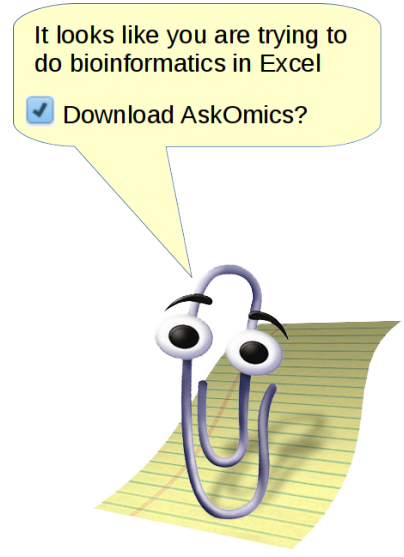
Integrating and querying data with AskOmics

Anthony Bretaudeau, Olivier Dameron,
Olivier Filangi, Xavier Garnier, Fabrice Legeai

IRISA, France

June 1, 2024



Version 1.0

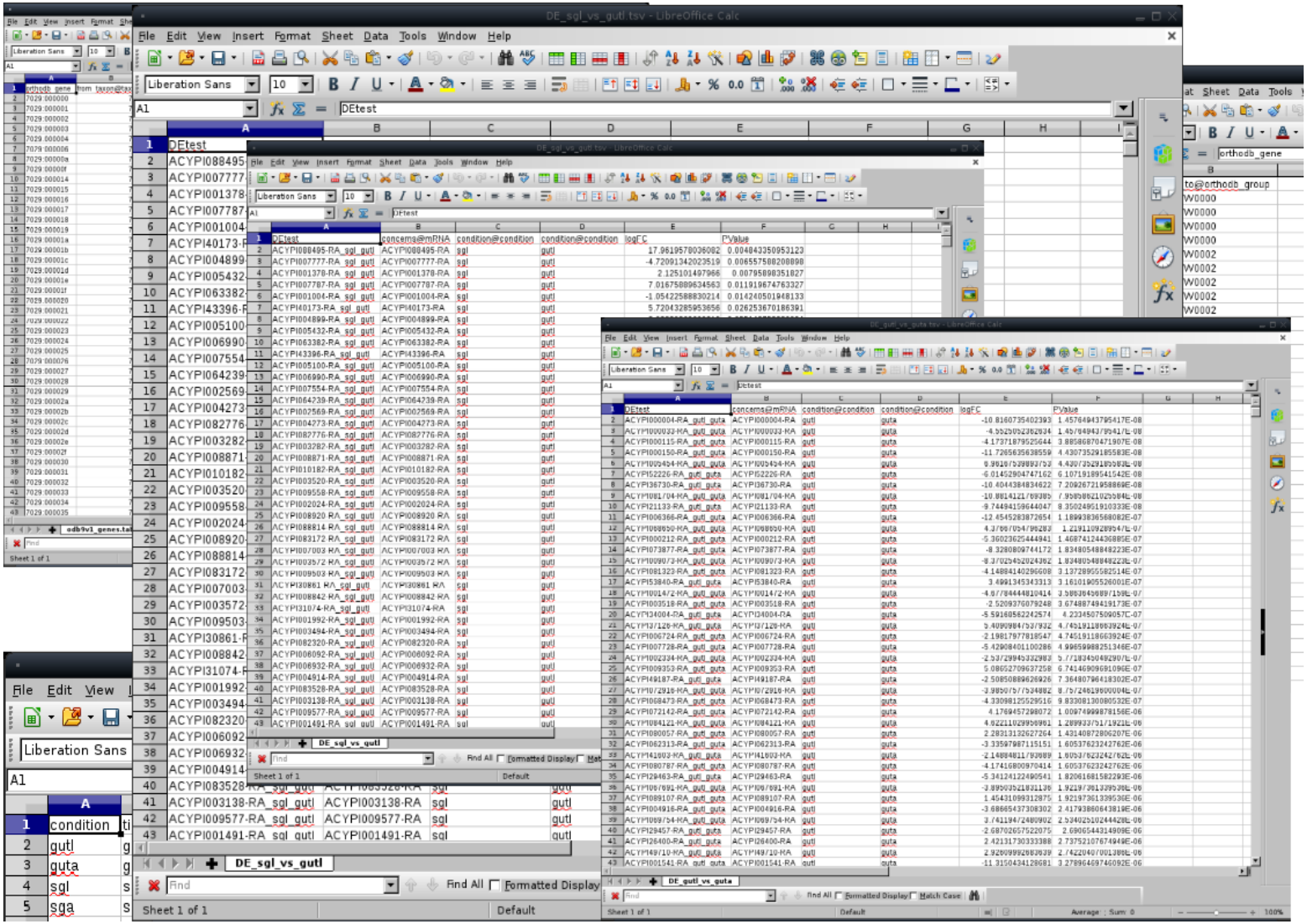
Outline

1. Integrating and querying data
   - Why it is difficult
   - Why RDF and SPARQL are relevant

2. What AskOmics is useful for
   - Integrating data
   - Querying data

Integrating and querying data

Data everywhere! (aka death by spreadsheet)



Death by spreadsheet: the worst is yet to come!

Data are distributed

Definition: Entity

Anything that can be identified (i.e. the things we can talk about)

- some informations about an entity in a repository
- other informations about the same entity in another repository

your question requires to combine entity descriptions from multiple datasets

Only possible if:

- Entities are identified
- Datasets use the same identifier to describe the same entity

Good luck with your spreadsheets! :-)

Data description involves hierarchies

- Data are precise
- Queries involve more general criteria

your question requires some reasoning (often based on hierarchies and ontologies) in order to reconcile the data and the criteria

Only possible if:

- Knowledge has been formalized (e.g. in ontologies)
- Query engines (more or less) gracefully handle
  - linking data and ontologies
  - reasoning on the ontologies

