

Changing the Data Paradigm: Turning 'Open Data' into 'Linked Data'

Dr. Erik Mannens, iMinds - Ghent University

A couple of years ago, governments started to make loads of information available on their websites (ranging from contact lists to image libraries, event lists, or urban planning files); it enables them to share data in an efficient way with citizens and companies.

Unfortunately, these data typically come in a wide variety of digital formats (CSV, PDF, XLS, etc.) – and sources are rarely linked with one another. It makes reusing or combining information very difficult and prevents the creation of more transparency, the stimulation of economic growth, or the improvement of public service.

That is where the concept of 'Open Data' comes in: it is about making data publicly and electronically available – in such a way that information from various sources can be shared, retrieved and combined easily and flexibly; it is an area in which the region of Flanders (Belgium) has recently started to lead the way!

In the last twelve months, Flanders has made big progress in the creation of an Open Data policy – and related ecosystem – thanks to the involvement of some local and regional governments, a number of software developers and the MMLab researchers from iMinds - Ghent University, which were sponsored by the Flemish Ministry of Innovation.

And it does not stop there: Flanders is ready to take the next step too, i.e. moving from Open Data to Linked Data – whereby information is not only easily retrievable, but also linked. To this purpose, the iMinds-MMLab team has developed two innovative Open Source platforms: *TheDataTank* and its accompanying *R&Wbase* platform.

TheDataTank

The US and UK governments currently operate Data.gov and Data.gov.uk, huge platforms for data publishing and retrieval. Both come with a number of shortcomings, though: their semantics capabilities are limited (or non-existing), they do not allow developers to easily re-use the data sets, and they are not equipped with a simple WYSIWYG tool for end-users to easily link diverse datasets to one another.

TheDataTank (<http://www.thedatatank.com>) from the iMinds-MMLab team provides an answer to these limitations: it is a low-barrier Open Data publishing platform that enables anybody to publish datasets in an easy and automated way, and connect them with other open data. Through its user-friendly interface, data from national, regional and local governmental departments can easily be uploaded, updated, re-used and aggregated by any third party.

TheDataTank has already passed its first major test. It was applied to data sets on academic publications and joint research projects, visualizing how researchers of different Flemish universities work together, how they perform vis-à-vis their peers, how different research labs share resources in joint projects, and how these labs perform vis-à-vis their peers. Easily combining and visualizing this information was previously unthinkable; now, the outcomes can be used by researchers, universities and the Flemish funding agency EWJ to optimize their individual and global excellence levels, as well as the related funding.

The R&Wbase platform

Today, the basic model for data processing is a 'one-way street': originators of data (such as governments) publish read-only data; there are no feedback loops; and temporal versioning is not supported.

This situation is far from ideal. Instead, 'data cycles' should be introduced – with users being able to flag errors or submit corrections. That way, data quality improves at

the source and a real ecosystem is created, thriving on collaboration and read-write open data.

The R&Wbase (read: *Rawbase*) cloud-based platform responds to this read-write web paradigm: it creates a feedback mechanism that logs any changes, identifies the contributor, and tracks the edits. Anyone who wants to open up data can install their own R&Wbase. Afterwards, if a third party wants to aggregate the different kinds of versioned read-write data, they only have to install a R&Wbase instance and link to other R&Wbases. It is a hierarchical and distributed environment where aggregation can be done in a matter of minutes.

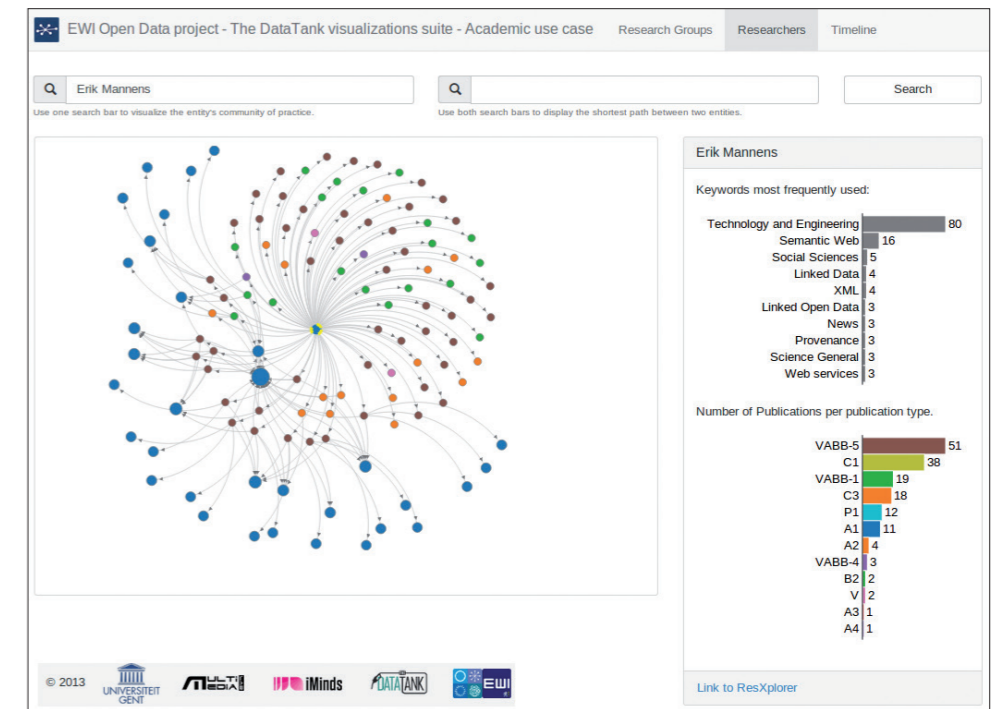
Application domains and challenges ahead

