DataGraft beta v2: New Features and Capabilities

Nikolay Nikolov[1], Dina Sukhobok[1], Stefan Dragnev[2], Steffen Dalgard[1], Brian Elvesæter[1], Bjørn Marius von Zernichow[1] and Dumitru Roman[1]

[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. In this demonstrator, we will introduce the latest features and capabilities added to DataGraft – a Data-as-a-Service platform for data preparation and knowledge graph generation. DataGraft provides data transformation, publishing and hosting capabilities that aim to simplify the data publishing lifecycle for data workers (i.e., Open Data publishers, Linked Data developers, data scientists). This demonstrator highlights the recent features added to DataGraft by exemplifying data publication of statistical data – going from the raw data published at a public portal to published and accessible Linked Data with the help of the tools and features of the platform.

Keywords: open data, linked data, data publication, data transformation, data management

1 Introduction

In the recent years, a large number of public organisations and governments have been increasingly committed to publishing data in reusable formats and under open licenses. Nevertheless, in many cases such organisations choose to prioritise e-government and open government initiatives due, to a large extent, to the high costs and domain-specific expertise required. DataGraft[1] started as an initiative with the goal to alleviate such obstacles by providing new tools and services for faster and lower-cost publication of Open Data. The lifecycle for the creation and provisioning of (Linked) Open Data typically involves raw data cleaning, transformation, and preparation (most often from tabular formats), mapping to standard Linked Data vocabularies and generating a semantic RDF graph. The resulting semantic graph is then stored in a triple store, where applications can reliably access and query the data. Conceptually, this process is rather straightforward; however, such an integrated workflow is not commonly implemented. Instead, publishing and consuming (Linked) Open Data remains a tedious task due to a variety of reasons, including:

[1] https://datagraft.io/

1. The *technical complexity* of preparing Open Data for publication is high – toolkits are poorly integrated and require expert knowledge, particularly for publishing of Linked Data;
2. There is *considerable cost* for publishing data and providing reliable access to it. In the absence of clear monetisation channels and cost recovery incentives, the relative investment costs can easily become excessively high for many organisations;
3. The *poorly maintained and fragmented supply* of Open Data reduces the reuse of data: datasets are often provided through disconnected outlets; sequential releases of the same dataset are often inconsistently formatted and structured.

DataGraft features and capabilities have been developed to tackle such Open Data challenges.

2 The DataGraft Platform (beta v2)

DataGraft aims to streamline the process of publishing, managing and accessing Open Data and related artefacts in order to support data workers and the users of their data. The beta v2 version of DataGraft supports the following features:

- Asset management – metadata, access management (public/private), sharing – for files, SPARQL endpoints, data transformations, queries;
- Automatic reliable storage of files and RDF databases;
- Data hosting and access services for all data stored on DataGraft;
- Interactive design of data transformations and RDF generation;
- Interactive visual exploration of RDF data in the SPARQL endpoints published;
- REST-based access to all platform assets;
- Layered security – encrypted user login information and SSL; asset security using OAuth2, API keys for semantic graph databases.

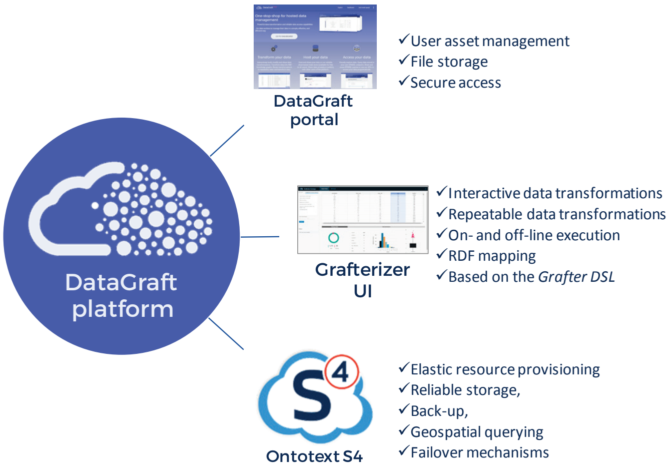


Fig. 1. Main components of DataGraft platform

The DataGraft platform mainly consists of three components – the DataGraft portal, Grafterizer and a cloud-enabled semantic graph database-as-a-service (as shown on **Fig. 1**), which is based on a dedicated instance of the Ontotext S4 platform[2].

Since the initial release of the DataGraft platform reported in [1][2], a number of new features and improvements have been implemented as part of the new release (beta v2):

- New asset types for the catalogue – SPARQL endpoints, queries and file pages sharing between users of the platform;
- Improved Grafterizer capabilities – including conditional RDF mappings, support various types and formats of tabular inputs;
- Versioning of assets – browsing, recording of provenance when copying assets;
- Visual browsing of data from SPARQL endpoints (using RDF Surveyor[3]);
- New dashboard with more control of user assets, instant search and various filters;
- Improved security (authentication) using OAuth2;
- REST API re-implemented using Swagger;
- Updated version of the semantic graph database, which now support geospatial queries and serialisation to GeoJSON.

