Laboratoire d'Informatique de Grenoble

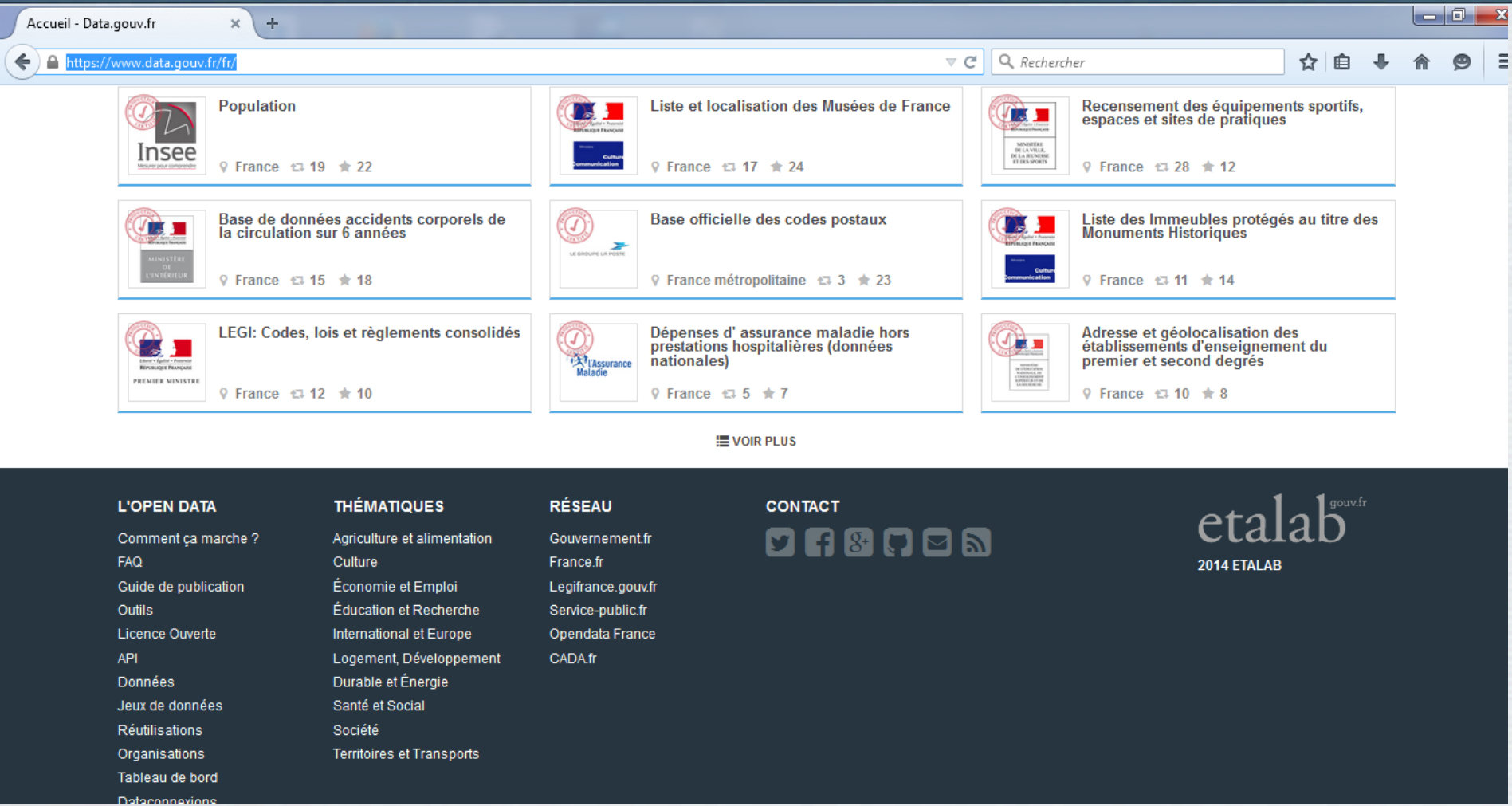
(Linked) Open Data: issues and opportunities for researchers

Marie-Christine Rousset
Laboratoire d'Informatique de Grenoble
Université Grenoble Alpes and Institut Universitaire de France

