DataGraft: A Platform for Open Data Publishing

Dumitru Roman[1], Marin Dimitrov[2], Nikolay Nikolov[1], Antoine Putlier[1], Brian Elvesæter[1], Alex Simov[2], Yavor Petkov[2]

[1]SINTEF, Forskningsveien 1a, 0373 Oslo, Norway
{firstname.lastname}@sintef.no
[2]Ontotext AD, Tsarigradsko Shosse 47A, 1784 Sofia, Bulgaria
{firstname.lastname}@ontotext.com

Abstract. DataGraft is a platform for Open Data management. It has the goals to simplify and speed up the data publishing process and to improve the reliability and scalability of the data consumption process. This demonstrator provides a summary of the key features of the current DataGraft platform as well as simple demo scenario from the domain of property-related data.

1 Introduction

DataGraft has the goal of providing tools and approaches for easier and lower-cost publication and reuse of Open Data (and Linked Data in particular). The lifecycle for publishing Open Data typically involves data *cleaning & transformation* (most often from tabular formats), *mapping* to standard Linked Data models and *generating a semantic RDF graph*. The resulting semantic graph is *stored in a triple store*, so that applications and services can easily access and query the data. While this process is rather straightforward, publishing and consuming of (linked) Open Data still remains a complex and time consuming task due to a variety of reasons:

1. The *technical complexity* of preparing Open Data for publication is high – toolkits are poorly integrated and often require expert knowledge;
2. There is a *considerable cost* for publishing data and providing reliable access to it. The required expertise & resources often become excessively high for many non-profit organisations;
3. The *poorly maintained and fragmented supply* of Open Data: datasets are usually provided through disconnected channels; inconsistently formatted and structured; poorly maintained.

2 The DataGraft Platform

DataGraft[1] provides a cloud-based platform for open Data publishing. Its key features are:

[1] http://datagraft.net/

- *Interactive design of data transformations*: transformations provide feedback to publishers on how data changes;
- *Repeatable data transformations*: data transformation processes often need to be repeatedly executed as new data arrives. Executable and repeatable transformations are a key requirement for a low cost data publication process;
- *Shareable and reusable data transformations*: Capabilities to reuse and extend data transformations created by other developers further lowers the data publication cost;
- *Reliable data access*: provisioning data reliably is another key aspect for the 3rd party data services and applications built on top of Open Data.

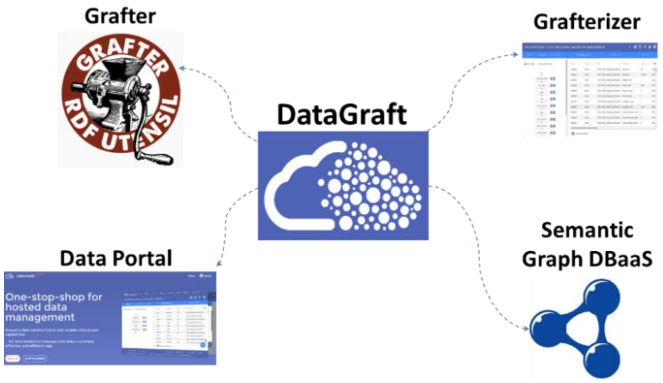


Fig. 1. Key DataGraft components

The key enablers of DataGraft are shown in **Fig. 1**. *Grafter*[2], which is an open source framework of reusable components designed to support complex data transformations. Grafter provides a domain-specific language (DSL), which allows the specification of transformation pipelines that convert tabular data or produce linked data graphs. The main advantages of Grafter over similar ETL frameworks include: 1) efficient support for very large datasets, due to its streaming approach for data processing; 2) its highly modular and extensible design; 3) the ability to serialize and execute transformations as services in a sandboxed environment.

Grafterizer is an open source web-based frontend for data cleaning and transformation built on top of Grafter. It provides an interactive user interface that supports the data transformation process: 1) forking of existing data transformations; 2) creating complex data transformation workflows by combining and configuring data transformation steps; and 3) live preview of the data transformation over sample data.

Another key enabler is the *semantic Graph Database-as-a-Service* (DBaaS) triple store, which is used for accessing the Linked Data on the platform. With a database-as-a-service solution, data publishers do not need to deal with administrative overheads such as installation, upgrades and maintenance, provisioning, etc. From the point of view of a data publisher or a data consumer, the DBaaS provides standard

[2] http://grafter.org/

APIs and endpoints for Linked Data access, querying, and management. These functionalities are based on a complex cloud architecture, which ensures the database scalability, extensibility and availability on large scale [1].