Requirements

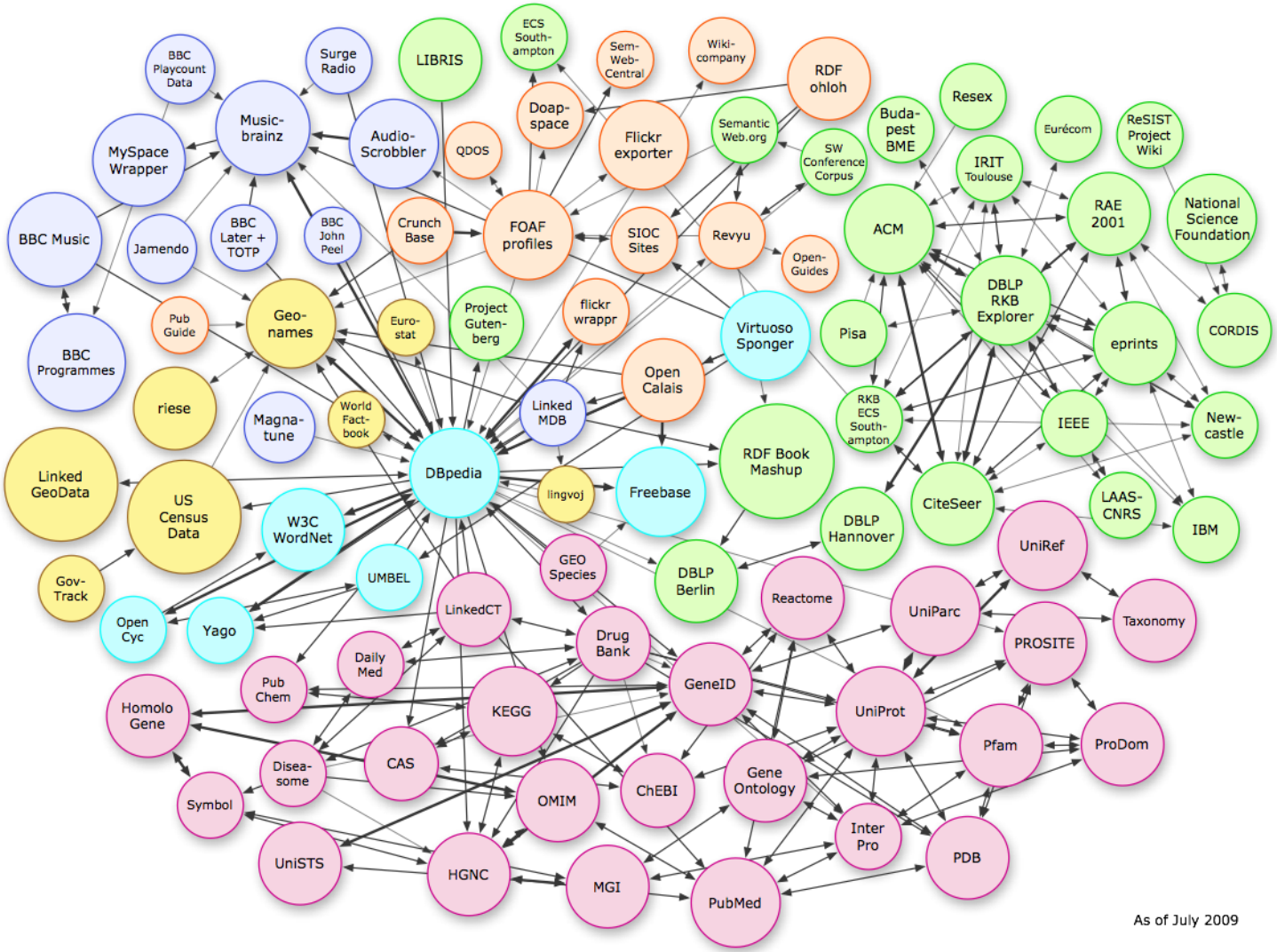
- Identify entities (uniformly) so that multiple repositories use the same identifier when they refer to the same entity
- Describe entities
  - their characteristics
  - their relations with other entities
  - both can be scattered in multiple repositories
- Query descriptions even if scattered in multiple repositories
- Perform reasoning (based on domain knowledge) over these descriptions

