# A three-level data publishing portal

Pieter Colpaert, Anastasia Dimou, Miel Vander Sande, Jonas Breuer, Mathias Van Compernolle, Erik Mannens, Peter Mechant, Rik Van de Walle

<firstname>.<lastname>@ugent.be
**UGent - iMinds - MMLab, MICT and SMIT**

## Contributors and their short CVs

### Pieter Colpaert

Pieter holds an engineering degree in informatics from College Ghent, Belgium. Before his position as a researcher in Linked Open Data at MultiMediaLab - Ghent University, he started his own company which visualises Open Data on digital information screens (FlatTurtle) and founded the Belgian chapter of the Open Knowledge Foundation. In order to lower the barrier for data owners to publish data, Pieter started The DataTank as part of his master's thesis in 2010, and continued working on it until today.

### Anastasia Dimou

Anastasia Dimou is a scientific researcher at Multimedia Lab, Ghent University, active in the research of (Linked) Open Data and Semantic Web technologies. She graduated from two Masters: Master in technologies for e-Government from University of Trento, Italy (UNITN) in 2009 and Master in Web Science from Aristotle University of Thessaloniki (AUTH), Greece in 2012. Her Master thesis for the Master in Web Science was entitled "Exploring Scientific Knowledge using Linked Data" dealing with the transformation of scientific Classifications into Simple Knowledge Organization System (SKOS) and the emergence of mappings between the distinct scientific Classification Schemes. Her research interests were focused on applying Semantic Web technologies in education and research and she has previous experience with Semantic Wikis and Semantic Content Management Systems. Her current research work is focused on mapping and interlinking procedures to integrate, semantically annotate and interlink Open Data while and after publishing.

### Miel Vander Sande
In 2008 **Miel Vander Sande** graduated as Bachelor in Multimedia and Communication technology and in 2010 as Master in Industrial Engineering: ICT. He wrote his Master

thesis in collaboration with the University of Valencia in Spain about analyzing and visualizing RFID tracking data. Finally, he ended his education with an extra year of teacher studies in Informatics. Since september 2011, Miel joined UGent-iMinds in the research group Multimedia Lab (http://multimedialab.elis.ugent.be) as a researcher. His main interest and expertise are (linked open) data publishing (a.o. in the context of Open Knowledge Foundation), technological support for open data legislation, ontology mapping and data transformation. He is very active as an activist of Open Data in the Belgian and European community, supporting policy making and the organization of events (such as App contests). Miel was involved in several Flemish and European research projects involving Semantic Web technologies and Open Data. Miel Vander Sande co-developed *The DataTank platform* (http://thedatatank.com) with the Belgian Chapter of the *Open Knowledge Foundation (OKFN)*, which he co-founded. The DataTank is a data publishing platform, allowing quick RESTful publishing and transformation of plain datasets. Miel's contribution involved on-the-fly Linked Data transformation. Together with his team, Miel developed the Everything is Connected (http://www.everythingisconnected.be) demo. This demo makes Linked Data tangible by finding paths between concepts using the DBPedia dataset, and presenting these paths as multimedia presentations. Currently, Miel is active in multiple Flemish government projects for creating read/write Open Data ecosystems. The goal is to create innovative platforms, that reuse the existing Web to produce, consume and benefit from Linked Data. The project aim towards a reference implementation for the W3C Linked Data Platform Working Group. Additionally, he participates in a Flemish innovation project for the digitalization of book publishers and eBooks: *Publishers of the future*.

## Jonas Breuer

Jonas holds a bachelor degree in European Studies from the University of Maastricht, Netherlands. After gaining work experience in the communication sector, he graduated as Master of Science in Communication Studies with excellent results in 2013.

Jonas works for iMinds-Smit since August 2013, within the unit of Market Innovation and Sector Transition. During this international Master program at the Vrije Universiteit Brussel, Belgium, he concentrated on the political economy of new media and ICT, focussing in particular on the Smart City. His main research interest still revolves around Smart Cities and related concepts, particularly considering the role that data assumes in this context. At the moment, Jonas works on business models that can be based on open data.

## Mathias Van Compernolle

Mathias holds a Bachelor degree in Social Work from Artevelde University College and graduated in 2012 as a Master in Public Administration from Ghent University College. Before starting as a researcher at the research group for Media and ICT (MICT) at Ghent University - iMinds, he was active as coördinator digital social innovation at iDROPS. Next to this he co-organizes and facilitates Open Data co-creation events in Flanders at the Belgian Chapter of the Open Knowledge Foundation.