Open Data

- General trend for governments and organizations to make available more and more data produced by public services
  - for transparency effort
  - for data reuse (cross-referencing with other data sources for data aggregation and information extraction)
- Open data principles :
  - Data formats for maximal technical access
    - providing data in structured formats using open standards
    - that can be accessed by proprietary or non-proprietary software means
    - to facilitate interoperability between machines and data processing for varied uses
  - No license-related barrier to the re-use of public information
    - made explicit using the appropriate Creative Commons licences
    - example: https://www.whitehouse.gov/copyright

A representative example: data.gouv.fr



Data.gouv.fr today

- 200 data producers
  - ministries, local authorities, associations, …
- 13.000 data sets
- Technical choice: CKAN
  - an open-source tool for constructing web data portals
  - used for the UK portal and also for the European commission portals

Issues

- Creating added-value and new usages as well as discovering new correlations require crossing data from different heterogeneous sources often stored into isolated data systems

⇒ **Linked Data :** initiative of W3C and of Tim Berners-Lee for promoting his idea of **semantic web**

  - where data, distributed all over the web, are linked and can be automatically queried (by humans and also by applications) **wherever they are stored and without having to duplicate them.**

The standards underlying Linked Open Data

- http, URLs and namespaces
  - For identifying and naming entities without ambiguity
    - URLs: Uniform Resource Locator
    - Namespace:
      - A name in a namespace consists of a namespace identifier and a local name.
      - No homonym within a given namespace
- RDF (Resource Description Framework)
  - For declaring facts on entities as triples

    <subject, relation/property, object/value>
- RDFS (RDF Schema) and OWL
  - For grouping entities into classes structured in class hierarchies
  - For providing semantics to the relations and properties
- SPARQL
  - For asking queries to endpoints accessible through web services
    - http://rdf.insee.fr/sparql

Illustration: Querying DBpedia.fr with SPARQL

- DBpedia:
  - RDF version of wikipedia pages on named entities(persons, locations, etc…): 4 millions entities, 470 millions RDF facts
    - Automatic exploration and extraction of named entities and of relations from web pages (Wikipedia)
  - french version since 2012 : http://fr.dbpedia.org
- http://fr.dbpedia.org/sparql
  - What are the municipalities of Ile de France having more than 100.000 inhabitants , and their mayors?

```
SELECT ?commune ?maire
WHERE {
?commune <http://dbpedia.org/ontology/region>  <http://fr.dbpedia.org/resource/Île-de-France> .
?commune rdf:type dbpedia-owl:PopulatedPlace .
?commune dbpedia-owl:populationTotal ?population .
?commune prop-fr:maire ?maire
FILTER (?population > 100000) }
```