These points exceed the "classical" relational model's capabilities

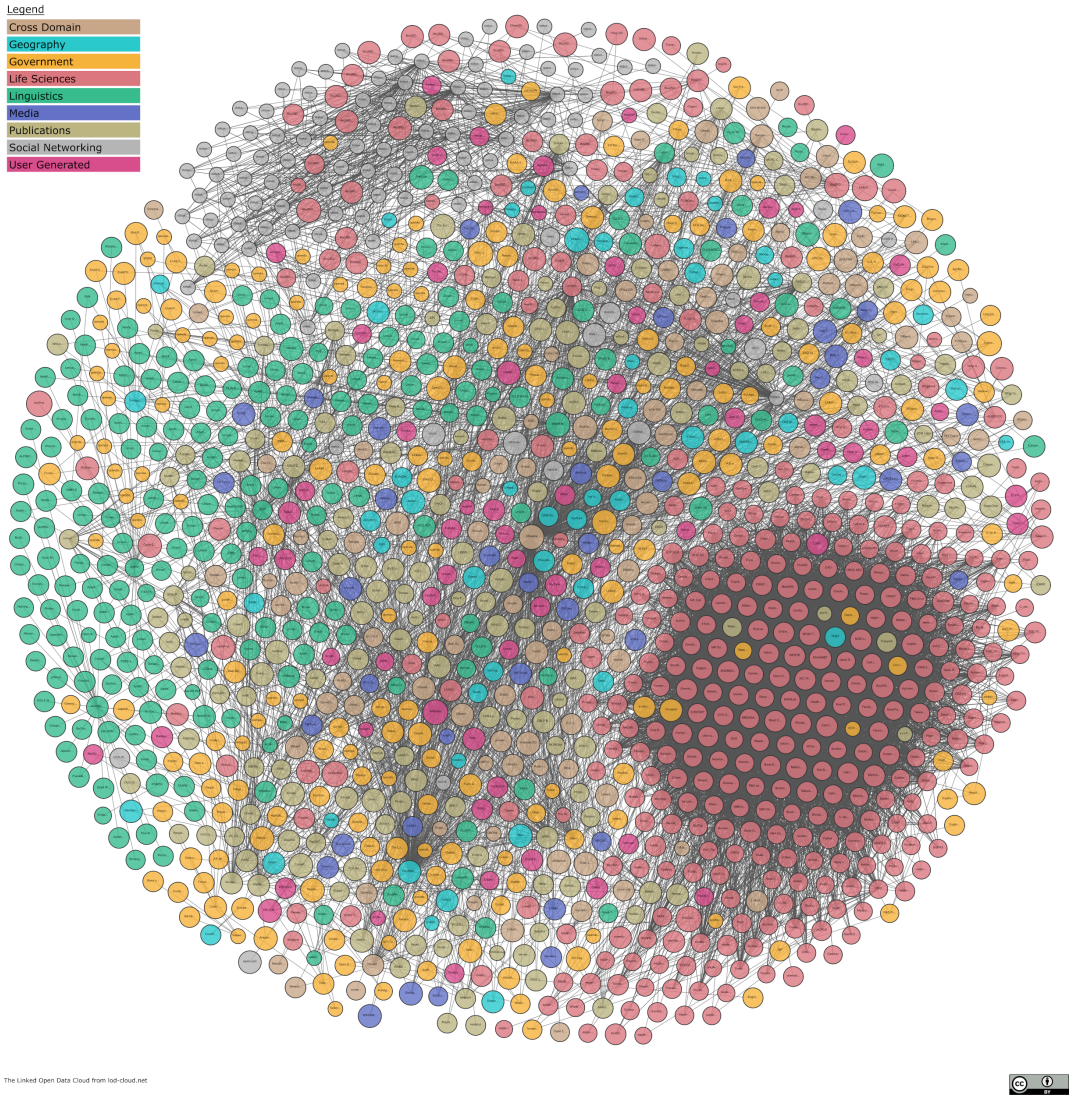
The good news (1/3)

Semantic Web technologies (RDF, SPARQL, OWL) address all of these

Linked Open Data (in 2009)



Linked Open Data (in 2023)



Linked Open Data

Semantic Web technologies
April 2016: (according to `http://stats.lod2.eu`)

- 149.10^9 triples
- distributed over 9960 knowledge graphs

A treasure at your fingertips (or a nightmare of spreadsheets)

Linked data are here... but still have to be adopted by end users

"Real" users

- do not contribute (yet) their data to the LOD cloud
- do not use the LOD cloud for analyzing their own data (yet)

RDF principles

Entities and relations are identified by their URI
RDF dataset = directed graph of triples

- **subject:** the entity we are talking about
- **predicate:** the relation used to describe the entity
- **object:** (one of the) relation's value for the subject

