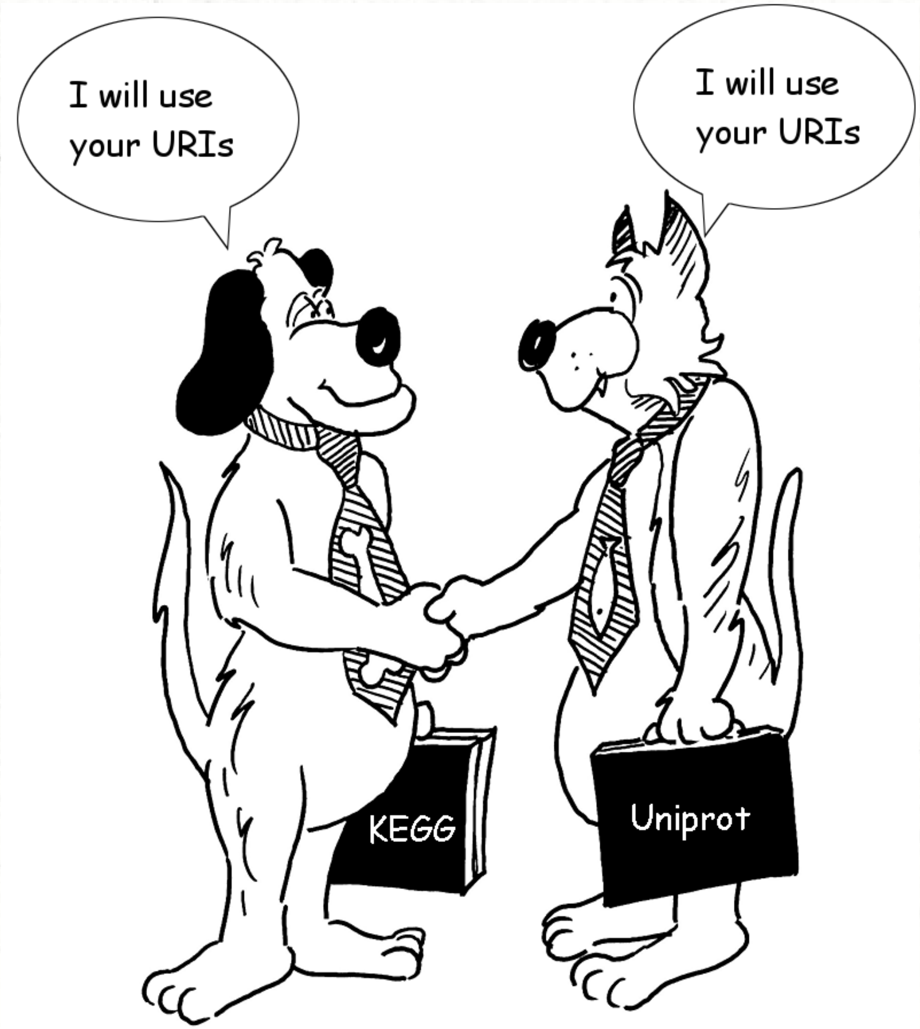
With all this, how is now the LOD for life sciences by original data providers ?

What is still missing !