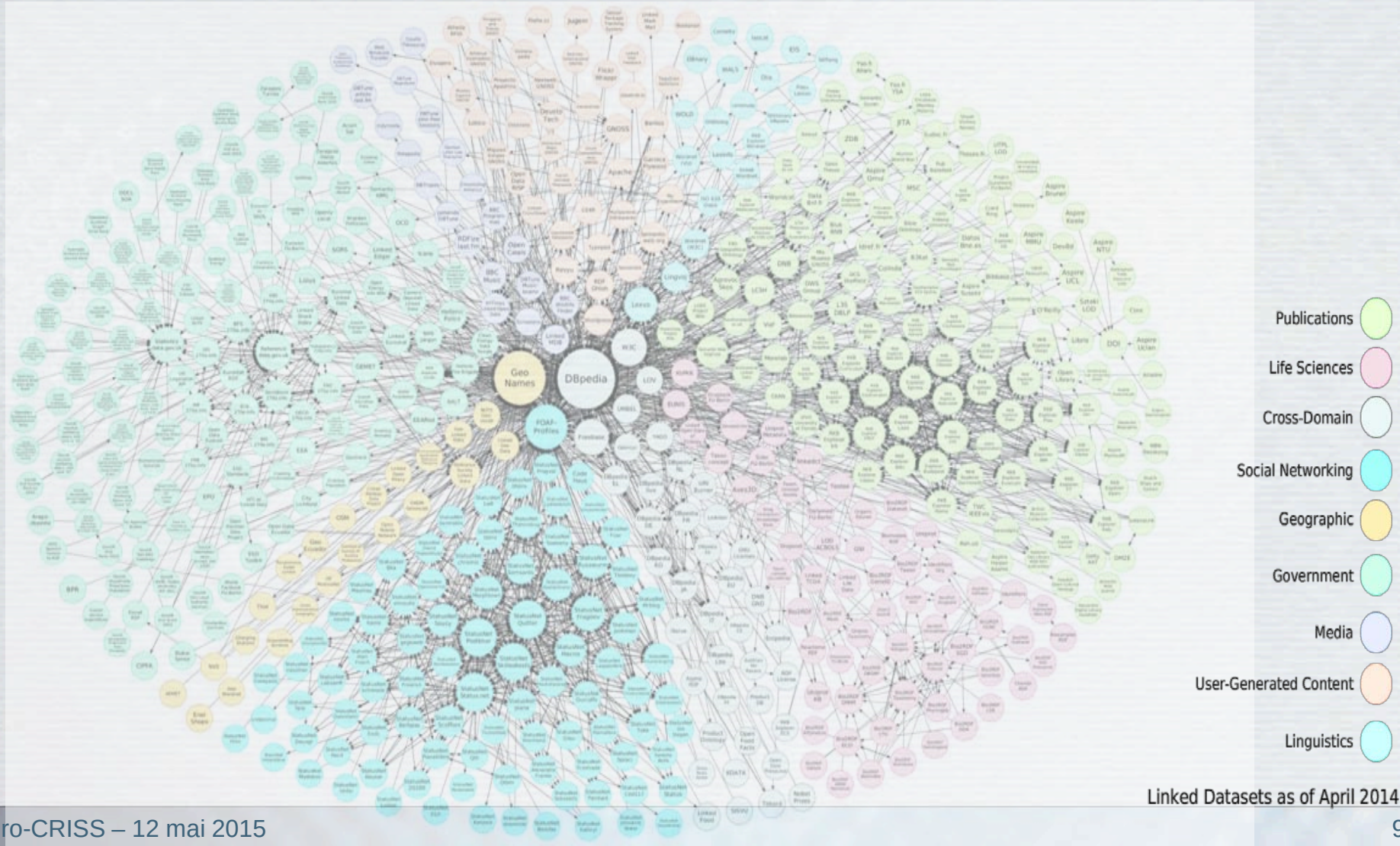
Illustration: Querying DBpedia.fr with SPARQL

- DBpedia:
  - RDF version of wikipedia pages on named entities(persons, locations, etc…): 4 millions entities, 470 millions RDF facts
    - Automatic exploration and extraction of named entities and of relations from web pages (Wikipedia)
  - french version since 2012 : http://fr.dbpedia.org
- http://fr.dbpedia.org/sparql
  - What are the municipalities of Ile de France having more than 100.000 inhabitants and their mayors?

commune	maire
http://fr.dbpedia.org/resource/Saint-Denis_(Seine-Saint-Denis)	http://fr.dbpedia.org/resource/Didier_Paillard
http://fr.dbpedia.org/resource/Boulogne-Billancourt	http://fr.dbpedia.org/resource/Pierre-Christophe_Baguet
http://fr.dbpedia.org/resource/Montreuil_(Seine-Saint-Denis)	http://fr.dbpedia.org/resource/Dominique_Voynet
http://fr.dbpedia.org/resource/Argenteuil_(Val-d'Oise)	http://fr.dbpedia.org/resource/Philippe_Doucet_(homme_politique)
http://fr.dbpedia.org/resource/Paris	"Bertrand Delanoë"@fr

Linked Open Data today

Thousands RDF data sources accessible on the Web, billions of triples



Different means for declaring links in LOD

- owl:equivalentClass
- owl:sameAs
- rdfs:seeAlso
- skos:closeMatch
- skos:exactMatch
- skos:related
- foaf:homepage
- foaf:topic
- foaf:based_near
- foaf:maker/foaf:made
- foaf:page
- foaf:primaryTopic

- Example:

http://dbpedia.org/resource/Canberra
owl:sameAs
http://rdf.freebase.com/rdf/en.canberra