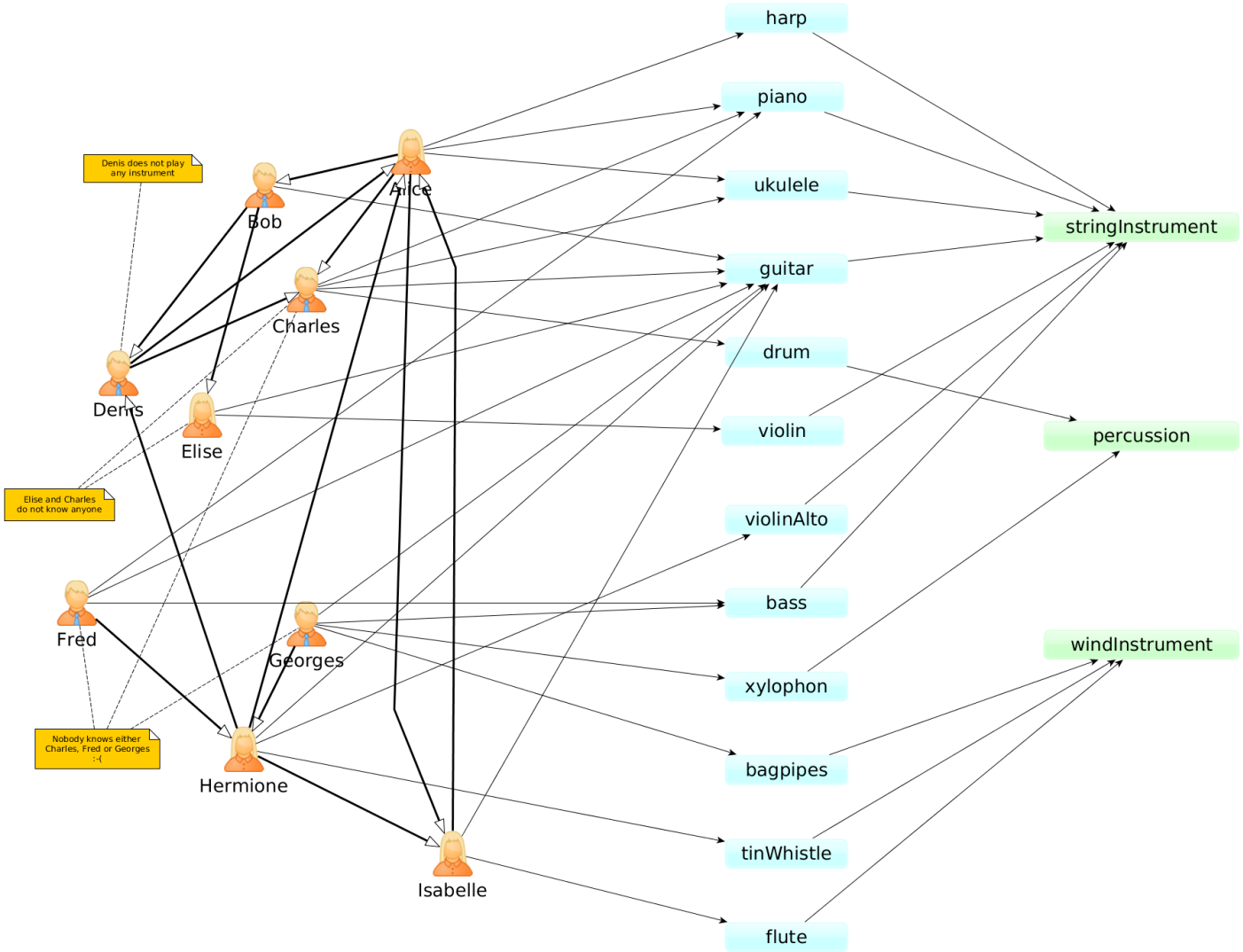
Example:

```
Alice plays violin
Alice plays guitar
Alice knows Bob
```

The good news (2/3)

Don't worry about RDF, AskOmics generates it from your csv

RDF as a directed graph



SPARQL principles

SPARQL = query language similar to SQL
Variable names start with a question mark
What instruments does Alice play?

```
SELECT ?instr
WHERE {
  p1:Alice mus:plays ?instr .
}
```

Returns:

```
violin
guitar
```

The good news (3/3)

Don't worry about SPARQL queries, AskOmics generates them for you (and runs them as well)

What AskOmics is useful for

AskOmics

Integrating data

- Import your data files
  - CSV or TSV
  - RDF
  - GFF
- Import public knowledge bases (GO, Reactome, NCBI taxon,...)
- Declare (remote) SPARQL endpoints (in progress)