Finally, the *Open Data portal* integrates the components together in a web-based interface. The entire process of publishing data is reduced to a simple wizard-like interface, where publishers can simply drop their data and enter some basic metadata. Currently, the platform provides a number of visualization widgets, including tables, line charts, bar charts, pie charts, scatter charts, bubble charts and maps.

3 Demo Scenario: Publishing Property-related Data

The simple demonstration scenario will highlight the capabilities of the DataGraft platform: transforming data by the State of Estate service for state-owned properties in Norway and publishing the data as Linked Data. The scenario workflow is summarised in **Fig. 2**.

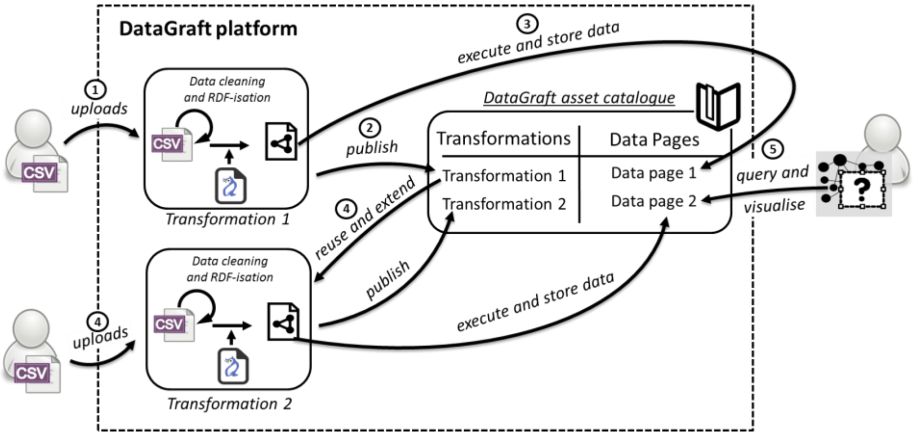


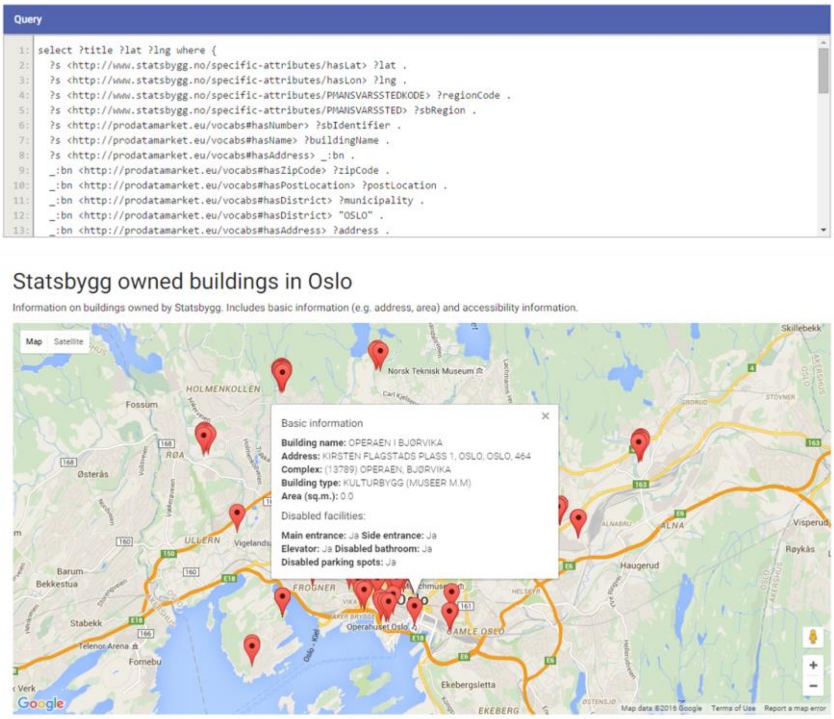
Fig. 2. Demo scenario

The scenario will demonstrate:

1. Interactive specification of tabular data transformations and mapping of tabular data to graph data (Linked Data);
2. Publication of data transformations on the DataGraft asset catalogue;
3. Execution and storage of transformed data on the semantic graph database-as-a-service on DataGraft;
4. Sharing, reusing and extending user-generated content;
5. Querying published data from the live endpoint and visualising query results (**Fig. 3**).

