# LOTED: Exploiting Linked Data in Analyzing European Procurement Notices

Francesco Valle[1], Mathieu d'Aquin[2], Tommaso Di Noia[1] and Enrico Motta[2]

1. Technical University of Bari, Electrical and Electronics Engineering Department
Information Systems Research Group
francescovalle84@gmail.com, t.dinoia@poliba.it
2. Knowledge Media Institute, The Open University, Milton Keynes, UK
{m.daquin, e.motta}@open.ac.uk

**Abstract.** The world of procurements and eProcurement generates daily large amounts of data, that represent knowledge of great economical value both for individual companies and for public organisations wishing to achieve a better understanding of a given market. However, such data remains difficult to explore and analyze as it is being kept isolated from other sources of knowledge, in dedicated systems. In this paper, we present an ongoing work on extracting and linking data from the European 'Tenders Electronic Daily' system, which publishes approximately 1,500 tenders five times a week. We specifically show how such information is dynamically extracted and linked to external datasets, and how the created links enrich the original data, introducing new perspectives to its analysis. We show tools we developed to support such 'linked data-based' analysis of data, and report on the lessons learnt from our experience in building a linked data application with potential for real-life use in knowledge extraction.

## 1 Introduction

The Tenders Electronic Daily (TED[1]) website is a portal maintained by the European Commission and dedicated to the public procurement in the countries of the European Union. As the name suggests, it is updated daily (5 days a week) with newly published tenders in 14 different sectors (e.g., Education, Technology and Equipment, Agriculture and Food). Each tender is an official document, containing information related to the public organisation it originates from (e.g., a town council, a public administration), its location, the type of activity it is related to, etc.

As such, this portal represents a rich source of information from which crucial strategic and economical knowledge could be extracted. Services exist that provide mail alerts and other mechanisms for companies on the basis of TED, but these tend to simply redirect the information from the original system to the user, without making any further analysis.

In this paper, we demonstrate how the principles of linked data can be used on the information exposed by the TED system, on the one hand to improve access

[1] `http://ted.europa.eu/`

to this data, allowing new applications to be built on top of it, and on the other hand, to investigate how links to external datasets can bring valuable additional perspectives into the data, enriching it with new dimensions and making possible new forms of analysis that exploit data and links to reach a better understanding of the global public market. In addition, we derive from our experience in building such a concrete linked data application lessons regarding the current limitations of linked data and of the supporting tools.

In the next section, we detail how we developed a platform realising a workflow from obtaining the data in the original TED portal, to exposing it as linked data and providing interfaces to it. In Section 3, we detail a prototype application of this platform to build visualisations combining dimensions from the original data and from external datasets, demonstrating how such an approach can provide endless possibilities for new perspectives on previously isolated data. Section 4 reports on our concrete experience in building such an application, showing in particular how additional work should be realised in the area of linked data to make applications such as ours easier to build and more efficient. Finally, we conclude the paper pointing out directions for future work in Section 5.

## 2 LOTED: Anatomy of a Linked Data Application

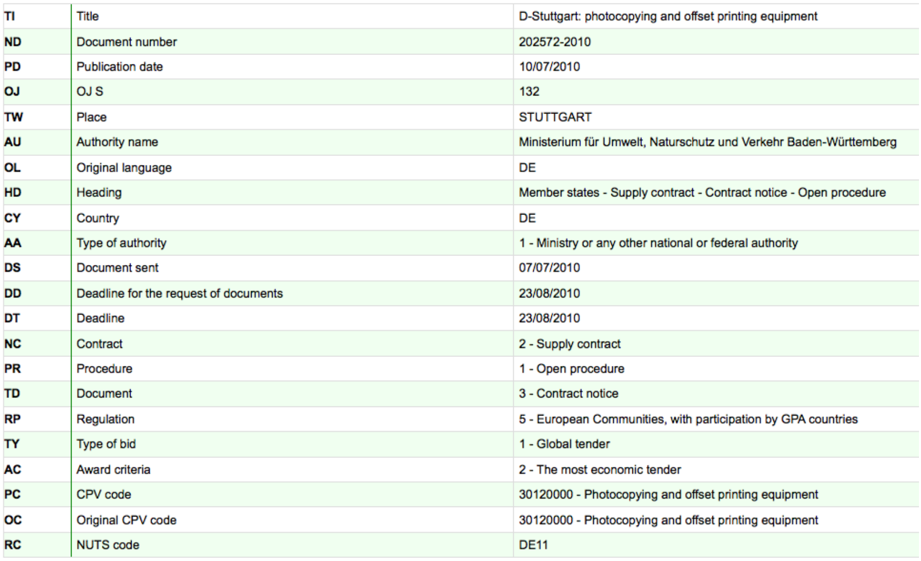
The TED portal updates its subscriber on newly published tenders daily through a set of RSS feeds, with an RSS feed for each combination of country (27 countries in total) and sector (14 sectors in total). Each RSS feed is updated at most once a day, generally five days a week, and is every time entirely replaced by the new tenders. For example, `feed://ted.europa.eu/TED/rss/en/RSS_tran_UK.xml` is the current URL of the RSS feed to the sector "Transport and related services" in United Kingdom.