Querying data

AskOmics

Integrating data

Querying data

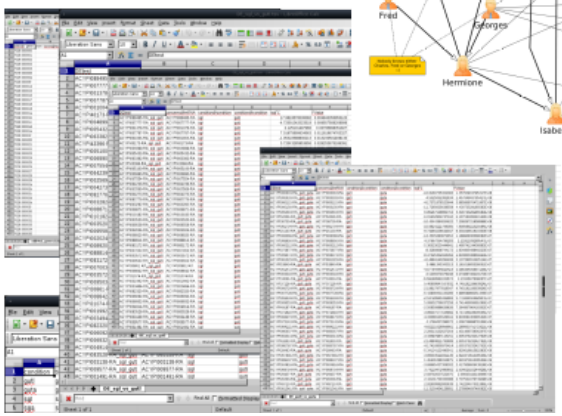
- Graph-based user-friendly SPARQL query composer
  - based on an abstraction of your data
    - depends on the data structure (small)
    - not on the data themselves (possibly huge)
  - can be simplified (hide a portion of the graph)
  - can be enriched (shortcuts and virtual links)
  - modular design (select/deselect datasets)
- Span multiple SPARQL endpoints (in progress)
- You do not have to see the SPARQL code
- Save the query result (obviously)
- Save or import SPARQL query

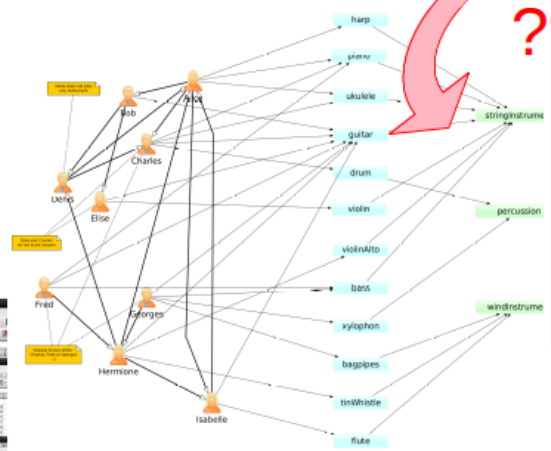
AskOmics principles

Abstraction (small)
facilitating query design
+ automagic SPARQL generation

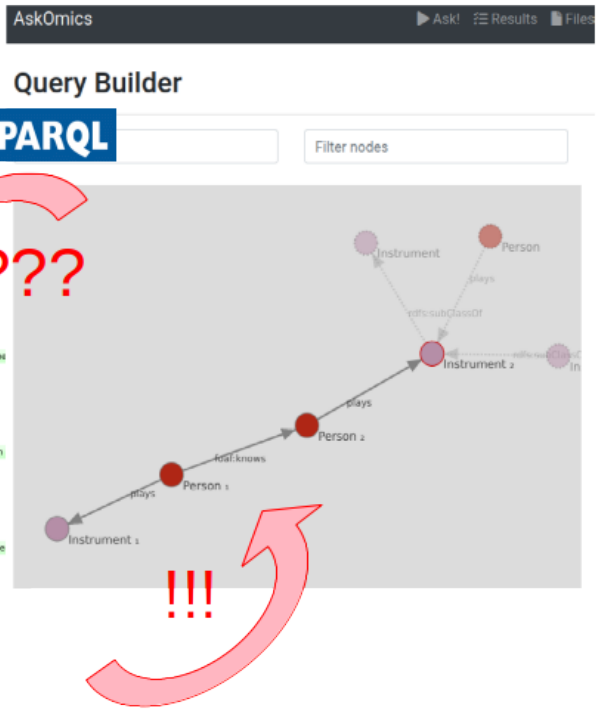
RDF
Knowledge graph
(large)

raw data
(large)





???



!!!

Acknowledgments

- Meziane Aite
- Arnaud Belcour
- Charles Bettembourg
- Matéo Boudet
- Anthony Bretaudeau
- Yvanne Chaussin
- Aurélie Évrard
- Xavier Garnier
- Maël Kerbiriou
- Colleagues from Nantes for their insight on federated queries
  - Pascal Molli
  - Patricia Serrano Alvarado
  - Hala Skaf
- Sylvaine Bitteur (INRA / Agrocampus Ouest) for the logo