The use of open data in the search of public financing and elaboration of studies for the development of projects on research and innovation

Short position paper for the Workshop "Open Data on the Web"

Mr Miguel García [a], Mrs Camino Correia [b]
(a, b) Zabala Innovation Consulting S.A.
(a) miguelgarcia@zabala.es (b) ccorreia@zabala.es

ABSTRACT

The search of public financing, management and tracking of scientific results derived from Research & Development and Innovation projects constitutes a business activity under which a remarkable number of consultancy firms, research centres and public organisations spend an important amount of resources. This expenditure of resources is motivated because the business processes involve working with an extensive and disperse knowledge on public financing, legislation and advanced web search abilities.

Technologies around open data (semantic web, standardisation of formats and final users' feedback) are an interesting starting point for the optimisation of processes and the cost effective management of data resources related with our business activity.

In this paper, we present the relation of the open data, from a users' point of view, as well as the business processes related to the search of public financing and elaboration of studies (many kinds) for the development of projects on research and innovation.

1. INTRODUCTION

Zabala Innovation Consulting (ZABALA) is a Spanish SME with around 150 employees [1]. ZABALA is an open data business user and sees the importance and benefit of open data not only for citizens but also for SMEs.

Our main activities with a direct relation with open data are 1) the search of public financing for research and innovation initiatives through the Spanish administration and the European Commission (EC) initiatives; and 2) the elaboration of studies for the development of projects on research and innovation for different sectors financed by the public administration within their innovation frameworks and policies.

Both activities do require the use and knowledge of open data sources to increase the benefit of the activities developed by the company.

2. OPEN DATA CASES

This section provides an overview of the cases related to open data on our business activity.

a) Search of public financing

One of the core business lines of the company is providing personalised information to a portfolio of over 700 customers regarding public funding opportunities in Spain (national, regional and province level) and Europe (FP7, CIP, Eranets, Eurostars and many other programmes).

The information related is published in different Official Gazettes from the different institutions financing the projects. The search among Gazettes can be considered an art itself given the number of public administration we are focused (over 60 gazettes only in Spain). The information on them is the open data we are targeting to.

Every single gazette has a different way to be published online. This refers not only to the formats but also on the division of the information per sections.

ZABALA is organised in a way that there are one or two people analysing the Official Gazettes (via RSS subscriptions, e-mail alerts and/or even with filters on Yahoo! Pipes) every day. They search any call, concession or modification in the related law ruling the financing programmes from a concrete number of gazettes. This process is time consuming and generates an important cost for the company.

Regarding the formats, they are usually PDF files with a division per sections. Some public organisms do also publish the information using other formats as JSON, RDF, etc. and the majority use TXT summarised versions of the PDF files earlier mentioned. TXT do not include tables with needed information such as percentages of reimbursement in projects, aid amounts per type of beneficiary, etc.).

Generating automatic extraction tools is a possibility but we have found that:

- The structure of the information published will practically need the development of tailored-made scrappers per source of information. This would mean over 60 developments only for the Spanish market (expensive).
- The information of gazettes in Spain is rarely published under the Linked Data paradigm (including semantic information). Nevertheless we've found interesting initiatives like the Public Contracts Ontology in LOD2 project [2], which could be of interest.
- Tailored-made scrappers would have to be adjusted every time a public organism decides to introduce modifications in their gazettes (new sections, new table styles, etc.).

On a wider approach, additional problems are found if the scope is opened to European programmes. The EC has been traditionally publishing aids at different websites depending on the programme under they were related to. This scope has changed lately as everything seems to be integrated in the Participant Portal [3]. Nevertheless the problem found here is related to the amount of information that needs to be manually processed to keep our internal databases updated.

Within our organisation we use a Document Management System with personalised developments integrated with other business processes. During the years, we have designed and implemented a database storing all the information categorised by programme, sub-programme and call, when related to public aids. Calls are stored in a table with around 50 different fields and relationships with other tables that are manually updated by the consultants every time a new opportunity is detected. Thanks to this, updated newsletters, lists and reports are generated under a different set of variables i.e. what aids in a region? What aids for SMEs in a sector? etc. An automatic filling of this information would be an important step for our processes.

b) Elaboration of studies for the development of Research and Innovation project.

The elaboration of this kind of studies is one of the main business lines of our company. These studies can be focused to present project proposals to the public administration, or performing sectorial studies to evaluate the impact of certain research results, etc.

Anyway, there are common areas which require the use of open data for these purposes:

Knowledge on the **state-of-the-art on a given subject**. The state-of-the-art is strongly related to earlier funded projects on the same subject. On the same way, listing past projects can give lots of information on research areas, results obtained, relevant partners to build new consortia, etc.

The EC publishes at different websites short summaries of every project funded under any Research or innovation programme which has ever existed. This information is under a Copyright which allows commercial use under the condition of due acknowledgment [4]. Moreover, each project usually publishes its results at their own website (mostly with Copyright advices).

Within ZABALA we have found the search functionality by the EC website is quite limited for our purposes, given the fact that listing of projects [5] does not show certain fields (it is impossible to list the real funds granted to a certain call in order to compare it with the planned ones and elaborate studies on that).

Besides this state-of-the-art, assessing the **impact of a project into the market or a financing policy** needs the use of different data published by the public administration.