A tender on TED is presented by a document, available in its original languages and in translated form. For example, `http://ted.europa.eu/udl?uri=TED:NOTICE:202572-2010:TEXT:EN:HTML&tabId=1` is a tender document (Contract Notice), for photocopying and offset printing equipment in Stuttgart, Germany. Such a document contains general, common information about the tender such as its type, requested products or services, the location, originating organisation, award criteria, etc. A summary of this data is available for each document, presented in a tabular format (see Figure 1) with normalised fields and values for these fields (for the previous tender, see `http://ted.europa.eu/udl?uri=TED:NOTICE:202572-2010:DATA:EN:HTML`). The availability of such a semi-structured summary of the document greatly facilitates the task of extracting data from the TED system, as shown in the following sections.

### 2.1 Overview

In this section, we give an overview of the platform for the publication and use of a linked data version of the information provided by the TED system, which we called LOTED[2] (Linked-Open Tenders Electronic Daily).

[2] `http://loted.eu`



| | | |
|---|---|---|
| TI | Title | D-Stuttgart: photocopying and offset printing equipment |
| ND | Document number | 202572-2010 |
| PD | Publication date | 10/07/2010 |
| OJ | OJ S | 132 |
| TW | Place | STUTTGART |
| AU | Authority name | Ministerium für Umwelt, Naturschutz und Verkehr Baden-Württemberg |
| OL | Original language | DE |
| HD | Heading | Member states - Supply contract - Contract notice - Open procedure |
| CY | Country | DE |
| AA | Type of authority | 1 - Ministry or any other national or federal authority |
| DS | Document sent | 07/07/2010 |
| DD | Deadline for the request of documents | 23/08/2010 |
| DT | Deadline | 23/08/2010 |
| NC | Contract | 2 - Supply contract |
| PR | Procedure | 1 - Open procedure |
| TD | Document | 3 - Contract notice |
| RP | Regulation | 5 - European Communities, with participation by GPA countries |
| TY | Type of bid | 1 - Global tender |
| AC | Award criteria | 2 - The most economic tender |
| PC | CPV code | 30120000 - Photocopying and offset printing equipment |
| OC | Original CPV code | 30120000 - Photocopying and offset printing equipment |
| RC | NUTS code | DE11 |

**Fig. 1.** Example of tabular summary of a tender on the TED portal.

LOTED essentially relies on a triple store[3] which is being updated daily with information extracted as linked data from the RSS feeds of the TED system, and exposed through a SPARQL endpoint (see Figure 2). The way RDF data is extracted from the original tender documents, and how such data is linked to external datasets (currently geonames[4] and DBpedia[5]), is explained in the next section.

As can be seen from Figure 2, at the heart of the system is the *LOTED Ontology*[6], which has been specifically developed for the needs of the platform. It is a lightweight ontology, that matches directly the semi-structured representation of the tenders from the TED system, while introducing an additional level of structure. It is worth also noticing that the labels in this ontology are available in three different languages. It can be argued that reusing existing ontologies would have better encouraged interoperability. However, we found that extracting existing information was made easier and less error-prone if realised in a target structure that matched the original data closely. Mapping and integrating this ontology with others, as well as evolving it towards a more expressive modelisation of the domain of procurement is planned as part of our future work.

[3] After a few tests, we found that the Jena system (`http://openjena.org/`) with a TDB persistent store (`http://openjena.org/TDB`) offered the best compromise between flexibility, robustness and performance for our scenario.

