COMSODE.eu : Components Supporting the Open Data Exploitation

Martin Necasky[(1)], Andrea Maurino[(2)], Miroslav Konecny[(3)]

[(1)]Charles University in Prague, Czech Republic, necasky@opendata.cz
[(2)]University of Milano Bicocca, Italy, maurino@disco.unimib.it
[(3)]ADDSEN, s.r.o., Slovakia,, konecny@addsen.eu

Short CVs:

Martin Necasky is an assistant professor at Charles University in Prague. He is also the founder of OpenData.cz initiative which promotes the ideas of (linked) open data in Czech Republic. He cooperates on the COMSODE project. He published more than 70 papers in international journals and conferences including Journal of Data Engineering, Journal of Systems and Software, ICWS, WISE, etc. His research interests include service oriented systems, semantic web and linked open data.

Andrea Maurino is an assistant professor at University of Milano Bicocca. He is the coordinator of the COMSODE project and he published more than 70 papers in international journal and conference including ACM computer surveys, VLDB, CAISE, AMCIS. His research interests include open data in particular in methodology for open data publication, data quality and in particular in temporal record linkage and temporal data assessment and innovation economy.

Miroslav Konecny is senior project manager and owner of ADDSEN. He is acting as dissemination & User Board manager of COMSODE project, having responsibilities for demand-driven approach and public engagement strategies. He has been working since 2006 as a project manager for FP7 and structural funds scientific projects. Miroslav has gained experience in FP7 project management in 5 collaborative projects, 1 capacities project and 2 CIPs.

Type of the presentation: Research contribution

Summary of the presentation: We present a new EU funded research project COMSODE. Its mission is to progress the capabilities in the Open Data re-use field with a strong emphasis on data quality. We will develop a publication platform and a methodology starting from the fusion of the best practice with the idea to produce a concrete and feasible solution supporting all the open data lifecycle. We cooperate with many public bodies ranging from small cities to large national institutions in several EU countries. The partnering bodies are represented in the COMSODE user panel which will be also presented during the talk.

1. Introduction

Computing and networking capabilities generate and store more data than any other time in history. Such trend combined with openness enhances the potential impact of the accumulated data, offering society an opportunity to drive massive social, political and economic change. Open government data transforms public administration into information providers and therefore may increase the transparency, participation and generate a relevant economic growth. The new EU funded project COMSODE (Components Supporting the Open Data Exploitation) started in October 2013. It is an SME-driven RTD project aimed at progressing the capabilities in the Open Data re-use field. To increase the usage of open data there is the need to completely rethink the way in which open data are currently provided. Actually, open data published by various open data catalogues are poorly integrated; quality assessment, and cleansing are seldom addressed; many existing open data portals do not take much attention to timeliness of provided data. Moreover, data consumers should be able to integrate the data before using them. There are initiatives for improving the publication of open data, but these are often isolated in a continuously changing landscape. The COMSODE project wants to develop a publication platform and an original methodology supporting all the open data lifecycle.

The paper is organized as follows. In Section 2 we present the most important goal of the COMSODE project and the research problem we address. In Section 3 we present the Open Data Node (the publication platform) we will deliver. In Section 4 we introduce the user board, the tool we use to put together all open data stakeholders. In Section 5 we draw conclusion and future work.

2. COMSODE goals and research problems

The project addresses several issues related to the publication of open data with emphasis of data quality. Two most important ones are:

(1) **Create a publication platform called Open Data Node** that builds on the results of previous research and development in the linked data field. A lot of inspiration came also from projects related to digitisation of cultural heritage and Europeana. Its mission is to bring the results from research environment into real-world for people, SMEs and other organizations.

(2) **Create a methodology framework** for easy use of technology in operating conditions of typical public bodies and rigorously tested for traceability, usability and sustainability in a public body environment. This is verified in three pilot implementations during the project. End-user communities are involved EU-wide to set a use case framework within which the requirements of heterogeneous organisations can be clearly understood. The provided feedback will be later processed into the final methodology and recommendations for re-use applications.

Concerning the publication platform, it is a project ambition to lay foundations for a data integration platform based on Open Data which allows the re-use of data not only between public bodies and end-users but also among public bodies themselves: Public bodies can exchange

information by using the same infrastructure and tools as end-users which will decrease costs of exchanging the data and in most cases also enhance the quality and speed-up the exchange. What is even more important, Open Data APIs can be used by integration projects among public bodies, again saving costs and enhancing the quality of the resulting solution. This in turn strengthens Open Data publishing, with end-users benefiting again – a self reinforcing loop.

The methodological framework will be developed by taking into account existing proposals integrating them with the most relevant and appreciated e-government strategies for the service publication. In the COMSODE project we want to offer to public decision makers the conceptual tools for the definition of an open data policy according to their general view and how it can be really aligned with the IT level with a continuous attention to all stakeholders (citizens, business companies, NGO, civil hackers, other PA). The introduction of tools supporting the maintenance and measurement of open data results could increase the open data policy adoption.

3. Open Data Node

We develop a publication software platform called Open Data Node (ODN). The main aim of ODN is to support public bodies during the whole process of publishing their open data in different formats. ODN is free to use and modify (Open Source) and provides two groups of services to a public body willing to publish its datasets according to open data principles.