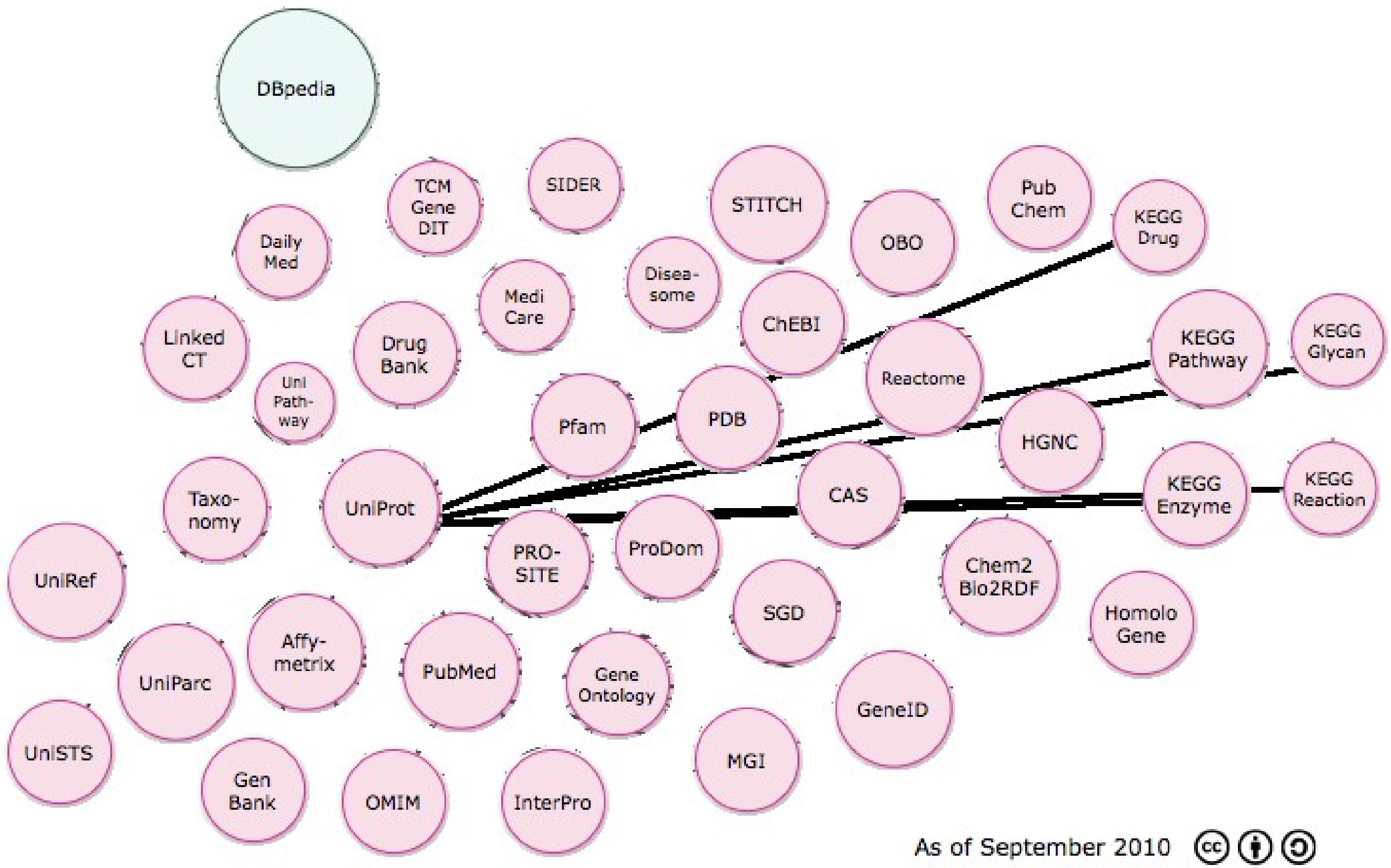
"Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/"

Some handshake



The new version of LOD for life sciences by original providers

Something is missing !



"Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/"

Bottom line

- That is not quite what we had at the beginning
- The original network made by third party provider still exist. Use it like if it was another provider.
- What about documents without an RDF version?
  - From NAR, there is 1230 databases. Too much for Bio2RDF or any other to convert entirely.
  - Use third party providers, like Bio2RDF, NeuroCommons, LinkedLifeData, Shared Names, etc.
  - Add these in addition to original data URIs
  - Some databases, for various reason, might never create RDF version of their data. So third party will continue to have their use by providing their data in RDF

By applying those conversion, publication and reuse rules to your data, we will witness the birth of a more stable network of linked data … and the death (in the very long term) of Bio2RDF.

Let's kill Bio2RDF together !!

Presentation Plan

- Bio2RDF
- Data conversion
- Data provision
- Reuse
- **Future work**

Future Work

- Data processing workflow
- New facet-based user interface to browse and formulate sophisticated queries
- Full text indexing for autocompletion support
- Exploring knowledge discovery possibilities in the linked data network

Thanks

- Bio2RDF community
  - Centre de recherche du CHUL
  - Dumontier Lab members
  - QUT eResearch Center
- Triplestore provided by Openlink Virtuoso
- François Belleau

Acknowledgment

- Marc-Alexandre Nolin funding provided by CANARIE via the C-BRASS project
- Servers in Quebec City are provided by Jacques Corbeil of Laval University

Contact Information

- Mailing list : bio2rdf@googlegroups.com
- URL : http://bio2rdf.org
- Wiki : http://sourceforge.net/apps/mediawiki/bio2rdf
- Blog : http://bio2rdf.blogspot.com