[4] `http://www.geonames.org/`

[5] `http://dbpedia.org`

[6] `http://loted.eu/ontology`

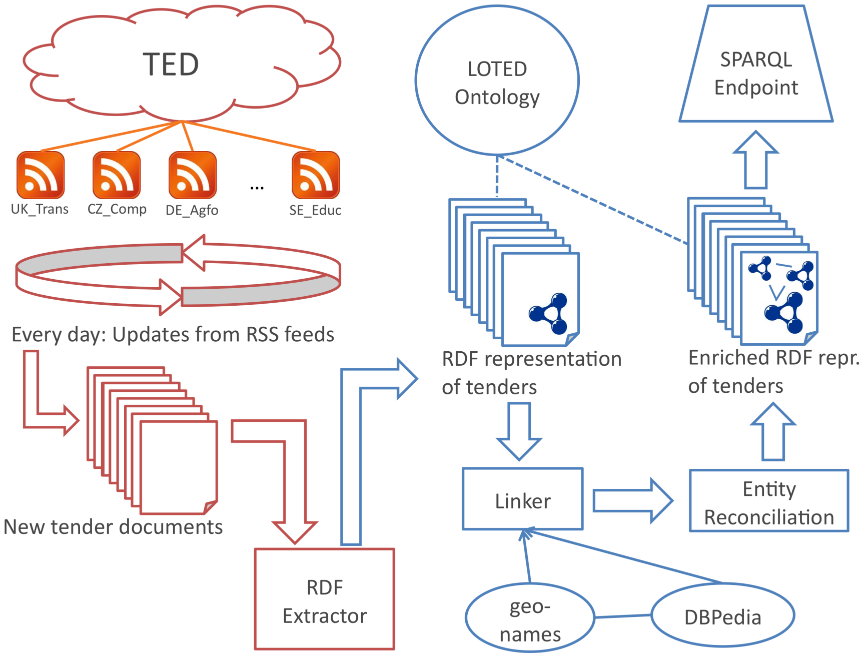


**Fig. 2.** Workflow for the daily update of linked data from the TED system to the LOTED endpoint. RDF representations based on the LOTED ontology are extracted from the TED RSS feeds, and enriched through automatically discovered links to geonames and DBpedia.

## 2.2 Extracting Information and Creating Links

The extraction of structured information is realised as a scheduled task, happening every day. It starts by checking whether any of the RSS feeds from the TED system had changed and downloading the English, tabular version of any new tender published on the portal. As explained above, extraction from these documents is facilitated on the one hand by the fact that they are formatted in a semi-structured way, and on the other hand, as the LOTED ontology has been explicitly designed to match this structure. We therefore developed a custom made RDF extractor which parses the table structure of the original document and transforms it into a structured RDF representation. Amongst the difficulties that are common to such processes is the issue of having to deal with special characters and unicode strings that have to be included in URIs of entities. We dealt with this problem by replacing any of such characters by its equivalent HTML entity.

Creating links is obviously a more challenging issue. Geographical information is well covered by available linked datasets and can provide useful information to associate to the tender documents, including the data about the places where the originating organisation is. Within the data, the information available is a string representing the name of the city where the organisation is located, and a two letter country code. Geonames is a data set of geographical places around the world. It provides simple identifiers for places, and information about their locations (coordinates) and their names in various languages. Coordinates can be very useful, as we will see later, simply to be able to place the city on a map. Another large source of information about cities is DBpedia. Indeed, DBpedia being extracted from Wikipedia[7], it can contain a large variety of interesting data about a given city, from its population to the region it is in or the current mayor.

To link the places found in the tenders to geonames, we make use of the search engine provided on the geonames website[8]. Having both the name of the city and the country code (being in Europe) makes the query sufficiently unambiguous, so that the first result from the search engine is in the large majority of the cases the one we are looking for. Actually, we never found any error in this linking process in the time the system has been running (more than 2 months). Another advantage of relying on geonames is that it is already well connected to other datasets. In particular, DBpedia includes in most of the resources it contains about cities a *sameAs* link to the corresponding URI identifying the place in geonames. Therefore, finding links to DBpedia is realised straightforwardly through a SPARQL query requesting resources that are the same as the discovered geonames objects.

When building a linked data based platform relying on links to external datasets, different choices can be made on the way to integrate and use these links. Indeed, a natural choice would be that any query sent to the system would automatically look up for any link involved and retrieve dynamically, at run-time, the corresponding information to be included in the results. However, while this appears to be the kind of process linked data applications would normally rely on, the tool support for realising it appears to be very weak. Only a few existing systems are currently able to realise live look-ups of external entities [1, 2], under specific conditions. In addition, these systems tend to ignore the links such as *sameAs* which, while having well defined semantics, are considered in the same way as any other relation by SPARQL engines.