The first group supports the management of datasets of a public body which will use ODN.

1. Maintenance of the internal catalogue of datasets, metadata for each dataset (including the provenance metadata), versioning, and information whether the dataset has been opened or not.
2. Integration with the local software environment of the public body for the dataset content extraction from internal data storages (databases, file system, etc.) and ensuring timely updates of the content of the datasets in defined time intervals by loading new content which originated in the environment.
    a. Support for various formats of input data (CSV, XLS(X), OpenDocument, XML, RDF)
    b. Support for various access methods (JDBC, Web Services, SPARQL)
3. Importing of the datasets published elsewhere.
4. Cleansing and transforming of the (updated) datasets and their linking to other datasets (inside or outside the public body) according to the defined rules.
    a. Many generic cleansing and transformation rules will be provided as inherent components of each ODN installation. However, it will be also possible to define new rules specific for the public body and its datasets.
    b. Anonymization will be addressed as a special category of transformations since it is important for publishers to be conformant to various Privacy protection regulations.

The second group of services supports publication of the datasets maintained in the ODN

instance. In particular, the following services belong to this group:

5. Automated export of the metadata about (updated) datasets (including the provenance metadata) to defined open data catalogues.
6. Publishing the datasets (and their historical versions) both for bulk downloads (i.e., in a form of datasets dumps) and through APIs.
    a. Support for various formats of output data - CSV, OpenDocument, XML, RDF
    b. REST APIs are automatically generated for the known format
7. Notification of registered consumers about updates in the published datasets.
8. Publication of "audit trail" (what, when, which way and by whom was updated) for managed datasets to general public to help build trust in the data.
9. Replication of the datasets so that for example 3rd party application developers are able to bridge performance bottlenecks or maintenance windows of the data publishers.

Note that even though ODN is primarily aimed for public bodies, various data aggregators and even application developers will find ODN useful. ODN enables to import open data from different sources and define processes which integrate those data and provide results through automatically generated REST APIs. Those APIs can be either offered by an aggregator to its users/customers or integrated to his or her application by an application developer.

ODN will also help with bridging of language barriers. With the support of RDF and methodologies, ODN will allow publishers to provide multilingual Linked Open Data which will then have much more potential for reuse, for example:

- broader international audience
- greater potential when linked with data from other countries

3.1 Strong Emphasis on Data Quality

According to the 5-stars open data schema[1], ODN will publish datasets with at least 3 stars. However, it is desirable to achieve 4 or 5 stars when possible. 4 stars mean that URIs are used to denote things present in the dataset. For this, ODN built-in components will be able to recognize current entities (e.g., geographical locations) and will enrich them with URIs assigned to them in the LOD cloud. It will also be possible to define own components which provide URIs for specific entities. Well-recognized portions of the datasets will be also converted to RDF according to Linked Data principles, i.e. to 5 stars. Widespread ontologies (e.g., schema.org or those listed in lov.okfn.org) will be used to represent the created Linked Data. Again, this will be done automatically by ODN using built-in components and partly it will be done by user specific components.

ODN will enable to define data refinement processes which will assess various dimensions of data quality and cleanse and transform datasets so that quality of the published dataset is

[1] HAUSENBLAS, Michael. 5 star Open Data. In: *5 star Open Data* [online]. Last update: 2012-04-03 [cit. 2013-09-20]. Available from: http://5stardata.info/

increased. Each data refinement process is a sequence of steps; each step solves some portion of the necessary quality assessment, cleansing and transformation, e.g., one step may examine the syntactical accuracy of the data, so that another step may correct certain syntactical errors in the data. The goal is to provide a quality assessment methodology to assess the quality of (Linked) Open Data sets and provide a set of components realizing various steps of the refinement processes.

4. COMSODE Use Cases and User Group

The COMSODE project contributes to methodological aspects of Open Data policy and deals with practical implications of Open Data Node deployment in real-life conditions. The project pilotes the ODN publication platform and related methodology in several use cases: For example, in the Czech Republic the project delivers to national bodies like Czech Telecommunication Office, Czech Trade Inspection Authority, Ministry of Interior; in Slovakia the Ministry of Interior is part of the consortium, the digital champion is an associated partner to the project, etc.

Representative users of COMSODE prospective results have been invited to constitute the COMSODE User Group. Currently the COMSODE User Group contains bodies from 10 EU countries. Their aim is to provide valuable feedback to the COMSODE interim results and helping to steer it towards fulfillment of user needs (with user being either data publishers or data reusers).

5. Conclusion

In this paper we shortly presented the COMSODE vision to support the reuse of open data by rethinking the whole open data lifecycle taking into account the best experiences available nowadays. Two are the most important and integrated contributions: a methodological framework that supports decision makers from the policy definition to the technical implementation and a publication platform built on the top of the best solutions available. The project has just started and authors and all consortium partners are strongly dedicated to support the open data movement and its impact into the European knowledge society by means of the vision just sketched above. We are also very interested to collaborate with all current open data initiatives in Europe to effectively support the cross-fertilization in such very dynamic field.