Open Data with DaPaaS:
From Tabular Files to Data Services

Contributor names and short CVs

Dr. Dumitru Roman works as a Senior Research Scientist at SINTEF (Norway). He is active in the Open Linked Data field where he currently acts as the project coordinator of the DaPaaS and proDataMarket projects---both dealing with innovative products and services using Open Data. He holds an adjunct associate professorship at the University of Oslo, Norway.

Marin Dimitrov is the CTO of Ontotext (Bulgaria), with deep expertise in text mining, Linked Open Data, Semantic Technology, cloud computing and Big Data. Marin is a member and active participant on several program committees and conferences in the fields of Linked Data Management and Text Analytics. Marin earned his Masters of Science in Artificial Intelligence from the University of Sofia (Bulgaria).

Rick Moynihan is is a software engineer and technologist who's spent over 12 years working with startups to help research and develop new technologies and try to bring them to market. He's long been a believer in the transformative nature of the Internet, open source software and open data, and is pleased to be working for Swirrl (UK) with the DaPaaS partners on data platforms that solve real user problems.

Amanda Smith is the Community Engagement Manager at the Open Data Institute (UK). She is primarily responsible for undertaking research, developing tools, services and best practice and managing the dissemination activities of the ODI's EU funded research projects. Before joining the ODI, Amanda worked in policing and government, and she holds a BA Hons in Criminology. She discovered her passion for open data when working with police forces

throughout the country to release open crime and justice data, developing the national crime mapping website, Police.uk and its data site Data.police.uk.

Ivan Berlocher is a Chief Scientist in Saltlux (South Korea). Ivan has an M.Sc. and is a candidate Ph.D. in IRISA, which is located in Rennes University, France. Ivan has led lots of industrial projects, and he designed the core architecture of most of Saltlux's products in semantics, text mining and information retrieval

Momchill Zarev is CEO and founder of Sirma Mobile (bulgaria). He is an expert in mobile technologies with an excellent technical background and more than twenty years expertise in different software technologies.

Type of presentation

Research contribution

Title of the presentation

Open Data with DaPaaS: From Tabular Files to Data Services

Summary of the presentation

DaPaaS (*Data- and Platform-as-a-Service approach to efficient Open Data publication & consumption*) is an FP7 EC-funded research project that has the goal to make publishing, consumption, and reuse of open data, easier and cheaper for SMEs and small public bodies which otherwise may not have sufficient technical expertise, infrastructure and resources required to do so. DaPaaS uniquely combines Open Data, Linked Data, and Data-as-a-Service concepts to offer innovative solutions for simplified and cost-effective open data access. The key elements of the DaPaaS platform which will be part of the presentation are: the tools for data cleaning and transformation of tabular data into RDF; the scalable data hosting layer providing access to open data services; the data access portal and a proof-of-concept demonstrator showcasing DaPaaS technologies in the smart cities domain.

Extended abstract

In the space of just a few years we've seen the transformational power of Open Data; both for transparency and accountability in public data, and efficiency and innovation with businesses in private data. In the recent years public institutions, start-ups and developers throughout Europe have been actively involved in releasing data and reusing this open data for economic benefit.

However, we are still at the beginning of the Open Data movement, and there is still more that can be done to make Open Data simpler to use and to make it available to a wider audience. The current Open Data landscape consists mostly of static files with tabular data, which is often difficult to access, integrate and interlink with other data, and often not sufficiently enriched with important metadata. Such tabular data files often have data quality issues which further make their effective reuse difficult. The market calls for better tools and approaches for simplified and cost-effective Open Data access for data workers and data developers.

DaPaaS (Data-and-Platform-as-a-Service) is an FP7 EC-funded research project, that provides a platform with associated tools for making publishing, hosting and consuming linked open data easier. The goal of DaPaaS is to provide Open Data publishers with the tools they need to create high quality Linked Data, which is then easily accessible via live data services. DaPaaS integrates the best technology from leaders in the fields of open and linked data publishing, semantic web & data consumption, to provide users with a one-stop solution for data publishing and associated transformation services.

Essentially, the platform aims to make it easy for open data publishers --- who may not have sufficient technical expertise, infrastructure or the resources required --- to push the data publishing process beyond static tabular files and into live data services, which can be easily and reliably accessed by 3rd party applications and data mashups.

The *key innovations* of the DaPaaS platform include:

- Moving beyond tabular data format and using Linked Data as the means to publish, interlink and reuse data. Data is easily queryable and not just available as a static data dump.
- Providing intuitive frontend tools that make the data transformation and data quality processes easy to a wide audience of data publishers. Data transformations are reusable and can be shared among different users.
- Providing simple and well documented APIs for developers, for accessing key platform services and the open data services.
- Providing support for scalable data transformations and handling even very large tabular files. The data hosting layer is scalable and capable of handling a very large number of live data services and concurrent users (applications).

The *key scientific & technical components* realizing the above innovations of the DaPaaS platform include:

- *Grafter* is an open source suite of tools for tabular data transformation & processing. Grafter can be used to transform tabular data formats to tabular data formats (for data quality / cleanup purposes), or to Linked Data format (for publishing a live data service on the platform). A key feature of Grafter is that the resulting transformations can be serialised, shared between users and repeatedly executed over data (e.g. data transformation services).
- *Grafterizer* is the frontend framework for the Grafter suite. It assists data publishers with creating (Grafter) data transformations and mappings to Linked Data ontologies and vocabularies.
- The *Database-as-a-Service layer* which turns the RDF-ised legacy data into live data services, easily accessible and queryable by developers and applications. A key feature of the data layer is scalability and reliability.
- The *Data Portal* which provides a catalogue of various datasets (data services) as well as reusable data transformation services.
- The *Personalized and Localized Urban Quality Index (PLUQI)* application, which is a demonstrator of the DaPaaS platform. PLUQI uses DaPaaS to integrate various open datasets (e.g. transportation, weather, crime statistics, financial indicators, etc.) and computes various indexes on well-being and sustainability of cities, and visualizes them through a fancy GUI.

By using DaPaaS, data workers and data developers will not only get access to *better tools and services for data transformation, hosting, and access*, but will do so in a *simplified and cost-effective manner*. DaPaas will enable them to increase the quality and consistency of their published data, while at the same time increasing the uptake of Open Data. With DaPaaS, we overcome some of the main challenges regarding uptake of Open Data publishing and consumption. DaPaaS specifically enables:

- A reduction in the cost for those who lack sufficient expertise and resources to publish open data;
- A reduction on the dependency of open data publishers on generic Cloud platforms to build, deploy and maintain their open/linked data from scratch;
- An increase in the speed of publishing new datasets and updating existing datasets through the provision of a sound methodology and integrated toolset that will support the full linked open data lifecycle.

The proposed presentation will give an overview of the DaPaaS platform with a specific focus on its key scientific & technical components, emphasizing key innovations and impact.