For this reason, in LOTED, we made the choice of including in the workflow an offline step of *entity reconciliation* (also called materialisation in [1]). The basic idea for this step is to aggregate locally, under one identifier, all the information from entities related with each other by *sameAs*. In our case, we decided to use the geonames URI to represent the location of the organisation in a tender, retrieving any information related to this URI from the geonames system. We then retrieve the URI of entities in DBpedia linking to the geonames object, and import all the information obtained by resolving this URI as

[7] `http://www.wikipedia.org/`

[8] `http://www.geonames.org/search.html`

attached to the geonames URI. In this way, all the information about a given place, from both geonames and DBpedia, ends up being aggregated in our triple store, under the corresponding geonames URI.

The creation and linking process described above has been running for more than 2 months at the time of writing (since the 12th May 2010). It collected around 55,000 procurement notices unequally distributed in the 14 sectors and 22 countries. These tenders relate to 5,000 places that are being linked to geonames and DBpedia.

### 2.3 Access Interfaces

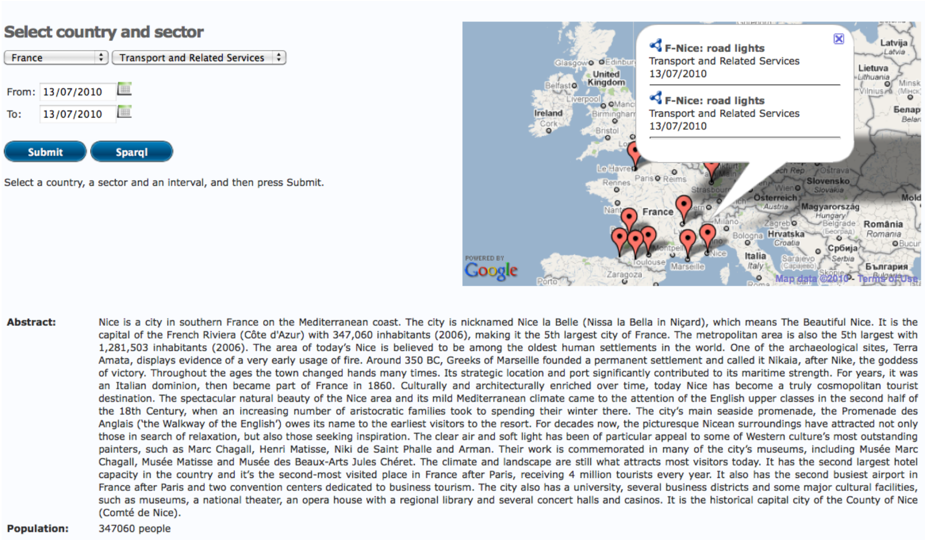
The primary goal of building a linked data platform such as LOTED is to make available data in a machine readable and connected way, so that this data can be further linked to and exploited in applications. Therefore, all the URIs used in tender descriptions in LOTED resolve, and provide a complete representation of the corresponding objects in RDF. For example, `http://loted.eu/data/tender/204339-2010` can be accessed to obtain the complete representation of the tender number *204339-2010*. Similarly, `http://loted.eu/data/authorityName/Ville_de_Nice` is the URI for the particular organisation (Nice City Council) that created the tender and the same pattern applies to other types of objects in the data. The same kind of information is also available through a SPARQL endpoint, located at `http://loted.eu/Sparql` and for which a basic interface is available linking from the front page of the LOTED portal.

The portal also implements a straightforward Web user interface, from which people can find, retrieve and obtain information about tenders (see Figure 3). A specific country, sector and range of dates can be chosen, that will make appear the selection of tenders on a map, focusing on the selecting country. In practice, this selection is translated into a SPARQL query to obtain the precise coordinates of the cities that are related to tenders in the given sector, country and range of dates. Clicking on the marker on the map corresponding to a given city makes appear the list of tenders available in this city for the given sector and dates, which are further linked to their original documents on the TED portal, and to their RDF representations. It also displays information about that city as described in DBpedia. This can in particular include data about existing companies in this location. The RDF description of the current selection of tenders can also be obtained, as well as the SPARQL query to generate this RDF description, which can be further edited by the user to obtain a more precise selection.

