# Linked Data and facets to explore text corpora in the Humanities: a case study

Christian Morbidoni

Semedia, Department of Information Engineering (DII), Università Politecnica delle Marche, Ancona, Italy

**Abstract.** Faceted search and browsing is an intuitive and powerful way of traversing a structured knowledge base and has been applied with success in many contexts. The GramsciSource project is currently investigating how faceted navigation and Linked Data can be combined to help Humanities scholars in working with digital text corpora. In this short paper we focus on the "Quaderni dal carcere" by Antonio Gramsci, one of the most popular Italian philosophers and politicians, we present our ongoing work and we discuss our approach. This consists of first building a RDF graph to encode different "levels" of knowledge about the texts and then extracting relevant graph paths to be used as navigation facets. We then built a first prototype exploration tool with a two-fold objective: a) allow non experts to make sense of the extremely fragmented and multidisciplinary text corpus, and b) allow Gramsci scholars to easily select a subset of the corpus of interest as well as possibly discovering new insights or answer research questions.

**Keywords:** Faceted browsing, Digital Humanities, Entity Extraction

## 1 The data

Gramsci's "Quaderni dal carcere" is an extremely fragmented corpus composed of more that 4,000 "notes' organized in 29 books (quaderni). Notes vary in length from single lines to several pages and span different domains, from sociology and politics to literature. They are available in the GramsciSource digital library[1] (created in the frame of the project) with stable URLs. We built a Linked Data graph by merging structured knowledge coming from different sources: the Linked Data Gramsci Dictionary, data coming from DBpedia Italia[2] and semantic annotations made with Pundit [1][3].

The **Linked Data Gramsci Dictionary**, is a dataset extracted from the Gramsci Dictionary [2], a recognized scholarly contribution within the international Gramsci scholarly community, which includes all the most important topics in the Gramsci's thought. Each topic in the dictionary is documented by a

[1] The DL is currently officially offline due to maintenance, but can be reached at the following address: http://89.31.77.216

[2] http://it.dbpedia.org

[3] http://thepund.it

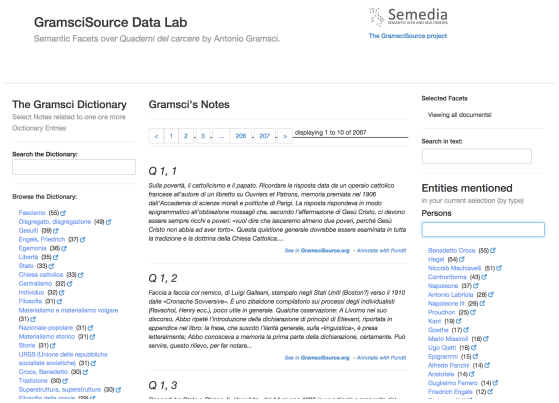


**Fig. 1.** A screenshot of the prototype

text with references to specific Gramsci's Notes from the Quaderni dal Carcere. We automatically processed such citations using regular expressions and producing RDF triples representing such connections, expressing relations among single notes and a number of dictionary topics they are relevant to.

**Entity extraction and linking.** Several approaches and tools for extracting and disambiguating relevant entities mentioned in a text appeared in the last years. Among them, DataTXT[4] is, to our knowledge, one of the best tools supporting Italian language. DataTXT derives from previous academic research [3], makes use of Wikipedia to disambiguate matched entities and to link them to the Italian DBPedia and proved to be highly performant even on very short texts. Running such a entity extraction tool on all the notes resulted in over 2,038 notes (50%of the total number) annotated with at least one entity and a total of 43,000 entities matched. After a manual revision of the results we removed around 30 entities that were clearly wrong matches. We then inspected 80 random notes (2% of the total number of notes) and measured an accuracy of around 85%. Extracted entities span 144 different entity rdf:types and 5,876 distinct dc:types (which can be considered as entities categories). A more accurate evaluation of the results a well as a better tuning of the tool are goals for the next stage of the project.

**Scholars annotations.** Pundit[5] is a semantic web tool that enables users to produce machine readable data in the form of RDF by annotating web pages. Annotations from a single scholar are collected in so called "notebooks", which can be private or public. For the purpose of our proof of concept we created a set of sample annotations by manually linking texts to DBpedia and Freebase entities. At data representation level, such annotations are equivalent to those produced by DataTXT, once imported they are naturally captured by the facet queries (discussed in the next section).

[4] https://dandelion.eu/products/datatxt/
[5] http://thepund.it

## 2 Faceted search prototype

Existing approaches to identify relevant facets to browse a RDF graph based on quantitative measures such as predicate frequency, balance and objects cardinality [4]. This kind of approaches do not account for informative content of a facet and only consider facets derived from a set of triples with the same predicate. In the general case, however, relevant facets could be derived from more complex paths in the graph. Approaches to automatic facets extraction in such a general case have been recently proposed [6] and we plan to investigate their applicability in the near future.