A visitor of the demonstration will learn how to:

- Use DataGraft to for simple data transformation and publishing;
- Easily create data transformations through the DataGraft’s GUI;
- Share and reuse data transformations already published in DataGraft;
- Run data transformations and publish the resulting data on DataGraft’s cloud-based semantic graph database;
- Access and query data published on DataGraft;
- Use DataGraft for real life applications (publishing property data).



Query

```
select ?title ?lat ?lng where {
  ?s <http://www.statsbygg.no/specific-attributes/hasLat> ?lat .
  ?s <http://www.statsbygg.no/specific-attributes/hasLon> ?lng .
  ?s <http://www.statsbygg.no/specific-attributes/PMANSVARSSTEDKODE> ?regionCode .
  ?s <http://www.statsbygg.no/specific-attributes/PMANSVARSSTED> ?sbRegion .
  ?s <http://prodatamarket.eu/vocabs#hasNumber> ?sbIdentifier .
  ?s <http://prodatamarket.eu/vocabs#hasName> ?buildingName .
  ?s <http://prodatamarket.eu/vocabs#hasAddress> _:bn .
  _:bn <http://prodatamarket.eu/vocabs#hasZipCode> ?zipCode .
  _:bn <http://prodatamarket.eu/vocabs#hasPostLocation> ?postLocation .
  _:bn <http://prodatamarket.eu/vocabs#hasDistrict> ?municipality .
  _:bn <http://prodatamarket.eu/vocabs#hasDistrict> "OSLO" .
  _:bn <http://prodatamarket.eu/vocabs#hasAddress> ?address .
```

Statsbygg owned buildings in Oslo

Information on buildings owned by Statsbygg. Includes basic information (e.g. address, area) and accessibility information.

Basic information

Building name: OPERAEN I BJØRVIKA
Address: KIRSTEN FLAGSTADS PLASS 1, OSLO, OSLO, 464
Complex: (13789) OPERAEN, BJØRVIKA
Building type: KULTURBYGG (MUSEER M.M)
Area (sq.m.): 0.0

Disabled facilities:

Main entrance: Ja **Side entrance:** Ja
Elevator: Ja **Disabled bathroom:** Ja
Disabled parking spots: Ja

Fig. 3. Data query and visualization in DataGraft

DataGraft is available via http://datagraft.net/ and further details can be found in [2].

4 Ongoing Work

DataGraft is currently under active development within the *proDataMarket* project[3] and new features and improvements are being added to the live platform on a regular basis.

Various new DataGraft features are already in development or planned to be delivered within the next 12 months:

[3] http://prodatamarket.eu/

- Extending the data hosting platform towards data science and analytics, with the ability to configure and run simple analytics directly on the platform (rather than downloading data and running the analytics locally);
- Ability to interlink the generated Linked Data to existing datasets in a semi-automated manner;
- Dealing with data streams (rather than static input data files);
- Extensions towards working with large geo-spatial datasets and queries;
- Ability to share and reuse other assets, such as data queries or visualization widgets;
- Improved error reporting in data transformations;

Acknowledgements. This work was partly funded by the European Commission within the following research projects: *DaPaaS* (FP7 610988), *SmartOpenData* (FP7 603824), *InfraRisk* (FP7 603960), and *proDataMarket* (H2020 644497).

References

1. M. Dimitrov, A. Simov, and Y. Petkov. *Low-cost Open Data As-a-Service in the Cloud*. In proceedings of the 2nd Semantic Web Developers Workshop (SemDev 2015), part of the Extended Semantic Web Conference (ESWC 2015), May 31st 2015, Portoroz, Slovenia.
2. D. Roman, N. Nikolov, A. Putlier, D. Sukhobok, B. Elvesæter, A. Berre, X. Ye, M. Dimitrov, A. Simov, M. Zarev, R. Moynihan, B. Roberts, I. Berlocher, S. Kim, T. Lee, A. Smith, and T. Heath. DataGraft: *One-Stop-Shop for Open Data Management*. Technical Report, January 2016. Available at http://www.semantic-web-journal.net/system/files/swj1285.pdf.