In the next section, we show how we can use the links to DBpedia and geonames to create the global visualisations of the tenders which are available under the *Charts* button of the interface.

## 3 Analyzing the Data: Visualization of 'Tender Profiles'

As a way to demonstrate how linked data can benefit the analysis of data, and obtaining a better understanding of a global domain where large data is involved,



Select country and sector

France | Transport and Related Services

From: 13/07/2010

To: 13/07/2010

Submit | Sparql

Select a country, a sector and an interval, and then press Submit.

F-Nice: road lights
Transport and Related Services
13/07/2010

F-Nice: road lights
Transport and Related Services
13/07/2010

**Abstract:** Nice is a city in southern France on the Mediterranean coast. The city is nicknamed Nice la Belle (Nissa la Bella in Niçard), which means The Beautiful Nice. It is the capital of the French Riviera (Côte d'Azur) with 347,060 inhabitants (2006), making it the 5th largest city of France. The metropolitan area is also the 5th largest with 1,281,503 inhabitants (2006). The area of today's Nice is believed to be among the oldest human settlements in the world. One of the archaeological sites, Terra Amata, displays evidence of a very early usage of fire. Around 350 BC, Greeks of Marseille founded a permanent settlement and called it Nikaia, after Nike, the goddess of victory. Throughout the ages the town changed hands many times. Its strategic location and port significantly contributed to its maritime strength. For years, it was an Italian dominion, then became part of France in 1860. Culturally and architecturally enriched over time, today Nice has become a truly cosmopolitan tourist destination. The spectacular natural beauty of the Nice area and its mild Mediterranean climate came to the attention of the English upper classes in the second half of the 18th Century, when an increasing number of aristocratic families took to spending their winter there. The city's main seaside promenade, the Promenade des Anglais ('the Walkway of the English') owes its name to the earliest visitors to the resort. For decades now, the picturesque Nicean surroundings have attracted not only those in search of relaxation, but also those seeking inspiration. The clear air and soft light has been of particular appeal to some of Western culture's most outstanding painters, such as Marc Chagall, Henri Matisse, Niki de Saint Phalle and Arman. Their work is commemorated in many of the city's museums, including Musée Marc Chagall, Musée Matisse and Musée des Beaux-Arts Jules Chéret. The climate and landscape are still what attracts most visitors today. It has the second largest hotel capacity in the country and it's the second-most visited place in France after Paris, receiving 4 million tourists every year. It also has the second busiest airport in France after Paris and two convention centers dedicated to business tourism. The city also has a university, several business districts and some major cultural facilities, such as museums, a national theater, an opera house with a regional library and several concert halls and casinos. It is the historical capital city of the County of Nice (Comté de Nice).

**Population:** 347060 people

**Fig. 3.** Web user interface of the LOTED portal.

we included in the LOTED portal visualisations of *tender profiles* according to various dimensions. The idea of a tender profile is that it corresponds to the proportion of each sector in terms of the number of tenders being published in a particular place or by a particular group of organisations.

To illustrate this idea, we consider the most straightforward of these charts, shown in Figure 4. This chart shows the tender profiles using the country of origin as the main clustering dimension over the entire period starting at the beginning of the LOTED system (12th May 2010). As can be seen, different countries tend to have similar profiles, with however some variations in the focus on some of the sectors. Therefore, from there, it is possible to focus on a specific sector, ordering automatically the different countries according to their contribution to this sector. Doing so, we can then realise that one of the greatest discrepancies between countries is on the sector of "financial and related services", with Belgium having a larger proportion of tenders in this area, and countries such as Malta and Slovakia being almost absent from such a market. In addition, a table is presented associated with the chart that shows the number of tenders for each country. This is useful to assess whether some countries have a sufficient amount of tenders for the results to be significant. For example, we can notice that Malta only has had 66 tenders during this period, while several other countries had thousands. For some reasons, France produces significantly more tenders than any other countries (more than 14,000, compared to 6,500 for the second country, Germany). Finally, it is also possible to manually order the country in the chart, to try to find correlations with other dimensions than the sector (e.g., ordering the countries from East to West, to see if any regularity appears).

