Analysing Knowledge Domains that Emerge from Linked Open Data

Luigi Asprino[1,2], Paolo Ciancarini[2], Valentina Presutti[1]

1. STLab, ISTC-CNR, Rome , Italy
2. Dept. of Computer Science, University of Bologna, Italy

Garr 2019

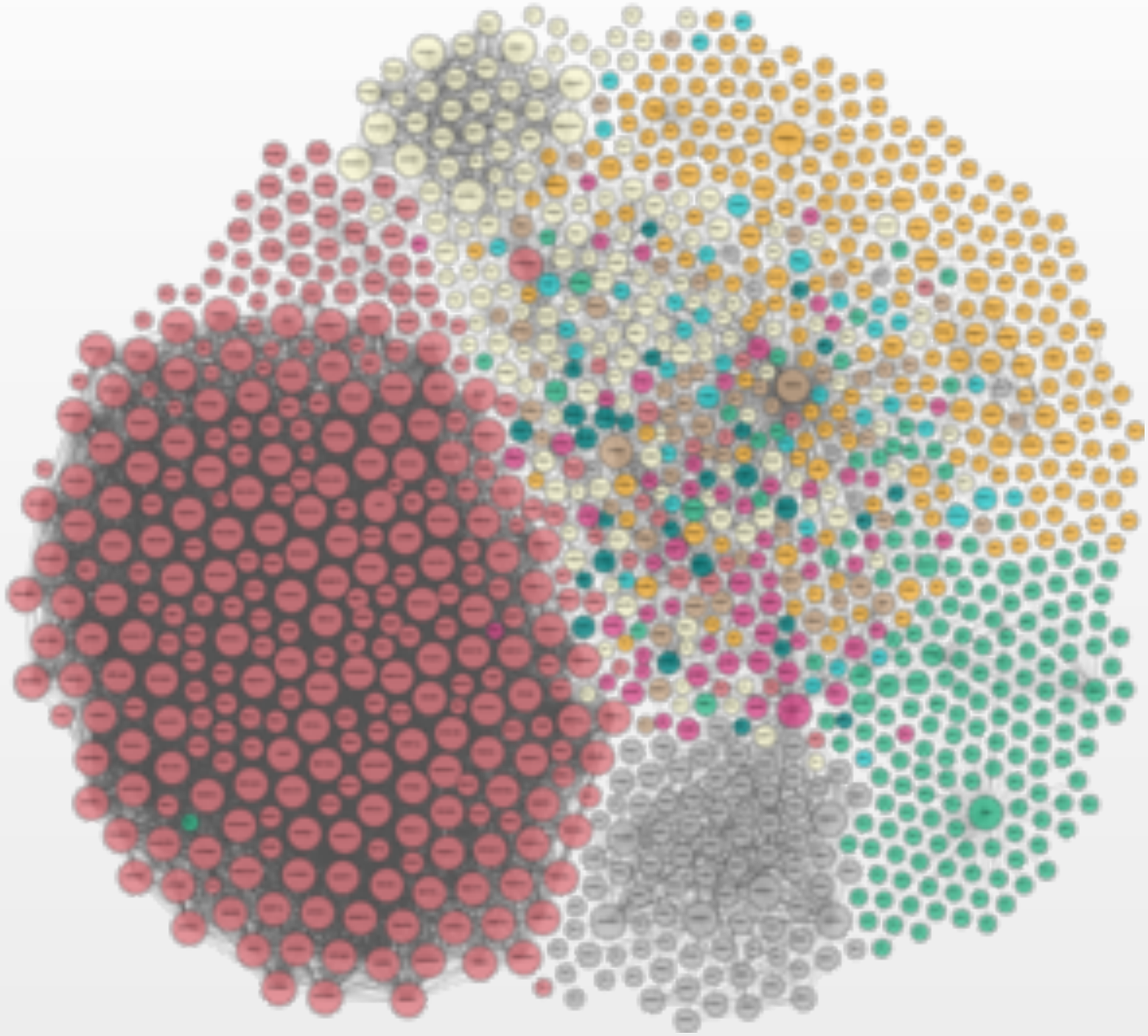
Torino, 4-6 Giugno 2019

Semantic Web and Linked Open Data

- Semantic Web: Initiative aimed at promoting standards (languages and protocols) for exchanging data through the Web

- Linked Open Data: **linked datasets** shared using Semantic Web standards with an **open** license

Linked Open Data Cloud



Huge more than 200B facts

Uniformly specified in RDF and OWL

Collaboratively built

Open

Miscellaneous

Empirical Analyses of Knowledge

- LOD gives a unique opportunity to study how knowledge behave at very large scale.
- Empirical Knowledge Representation:

➡ How KR's principles are applied in the wild?

➡ How languages are actually used?