Getting into an example, we could be considering performing an “impact study on a technology focused in the reduction of CO_2 in the industry”. This will take us to search for published information on CO_2 emissions in a concrete geographic area. We will find problems like:

- **Different sources of information related to the same topic**, which takes a considerable amount of time just to gather the raw sources.
- **Granularity**: meaning the same information is published at very different levels depending on the publisher. We could find CO_2 emissions per month in a region or CO_2 emissions in a concrete city of that region for the whole year. This lack of uniformity ends up on estimations which, inevitability, bring to an inaccurate analysis.
- **Update status**: much information is published with years of delay since their generation, depending on the needs of the users; this could mean the data is useless.
- **Legal restrictions**: we have found public open portals which are not very clear when defining the re-use of data for commercial purposes [6], which for ourselves it means, impeding using the data either way.
- **Multilingualism**: if the study is performed under different countries, we need to know a basis of the language under which the data is published.

3. TECHNICAL PERSPECTIVE AND NEEDS

Given our regular activities, ZABALA is not a technical company; we need technology to improve our open data based processes. The main needs of a company like ours are not different than the general ones for data re-users (automatic detection of relevant data sources, automatic integration of these sources with the company’s databases, application of the open definition [7] without any re-use restriction, foster the update of the information on a short-time basis by the public administration, etc.). Nevertheless some points are stressed:

- Easier formats for our analysis purposes (like CSV or any sort of editable tables). We are not so focused on the standard utilised but on the utilisation of editable information (PDFs are not workable formats, in terms of edition or manual combination with other formats, unless scrapers are designed for them). PDFs from gazettes constitute an important source of information.
- Standardisation on data granularity. We firmly believe the publication of certain information by the different administration (economic indicators, environmental figures, etc.) should be done following a structure on its content (common units, equal aggrupation of data, etc.) for its correct analysis.
- Apply multilingualism techniques to facilitate the multi-language sourcing data acquisition. This would be a key for boostering the internationalisation processes of our customers.

4. FUTURE OF OPEN DATA AT EUROPEAN LEVEL

Under our area of expertise, public funding and research, there has been an increasing interest on open data by the EC. From the Digital Agenda Assembly in 2011[8], open data has found a place in the EC and it has become a very relevant issue within the European policies. The workshop dedicated to open data gave birth to the most relevant communication from the EC [9] about open data.

This interest has been translated also in the launch of different public calls to fund research and innovation initiatives within the 7th Framework Programme (7FP) and the Competitiveness and Innovation Programme (CIP).

The future policies and trends for open data pushed by the EC are focused on three main topics [10]:

a) Creation of "data value chain friendly" policy environment.

This area is specially focused in the adoption of the Directive on the re-use of Public Sector Information (PSI) and the Commission decision on re-use of its own information, but also on the involvement and engagement of the stakeholders.

The definition of stakeholders, or actors in the open data value chain, sometimes forgets to take into account the participation of the data users' industry around (companies like ours) [11].

We disagree with the view that open data users are only citizens, as legitimate owners of the published data. Companies have been also populating the vast amounts of information that the public administration stores.

In this sense, we firmly believe the industry, understood as users, and not only as ICT data experts, should be included in the list of stakeholders.

b) Building of Multilingual (Open) Data infrastructure.

We will be really interested in knowing how these multilingual infrastructures are implemented, given our increasing needs in the search of opportunities abroad because of the internationalisation processes that many of our customers are now involved in.

c) Supporting Research and innovation.

From 2014, the EC will launch a new Research programme called Horizon 2020. Their plans are establishing data handling technologies as a priority within it. Besides this technological scope, the initiation of a European Data Forum and road-mapping is also other actions. Our interest will be here at two sides:

- Funds for research projects: both for customers and ourselves.
- Involvement of users in the road-mapping to have a clear view on their needs.

5. OUR INITIATIVES ON OPEN DATA

ZABALA has also its own roadmap in terms of research focused on the exploitation of open data. In this sense, ZABALA lead a proposal presented to the EC related to the search of public financing opportunities worldwide. The project, still under evaluation, would develop a software component which will help the publication of official gazettes under the Linked Data paradigm, in order to integrate its results into a public financing search tool for companies like ours and their customers.

Besides this, ZABALA is strongly committed to be part of the open data revolution in the participation in seminars, workshops and any sort of events where the implication of the users' industry would be advisable.

REFERENCES

[1] Zabala Innovation Consulting website http://www.zabala.es

[2] Public Contracts Ontology http://lod2.eu/BlogPost/1088-modelling-public-procurement-for-the-linked-open-data-cloud-release-of-the-public-contracts-ontology.html

[3] Participant Portal website http://ec.europa.eu/research/participants/portal/

[4] CORDIS website legal notice http://cordis.europa.eu/guidance/legal-notices_en.html

[5] Project search functionality provided by CORDIS. http://cordis.europa.eu/projects/home_en.html

[6] INE (Statistics National Institute in Spain) does not specify if the reuse of data can be done for commercial purposes: http://www.ine.es/ss/Satellite?c=Page&p=1254735849170&pagename=Ayuda%2FINELayout&cid=1254735849170&L=1#

[7] Open definition http://opendefinition.org/okd/

[8] Open data and re-use of public sector information https://ec.europa.eu/digital-agenda/01-open-data-and-re-use-public-sector-information

[9] COM(2011) 882 final. Open data. An engine for innovation, growth and transparent governance http://ec.europa.eu/information_society/policy/psi/docs/pdfs/opendata2012/open_data_communication/en.pdf

[10] European policies on open data- Presentation by Szymon Lewandowski, Policy officer - "Data Value Chain" Unit http://ec.europa.eu/information_society/newsroom/cf//itemdetail.cfm?item_id=9692

[11] Open Data Approach and Innovation in Brazil – Position paper http://www.w3.org/2012/06/pmod/pmod2012_submission_25.pdf