Related Work. In the current state-of-the art there are several software tool ecosystem solutions that provide support for publication of Linked Data (data extraction, RDF-isation, storage, access). Examples of such are Linked Data Stack[4], LinDA[5], and COMSODE[6] projects, which provide software tools and methodologies for Linked Data management. Such tools come functionally close to the features provided by DataGraft, but are not provided "as-a-service", which puts the burden of deploying and managing the infrastructure on the data publisher. OpenRefine[7], with its RDF plugin (including powerful RDF mapping features such as RDF reconciliation) comes closest to the data transformation approach implemented in DataGraft. Nevertheless, OpenRefine is unsuitable for usage in a service offering and the data transformation engine is memory intensive with large data volumes.

3 Demo Scenario: Statistical Data Publication and Access

The usage scenario will demonstrate the core capabilities of DataGraft using statistical data from Statistics Norway[8]. Firstly, it will demonstrate how to use the *DataGraft dashboard* to explore published data and user assets using the new filters and search

[2] https://console.s4.ontotext.com

[3] https://github.com/guiveg/rdfsurveyor

[4] http://stack.linkeddata.org

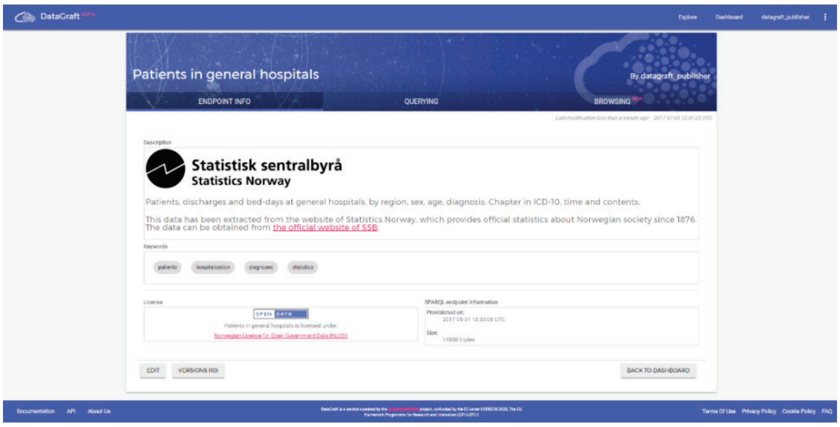
[5] http://linda-project.eu

[6] http://www.comsode.eu

[7] http://openrefine.org

[8] https://www.ssb.no/

functionalities. We will then show the different types of assets that are now supported by the platform along with the features related to browsing assets versions. Furthermore, we will demonstrate how to use the DataGraft wizards to publish Linked Data based on raw tabular data with the improved Grafterizer tool. Grafterizer allows to interactively specify data transformations and mappings of tabular data to RDF. We will demonstrate how transformations can be applied on-the-fly or offline. Then the resulting Linked Data will be published in a dynamically created semantic graph database and registered in the DataGraft catalogue in a SPARQL endpoint page (such as the one shown in **Fig. 2**). We will demonstrate how the endpoint can be explored using user-defined queries (as free-text or using pre-defined ones) or the new browsing tool integrated in DataGraft. Finally, we will showcase how the new interactive APIs can be directly used by users to access data hosted on the platform.



Patients in general hospitals

By datagraft_publisher

Last modification less than a minute ago · 2017 07 03 12:31:25 UTC

Description

Statistisk sentralbyrå
Statistics Norway

Patients, discharges and bed-days at general hospitals, by region, sex, age, diagnosis: Chapter in ICD-10, time and contents.

This data has been extracted from the website of Statistics Norway, which provides official statistics about Norwegian society since 1876. The data can be obtained from the official website of SSB.

Keywords

patients · hospitalisation · diagnoses · statistics

License

OPEN DATA

Patients in general hospitals is licensed under:
Norwegian Licence for Open Government Data (NLOD)

SPARQL endpoint information

Provisioned on:
2017 05 31 13:33:08 UTC

Size:
14000 triples

Fig. 2. SPARQL endpoint page in DataGraft

As of September 2017, DataGraft is available via http://datagraft.io and more details can be found in the platform online documentation[9].

Acknowledgements. This work is partly funded by the EC H2020 project proDataMarket (Grant number: 644497).

References

1. Roman, D., Nikolov, N., Putlier, A., Sukhobok, D., Elvesæter, B., Berre, A., Ye, X., Dimitrov, M., Simov, A., Zarev, M., Moynihan, R., Roberts, B., Berlocher, I., Kim, S., Lee, T., Smith, A., & Heath, T. (2017). DataGraft: One-Stop-Shop for Open Data Management. (2017). DOI:10.3233/SW-170263. Journal: Semantic Web, vol. Preprint, no. Preprint, pp. 1-19.
2. Roman, D., Dimitrov, M., Nikolov, N., Pultier, A., Sukhobok, D., Elvesæter, B., & Petkov, Y. (2016, May). DataGraft: Simplifying open data publishing. ESWC (Satellite Events) 2016: 101-106.

[9] https://github.com/datagraft/datagraft-reference/blob/master/documentation.md