➡ Quality of datasets' interlinking

➡ Common patterns in KR

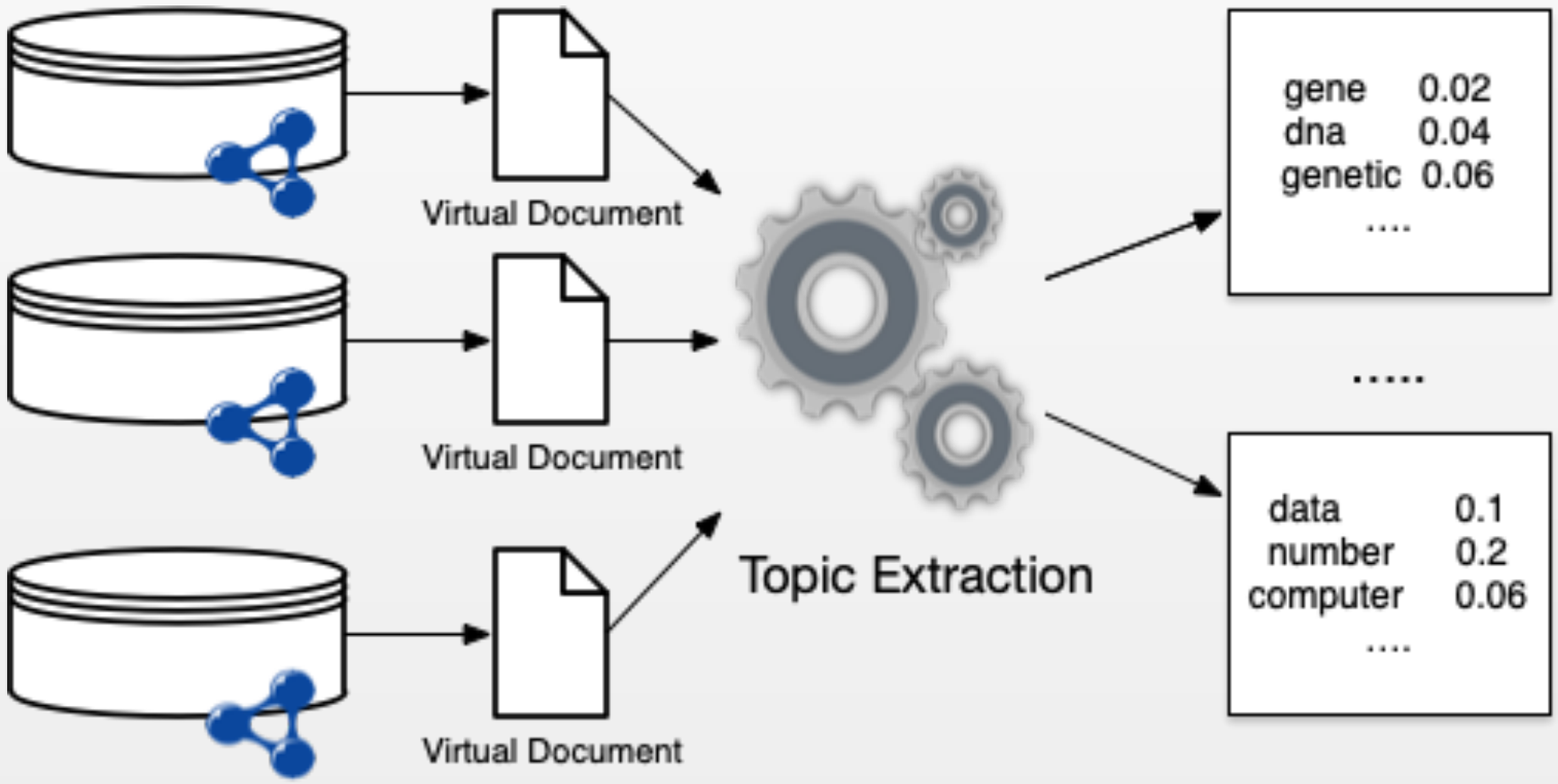
➡ …

Analyses by Knowledge Domain

- Each knowledge domain has its peculiarities
  - ➡ e.g. Biology, Medical
- How principles, languages, best practices are applied in each knowledge domain (e.g. (Schmachtenberg 2014 et al.)
- Most of these analyses datasets' metadata
  - ➡ Poorly represent knowledge domains
  - ➡ ~25% of datasets do not declare their domain
  - ➡ Single label for dataset
  - ➡ Datasets are labeled in a top-down approach

Proposed Approach

- Bottom-up approach: make domain emerge from data



Status of the work

- LOD Crawl: LOD-laundromat (28B triples ~550GB) http://lodlaundromat.org/
- m1.xxl instance (64 GB RAM, 16 vCPUs)
- 11,5 hr for extracting corpus openly available at https://tinyurl.com/y4qw7rpg (~20GB)

What's next

- Extracting topics using gensim