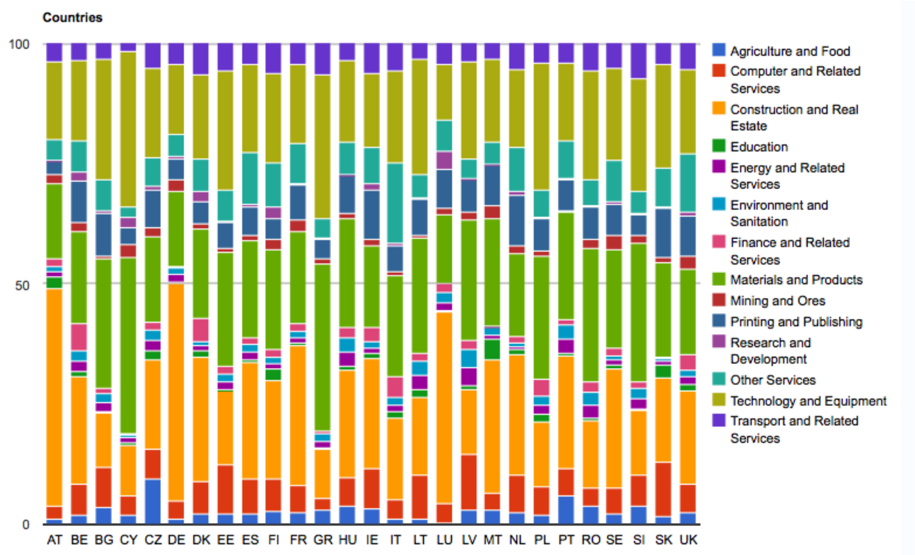


**Fig. 4.** Chart representing the tender profiles of the 22 countries in the LOTED system.

The chart relating tender profiles with country is very useful to obtain a general idea of the distribution of tenders across the entire set. However, what we want to show here is how the links to other datasets can provide additional, and more granular ways of analysis. Another chart included in the LOTED portal concerns specific regions within countries (see Figure 5 for the chart of the tender profiles by region in Italy). Indeed, for some countries, DBpedia provides the information on which sub-division a city is in. For Italy, we can identify 20 different regions with, like for countries, different amounts of tenders being published and different numbers of cities represented (this information is available in the table attached to the chart on the website).

In this case as well, tender profiles can be ordered by sector and manually to try to extract correlations between the regions and the sector in which they tend to publish tenders. This can be used to find the best place for different sectors, as the ones from which the most tenders originate, or the ones where a particular market has not developed yet. Using such a process, we can in particular devise the following table (Table 1) showing which region is the most and the least represented in each sector[9] (ignoring regions with less than 15 tenders):

Of course, one would need to be an expert in the economy of Italy to interpret these results appropriately. However, it appears unlikely for it to be a coincidence for example that Emilia-Romagna is first in both sectors of "education" and "research and development", while Umbria is last in these two, as

[9] In the spirit of http://www.informationisbeautiful.net/visualizations/because-every-country-is-the-best-at-something/