# Peter Mechant

Peter Mechant (PhD Communication Sciences) works at the research group for Media and ICT (MICT) at Ghent University and iMinds. He has been involved in various research projects focusing on e-culture, e-learning, Web 2.0 and smart cities. Peter has a keen interest in 'Big' (Social Media) Data and the additional value it can provide to social research, investigating these opportunities from a theoretical as well as a practical perspective. Peter has published in journals such as Observatorio, International Journal of Web-based Communities, International Journal of Interactive Communication Systems and Technologies, Contemporary Social Science or European Journal of Communication Research and coauthored papers in journals such as New Media and Society, Journal of Computer-mediated Communication and Cyberpsychology, Behaviour and Social Networking or Communication&Strategies.

# Erik Mannens

**dr. ing. Erik Mannens** is Research Manager - Future Media & Imaging Dept. / experienced Project Manager at iMinds (formerly known as IBBT) since 2005 where he has successfully managed +30 projects. He received his PhD degree in Computer Science Engineering (2011) at UGent, his Master's degree in Computer Science (1995) at K.U. Leuven University, and his Master's degree in Electro-Mechanical Engineering (1992) at KAHO Ghent. Before joining iMinds-MMLab in 2005 as project manager, he was a software engineering consultant and Java architect for over a decade. His major expertise is centered around metadata modeling, semantic web technologies, broadcasting workflows, iDTV and web development in general. He is involved in several projects as senior researcher and just finished up his PhD on Semantic News Production; he is co-chair of the W3C Media Fragments Working Group and actively participating in other W3C's semantic web standardization activities (Media Annotations, Provenance, and eGovernment). Since 2008 Erik is paving the Open Data path in Flanders. He stood at the cradle of the first Hackatons and is a founding member of the Open Knowledge Foundation (Belgian Chapter). Since then, he is frequently invited as Open Data evangelist at national and international events. He currently actively participates in W3C's eGov and Linked Data Platform working groups. Furthermore his team is owner of the Open Sourced Linked Open Data Publishing frameworks TheDataTank and R&Wbase. On all of these subjects he has published several papers and book chapters. He is also member of the technical committee of MTAP, ACM Multimedia, MareSO, CCNC, and SAMT. His full bio/CV can be obtained from both http://www.mmlab.be/emannens or http://www.linkedin.com/in/erikmannens. *Specialties*: (Linked) Open Data, Big Data Analysis, (Semantic) Web development, project management, W3C standardization (MMSEM, Video WG - Media Fragments & Media Annotation, Provenance WG), Java architect, iDTV.

# Type of presentation

Research contribution

# Summary of the presentation

Not all data published on Open Government Data Portals reaches the same level. In this talk, a distinction is made between non-machine readable data (1st level), data about which only the serialization format is known (2nd level) and data about which both serialization format is known as well as the model (3d level). At each of these three levels, the barrier can be lowered to get to the next star of Linked Open Data. The importance of getting data to the 3d level is illustrated by three use cases in development for the Flemish government.

# Extended abstract

See next page

# A three-level data publishing portal

Pieter Colpaert, Anastasia Dimou, Miel Vander Sande, Jonas Breuer, Mathias Van Compernolle, Erik Mannens, Peter Mechant, Rik Van de Walle

<firstname>.<lastname>@ugent.be
UGent - iMinds - MMLab, MICT and SMIT

## Introduction

Following European directives such as the Public Sector Information[1] (PSI) directive or INSPIRE[2] (Infrastructure for Spatial Information in the European Community) directive, governments across Europe are structuring and publishing data in such a way that it is 'open' to the public. In this context, Flanders (the northern part of Belgium) is creating a Flemish Open Data Portal. This paper describes the architecture developed for this portal. Similar to the "5 stars of Linked Open Data"-system - introduced by sir Tim Berners-Lee - which advocates Linked Open Data, the Flemish Open Data Portal affords to publish data under an open license as a first step and tries to lower the barrier for datasets to get closer towards the 5 star Linked Open Data. It targets to transform data into RDF in the long run and wants to link it to other datasets. However, not all datasets will be published as Linked Open Data because of two reasons; firstly, there is no one-size-fits-all formula [1] available, and secondly building linked data cycles at organization-level involves a long process.

After elaborating on the architecture of the envisioned Flemish Open Data Portal, the paper continues to describe the use cases that will be built on top of this portal as well as its socio-economic potential.

## The architecture

Open Data enables a dialog between governments and citizens. Data is shared with the public, that can consume data using applications (e.g., mobile apps, statistical websites) that can span any domain. The collected feedback from data consumption, can be embodied as enrichments, comments, or corrections. However, in order to close this feedback loop, a data portal must provide full writing, as well as reading functionality. In addition, governments need to manage incoming feedback, as contributions by multiple participants need to verified, and decide what is being republished. Tackling these challenges affects the complete technology stack of such a portal, a solution needs to affect the data storage level.