Governments are natural sponsors of Open Data initiatives – both as sources of data and as regulators of Open Data usage. But according to McKinsey¹, the concept of Open Data also unlocks a whole new range of other opportunities:

- Businesses are blending Open Data with proprietary information to get better insights into their target markets, or raise productivity by using more refined benchmarks
- Consumers benefit from Open Data as it allows them to take fully-informed decisions – based on all available info from a variety of sources (when buying a new house, for instance, they can easily retrieve and combine information on crime statistics, traffic flows, etc. in their new neighborhood)

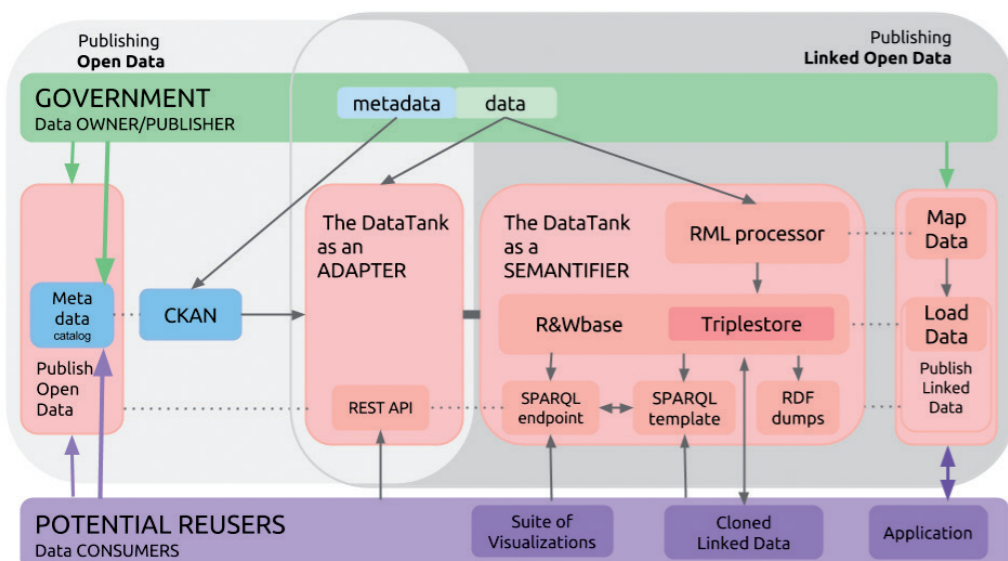
A practical case of how Open Data are already used today, can be found in Japan where citizens were mobilized to create an open database of radiation readings in different parts of the country following the Fukushima earthquake.

"But there are also some challenges ahead," says Dr. Erik Mannens, iMinds - Ghent University - MMLab. "Especially a shift in mindset is required. Generally, people are reluctant to open up (and give away) their data. We have to make it clear, though, that – if everyone does it – everybody wins."



"Of course, there are other, more technical obstacles too," he adds. "Not in the least the ability to produce results in real time when dealing with gargantuan amounts of data. This will be a challenge for TheDataTank and the R&Wbase platforms – but we already have a good idea of how to tackle Streaming Big Data too."

¹ McKinsey Global Institute: *Open data: Unlocking innovation and performance with liquid information* - October 2013 - <http://bit.ly/1hq8CIT>



Dr. Erik Mannens is Research Manager of the Future Media & Imaging Dept. within iMinds and is involved in several projects as Sr. researcher/project manager. His major expertise is centered around Big Data Analytics, Linked Open Data, Semantic Web Technologies, Broadcasting Workflows, Media & iDTV, and Web Development in general.