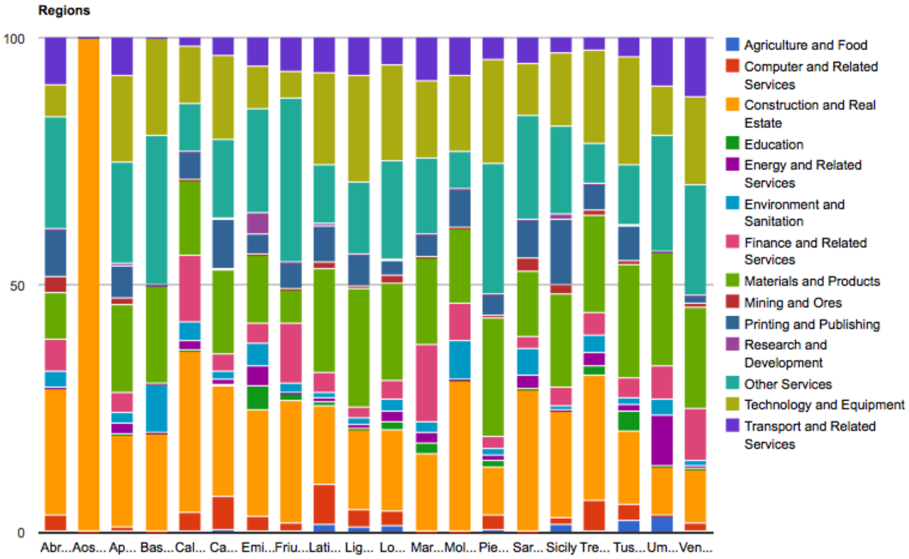


**Fig. 5.** Chart representing the tender profiles of 20 regions of Italy in the LOTED system.

**Table 1.** Most represented and least represented regions of Italy in each sector of tenders in LOTED.

| **Sector** | *First region* | *Last region* |
|---|---|---|
| Agriculture and Food | Umbria | Abruzzo |
| Computer and Related Services | Latium | Marche |
| Construction and Real Estate | Calabria | Piedmont |
| Education | Emilia-Romagna | Umbria |
| Energy and Related Services | Umbria | Sicily |
| Environment and Sanitation | Sardinia | Sicily |
| Finance and Related Services | Marche | Liguria |
| Materials and Products | Piedmont | Friuli-Venezia Giulia |
| Mining and Ores | Abruzzo | Friuli-Venezia Giulia |
| Printing and Publishing | Sicily | Umbria |
| Research and Development | Emilia-Romagna | Umbria |
| Other Services | Friuli-Venezia Giulia | Trentino-Alto Adige/Sdtirol |
| Technology and Equipment | Tuscany | Friuli-Venezia Giulia |

well as in “printing and publishing”.

To illustrate further how additional data brought into the initial dataset of tenders can be used to add originally unintended dimensions for data analysis, we also computed the chart of the tender profile in a given country, depending on the political affiliation of the cities the tender originates from (see Figure 6 for the chart for France). The political affiliation of a city can be found in DBpedia, either directly attached to the city (under the property `party`), or as a characteristic of the mayor (through the property `mayorParty`). Such information is however very heterogeneously present across countries and cities.

The basic idea here is that it would appear natural that different parties would have a different focus when it comes to public spending and that making emerge these different profiles can show users the influence of local politics. However, as can be seen from Figure 6, the tender profiles of the 2 major political parties in France are surprisingly similar. This is valid also for other countries where sufficient data can be obtained. This consistency within a country is even more surprising considering that, as shown previously, it is a lot less apparent across different countries.

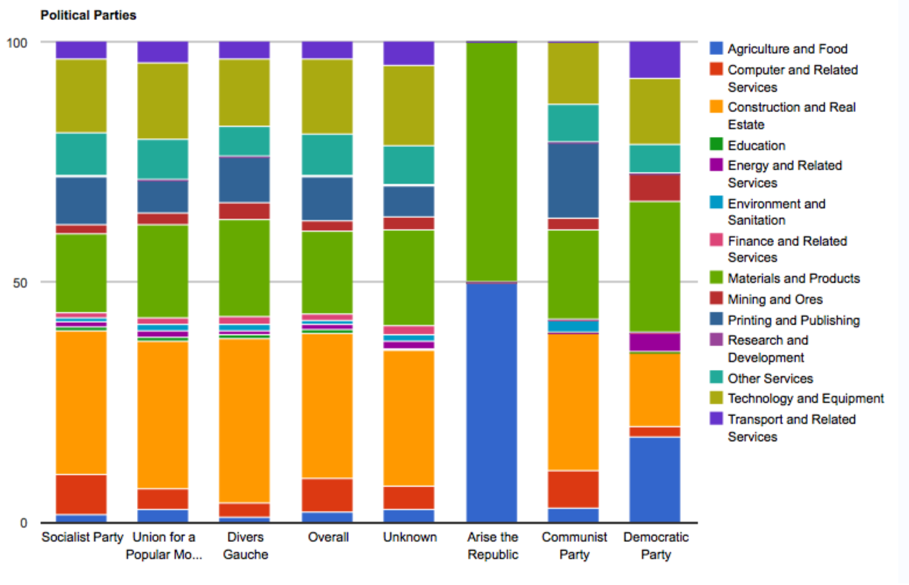


**Fig. 6.** Chart representing the tender profiles of political parties in France.

It is worth noticing here two additional columns that have been added to the chart: “Overall” and “Unknown”. Overall corresponds to the average tender profiles of all the cities for which the political party is known, while Unknown correspond to the ones for which the political party is not provided by DBpedia.

This allows us to measure the bias introduced by incomplete information, by looking at the dissimilarities between these two profiles.