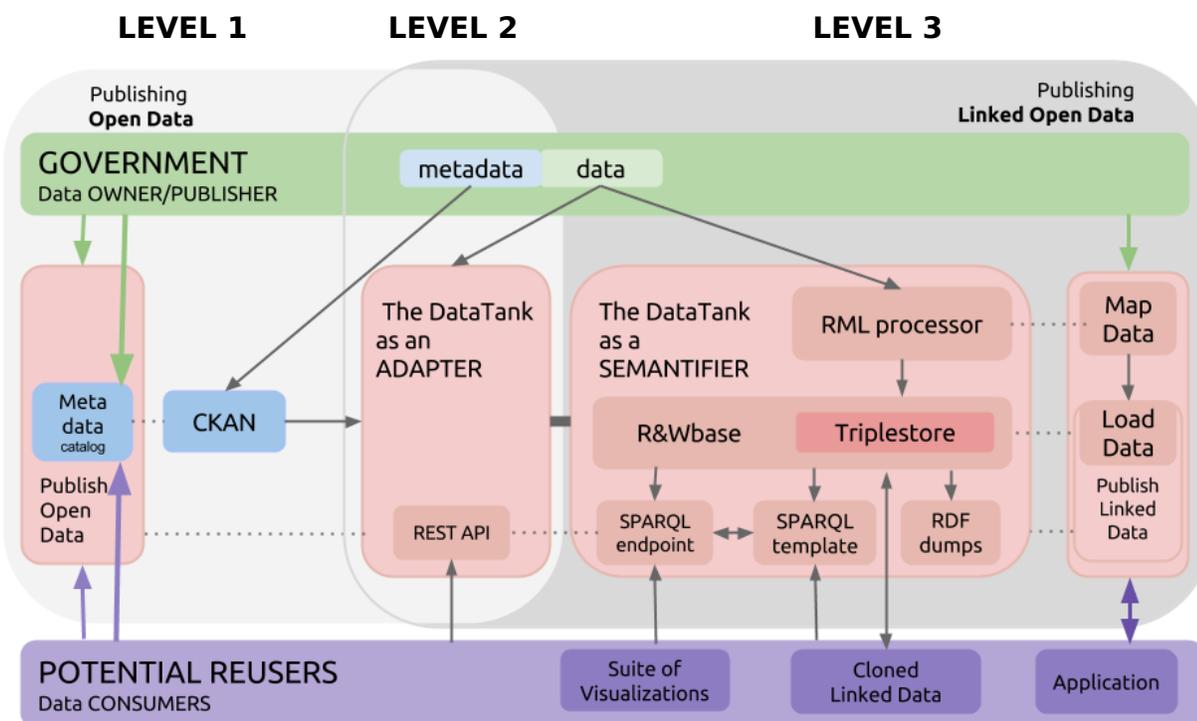The Flemish Open Data Portal can open up data on three levels which are in line with the

---

[1] http://epsiplatform.eu
[2] http://inspire.ec.europa.eu/

5 stars of Linked Open Data:

*1. Non-machine readable data (level 1).* In this case, only data about the data can be stored. The metadata are indexed and published online so that the dataset becomes discoverable. The Flemish Open Data Portal uses CKAN to maintain a list of all the available datasets inside the Flemish government.

*2. Data about which only the serialization format is known (level 2).* This type refers to data available in a machine readable format, such as JSON, CSV or XML. Here the DataTank [3], an essential component in the architecture of the Flemish Open Data Portal, can provide extra features. It can adapt a machine readable file over an HTTP interface (URIs, HTTP caching, HTTP errors, HTTP headers, etc.) towards other machine readable formats.

*3. Data about which both serialization format is known as well as the model (level 3).* In this case, an extra intermediate step will be taken. As long as the model is of interest to be captured, RML language [4] is used to interpret the model in RDF and semantically enrich the data. While in most of the mapping tools, the focus is put on semantically enriching the data as independent resources, RML deals with all resources in a uniform and collective way. The per-source mapping model followed so far gets surpassed, leading to data integration and interlinking at a primary stage. As a result of this mapping, the RDF representation of the data is generated. The generated RDF data are fed in R&Wbase, a Distributed Version Controlled Triple Store. R&Wbase (read: rawbase) [5] is an interpretation layer on top of existing triples stores, that serves interfaces such as SPARQL. It is natively compatible with the SPARQL and reuses the FROM clause to select existing versions by their virtual graph name. Furthermore, it is able to commit to a local branch and can push/pull towards/from a remote server.