Our simple approach is to derive facets from SPARQL queries of the form:

```
select distinct ?url ?facet ?value where { CUSTOM_QUERY }
```

Where ?url is a resource of interest (notes in our case), ?facet is a facet name and ?value is a possible value of such a facet. Such a simple approach is also quite flexible and allows, for example, to easily turn all the datatype properties of a resource to facets, e.g with the following query:

```
select distinct ?uri ?facet ?value where {
?uri rdf:type gramsci:Note. ?uri ?facet ?facet.}
```

Deriving facets from SPARQL queries is an approach already explored in literature [5]. For the purpose of our proof of concept we chose candidate graph paths by inspecting the data and accordingly to scholars preferences. The facets we implemented in our prototype are:

- Gramsci Dictionary topics. This facet lists all the dictionary topics where a note is referenced;
- DBpedia entities. A set of facets where entities mentioned in a note are grouped according to their rdf:type. Relevant rdf:types individuated are Persons, Books, Languages, Places and Events, but they could be more specific (e.g. Politicians, Artists, Magazines, etc.);
- Categories. A facet listing all the dc:types associated to entities mentioned in a note;
- Scholars Notebooks. This facet lists all the scholars (Pundit users) who manually annotated a note.

To enable navigation of the corpus along the different "dimensions", we implemented a faceted browser based on Apache Solr[6]. Solr, along with its Ajax-Solr[7] frontend provides a relatively easy way to build a performant faceted browser on top of Lucene. We built the solr index by running the SPARQL queries (described in the previous section) and using results associated to the ?uri variable as document ID, ?facet as index field and ?value as field values. The prototype is available at http://purl.org/gramscisource/quaderni.

[6] http://lucene.apache.org/solr/
[7] http://github.com/evolvingweb/ajax-solr/wiki

Some usage patterns have been individuated by scholars involved in the project: a) Using the Dictionary facet to intersect two or more topics from the vocabulary. This is a simple but useful "advanced search" feature; b) Choose one or more Dictionary topics (e.g. Storia), then use the facets on the right (DBpedia entities) to provide additional context (e.g. Hegel, Croce and Plechanov are the main persons related to History, "Teoria e storia della storiografia" and "Misre de la philosophie" are two related books, etc.); c) Start from a full text search or from a DBPedia entity (e.g. "Conte di Montecristo") and discover related topics.

## 3 Conclusions and Acknowledgements

In this short paper we discussed preliminary results in leveraging Linked Data in the GramsciSource project and we presented a proof of concept prototype. Feedback from Humanities scholars involved in the project (and in related projects, such as DM2E[8]) was positive and encouraged us to move further. End user evaluation will be run in the next months. We are currently evaluating automatic methods to derive entities and facets (e.g. based on language analysis tool such as [7]), with the aim of making the approach easily applicable to different texts corpora.

This work is supported by the GramsciSource project funded by the Italian Ministry of Education under the FIRB action.

## References

1. Marco Grassi, Christian Morbidoni, Michele Nucci, Simone Fonda and Francesco Piazza. Pundit: Augmenting Web Contents with Semantics. Literary & Linguisting Computing, 2013
2. Dizionario gramsciano 1926-1937, Curated by Guido Liguori, Pasquale Voza, Roma, Carocci Editore, 2009, pp. 918.
3. Paolo Ferragina, Ugo Scaiella, TAGME: on-the-fly annotation of short text fragments (by wikipedia entities), Proceedings of the 19th ACM international conference on Information and knowledge management, New York, 2010
4. Eyal Oren, Renaud Delbru, Stefan Decker, Extending Faceted Navigation for RDF Data, The Semantic Web - ISWC 2006, Lecture Notes in Computer Science Volume 4273, 2006, pp 559-572.
5. Philipp Heim, Jrgen Ziegler, Faceted Visual Exploration of Semantic Data, Human Aspects of Visualization, Lecture Notes in Computer Science Volume 6431, 2011, pp 58-75.
6. Bei Xu, Hai Zhuge, Automatic Faceted Navigation, Future Generation Computer Systems archive, Volume 32, March, 2014, Pages 187-197
7. Dell'Orletta F., Venturi G., Cimino A., Montemagni S. (2014) T2K: a System for Automatically Extracting and Organizing Knowledge from Texts. In Proceedings of 9th Edition of International Conference on Language Resources and Evaluation (LREC 2014), 26-31 May, Reykjavik, Iceland.

[8] http://dm2e.eu