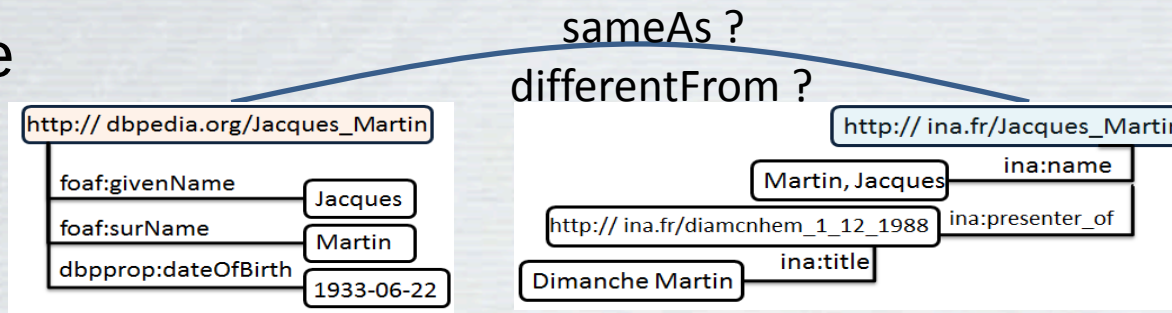
45 millions sameAs facts in DBpedia (linking Dbpedia URLs with entities of other data sources of Linked Data).

Applications developped on top of Linked Data: examples

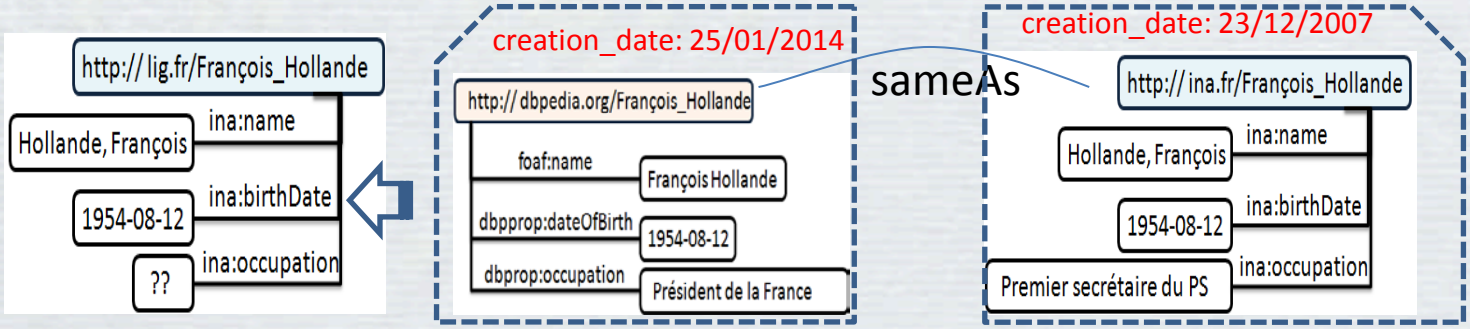
- BBC Olympics Data Service
  - Create on-the-fly interactive content on the different events by real-time integration of a rich content (existing in the RDF platform of BBC) associated with the named entities participating to such or such event.
- EPA Linked Data
  - On-demand visualization on a map of facilities likely to produce chemical substances in a certain geographical area, and links with scientific documents reporting varied statistics relevant to these substances and their effect on environment or health,…
  - Requires to cross data from Environmental Protection Agency , geonames, and data.gov

Open challenges of Linked Open Data

- Automatic Data Linkage



- Data fusion



- Data quality

Opportunities for researchers

- In more and more disciplines, varied and massive data are becoming objects of study for research:
  - raw low-level data produced by sensors, logs, traces
  - data produced by users (tweets)
  - results from experiments, polls or simulation
- filtered, analyzed, aggregated, mined, using methods that are also varied and the results of which are published :
- in bigger and bigger bibliographic databases,
- providing novel objects of study for other researchers

Towards Linked Open **Research** Data ?

- Linked Open Data offer a flexible infrastructure and robust tools
  - for sharing research data at large scale
  - based on **decoupling** **(massive) data storage** in data centers, and of **associated metadata** in RDF triplestore servers
  - for a fine-grained description in a uniform setting of :
    - the stored data sets
    - the algorithms and experimental protocols applied for analyzing them
    - the associated scientific articles

⇒ « data labs » for researchers to test and compare different methods of data analytics on data of different size and nature