For each level of the levels outlined above, the Flemish Open Data Portal (see Figure 1) will thus add extra features to the data, lowering the barrier towards the next level [2].

*Fig. 1: architecture of the Flemish Open Data Portal*

## Three use cases

In order to validate the Flemish Open Data Portal across various sectors and in different context and in order to further broaden the semantic linking across different data types, three demonstrator use cases are being developed: (1) an application that serves as a guide for civil servants and the public, (2) GISStory, a tool for data-driven journalism and, (3) a visualization and exploration tool in the context of scientific research.

Developed from a multi-user point of view, the *first* demonstrator application serves as a guide for civil servants and the public, in case that citizens and companies need to find contact information about services and persons within Flemish local governments and central administrations. It enables civil servants to consult personal contact details from colleagues with the same function, job activity or from those active in the same policy field. For companies and citizens the tool generates generic contact info about services or responsible officials. Going beyond administrative borders and intra- and inter-organizational interoperability issues, the tool can distinguish between e.g. generic and personal phone numbers or email addresses.

The *second* use case demonstrator describes GISStory, an online tool for data-driven journalism. Created for traditional and citizen journalists lacking sufficient statistical or analytical expertise, the app supports data analysis and visualization and in journalists' storytelling processes. The software enables combining and linking different layers of Open Data from Geographical Information Systems into clear visualizations that journalists can use as infographics or as starting points for investigative journalistic articles. For example, a journalist – interested in the relation between spatial planning, floodplains and the socio-economic status of citizens living in a specific region could select, in a user-friendly manner, various datasets from a (Open) data repository containing this information and visualize these layers on top of each other on a map.

The *last* demonstrator use case explores the existence and the nature of Flemish research communities and the diversification of knowledge and expertise within universities and research institutes. Academics and staff can explore their research network and discover the links between themselves and any other researcher. Furthermore, they can monitor emerging communities of practice and interest based on research groups active in a discipline, the collaborations between these research groups and the strengths of their collaborations. Finally, they are able to monitor the evolution of research over time based on corresponding timelines.

## Conclusion

Open Data's potential benefits for economy and society are leading arguments of advocates. However, in order for all affected parties to truly benefit of Open and Linked data, careful attention must be paid to establishing ecosystems around Open Data supply (provision) and demand (re-use & use).

Clearly the value of Open Data arises only, if the data is used; and to be usable, it has to

be accessible, both for machines and humans. Only then is knowledge in society (in very broad terms) fostered. For example, visualizations that demonstrate connections within complex data are a valuable tool in this regard. Providing data in ways that can be harnessed by journalists and citizens to generate more knowledge. They can even contribute to democracy and better governance.

At its core, Open Data essentially concerns efficiency; public-authorities can, for instance, improve their task of public-service-delivery, economizing at the same time. Examples illustrate that involved actors' benefits are mutually dependent, as social value, economic growth, and better public services emerge together. Possible business models for the Flemish Open Data Portal demonstrator applications described above could revolve around data-broker roles for private players (handling technical operations, providing actual accessibility, or adding value to data through enriching, sorting etc.). Still, revenue generation based on Open Data is not straightforward, and true economic potential is hard to measure.

Open Data is still in its early days, but it is undeniable that its hailed socio-economic potential depends on well-balanced ecosystems and their sustainability. Summarizing the contribution of the Flemish Open Data Portal and the experimental applications being developed: with the CKAN integration, the data is made discoverable, with The DataTank, the barrier for data owners to publish machine readable data is lowered, with RML, the quality of the semantic annotations can be improved and with R&Wbase, the different versions of the data is captured. Together, these components offer the technical potential to reach new data governance models.

## Acknowledgements

## References

1: Boris Villazón-Terrazas, Daniel Vila-Suero, Daniel Garijo, Luis M. Vilches-Blázquez, María Poveda-Villalón, José Mora, Oscar Corcho and Asunción Gómez-Pérez, Publishing Linked Data - There is no One-Size-Fits-All Formula, 2013
2: Pieter Colpaert, Sarah Joye, Peter Mechant and Erik Mannens, The 5 stars of Open Data Portals, 2013
3: Miel Vander Sande, Pieter Colpaert, Davy Van Deursen, Erik Mannens, Rik Van de Walle, The DataTank: an open data adapter with semantic output., 2012
4: Anastasia Dimou, Miel Vander Sande, Pieter Colpaert, Erik Mannens, Rik Van de Walle, Extending R2RML to a source-independent mapping language for RDF, 2013
5: Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Sam Coppens, Erik Mannens, and Rik Van de Walle, R&Wbase: git for triples, 2013