COMSODE: Component Supporting the Open Data Exploitation

Andrea Maurino

University of Milano Bicocca
maurino@disco.unimib.it

Extended Abstract

The EU funded project COMSODE (Components Supporting the Open Data Exploitation)[1] started in October 2013 is an SME-driven RTD project aimed at progressing the capabilities in the Open Data reuse field. The COMSODE project wants to develop a publication platform and an original methodology starting from the fusion of the best practice, and solution available with the idea to produce a concrete and feasible global solution supporting all the open data life-cycle from the open data policy definition to the technological platform.
COMSODE goals and research problems. Two are the most important problems faced in COMSODE :

1. Create a publication platform called Open Data Node that builds on results of previous research and development in the linked data field. 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. End-user communities are involved EU-wide to set a use case framework within which the requirements of heterogeneous organisations can be clearly understood. Provided feedback will be later processed into the final methodology and recommendations for reuse applications.

The methodological framework will be developed by taking into account existing proposals integrating them with most relevant and appreciated egovernment strategies for the service publication. In the COMSODE project we want to offer to public decision makers the conceptual tools for the definition of a 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.
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

[1] www.comsode.eu

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 public body in the management of its datasets.

1. Maintenance of the internal catalogue of the datasets of the public body, metadata for each datasets (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 extraction of the content of the datasets from internal data storages of the public body (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. Support for various data formats on the input (CSV, XLS(X), OpenDocument, XML, RDF), 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 of the body) according to defined rules. 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 body and its datasets. Anonymization will be addressed as a special category of transformations as 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:

1. Automated export of the metadata about (updated) datasets (including the provenance metadata) to defined open data catalogues.
2. Publishing the datasets (and their historical versions) both for bulk downloads (i.e., in a form of datasets dumps) and through APIs. Support for various data formats on the output CSV,OpenDocument, XML, RDF REST APIs are automatically generated for the known format
3. Notification of registered consumers about updates in the published datasets.
4. 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.

Strong Emphasis on Data Quality. According to the 5stars open data schema[2], ODN will publish datasets at least with 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 most often present entities (e.g., geographical locations) and will enrich them with their URIs in the LOD cloud. It will also be possible to define own ODN components which provide URIs for specific entities.

[2] http://5stardata.info/