## 4 Lessons Learnt

Generally, one of the most important lessons learnt from building an application such as LOTED is that there are obvious challenges in trying to *extract high level knowledge from interlinked datasets.* First of all, the incompleteness of the linked data cloud, and the general uncertainty regarding this incompleteness, appear clearly as major problems in analysing tender data using the visualisations presented above. Indeed, information as basic as the region in which a particular city is cannot be assumed to be always present. Heterogeneity also appears as a major issue, with many different properties used to represent the same information in DBpedia, as well as redundancies and variations that have to be manually cleaned up. Of course, it can be argued that such issues are specific to the considered dataset, DBpedia, and that geonames for example does not suffer from such issues. Indeed there seem to be two distinct sorts of linked datasets currently available, of which DBpedia and geonames seem to be the stereotypes: general purpose, heterogeneous, incomplete datasets, and focused, homogeneous and clean datasets. Having said that, ways for an application developer to realise in which category a dataset is and what can be expected from it seem crucially needed. This element appears especially critical in application such as LOTED which intend to provide some form of linked-data based data analysis, where the discrepancies in the representation of different resources can introduce a bias rendering the results of the analysis impossible to interpret.

At a lower level of granularity, our experience also made emerge the lack of support for the lifecycle of linked data applications. Indeed, we already discussed the need for us to realise the task of entity reconciliation offline and in an ad-hoc manner, due to the unavailability of tools to exploit links between datasets at run-time. This has obvious disadvantages, but also shows a need for generic frameworks to support common tasks in linked data applications, including data cleaning, URI creation (generating valid, consistent URIs from arbitrary unicode strings), data discovery, linking, link storage, link exploitation, etc. There have not yet been many applications exploiting linked data to a significant level. One of the reasons might be that, as we experienced, a lot of efforts need to be spent on setting up the basic underlying infrastructure, with every application starting almost from scratch.

The availability of such a generic framework for linked data applications, including reusable components implementing common tasks such as the ones listed above beyond the simple query engine/triple store, would then make possible the development of data analysis framework for linked data, able to find relevant links, and new dimensions to enrich an existing dataset, making possible the discovery of new knowledge from the created connections.

## 5 Conclusion

In this paper, we have presented the LOTED linked data application for European public procurement. There have not been many applications of linked data until now that where able to really exploit links to external dataset to provide additional functionalities. We can for example mention DBRec [3], a music recommender system exploiting DBpedia to help users in finding music, or simpler applications such as described for example in [4, 5]. While relatively simple, LOTED demonstrates how data analysis can be supported by linked data, including external, originally unintended perspectives into a dataset to potentially make emerge new high level knowledge about the considered domain. While this application and the general idea are still at an early stage, the obtained results have highlighted the new possibilities associated with the approach, showing the potential of being able to create data visualisations seamlessly combining dimensions from various datasets on the Web of Data.

We also report in this paper on how the current state of linked data and of the corresponding tools is hampering the development of such applications, leading to a need for a generic framework, a platform or toolkit to support the developers in realising the common, necessary tasks. Such a framework would allow applications such as LOTED to evolve from 'proofs-of-concept' of what the Web of data can help achieve, to concrete, real-life applications. In relation to this, we are currently exploring new analysis mechanisms on top of the data, extracting and explaining trends in the various sectors covered by LOTED tenders, in order to support the real needs of the potential users of the data, in relation with existing work in the area of business intelligence.

## Acknowledgement

The authors would like to thank Stefano Bertolo, Aldo Gangemi, Jose Manuel Gomez Perez and Jeff Pan for their ongoing contributions to the work on LOTED.

## References

1. Harth, A., Hose, K., Karnstedt, M., Polleres, A., Sattler, K.U., Umbrich, J.: Data summaries for on-demand queries over linked data. In: 19th International World Wide Web Conference (WWW2010). (April 2010)
2. Hartig, O., Bizer, C., Freytag, J.C.: Executing SPARQL queries over the web of linked data. In: ISWC 2009: Proceedings of the 8th International Semantic Web Conference, Chantilly, VA, USA. (2009) 293–309
3. Passant, A., Decker, S.: Hey! ho! let?s go! explanatory music recommendations with dbrec. In: 7th Extended Semantic Web Conference, ESWC Demo session. (2010)
4. Hausenblas, M.: Exploiting linked data to build web applications. IEEE Internet Computing **13** (2009) 68–73
5. Bizer, C., Heath, T., Berners-Lee, T.: Linked data – the story so far. International Journal on Semantic Web and Information Systems **5**(3) (2009) 1–22