EU H2020 Research and Innovation Project

HOBBIT – Holistic Benchmarking of Big Linked Data

Project Number: 688227 Start Date of Project: 01/12/2015 Duration: 36 months

Deliverable 8.5.2
Intermediate Data Management Plan

Dissemination Level	Public
Due Date of Deliverable	Month **18**, 31/05/2017
Actual Submission Date	Month **18**, 28/05/2017
Work Package	WP 8
Task	T 8.1
Type	Report
Approval Status	Final
Version	**0.9**
Number of Pages	10
Filename	D8.5.2_IntermediateDataManagementPlan .pdf

Abstract: This report describes the intermediate data management plan for the project.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability.

Project funded by the European Commission within the H2020 Framework Programme

History

Version	Date	Reason	Revised by
0.9	19/05/2017	Draft	Ruben Taelman

Author List

Organisation	Name	Contact Information
imec	Ruben Taelman	ruben.taelman@ugent.be

Executive Summary

This report improves and extends D8.5.1, which described the initial data management plan.

This report describes the intermediate data management plan. This plan is used as a guideline when handling the data submitted by members of the HOBBIT community to the benchmarks.

In this document, we discuss the data management lifecycle to answer questions like: how can data be added to the platform; how can it be accessed; and how long will it be kept. The data management plan is detailed as it has been agreed upon by the consortium at the time of publication of this report.

Most important changes:

- Adhere to the EU DCAT-AP profile in RDF serializations;
- Abandoned the more heavy requirement of a SPARQL endpoint in favor of a lightweight queryable linked data interface, triple pattern fragments (TPF);
- Refined the section on archiving and preservation;
- More details on the group management within the HOBBIT platform;
- Current status update.

Table of Contents

List of Figures

1. Data Management Lifecycle

HOBBIT will continuously collect datasets (i.e., not limited to specific domains) as the base for benchmarks. Those datasets are provided by both the project industrial partners and members of the HOBBIT community.

To make the data **discoverable** and **accessible**, besides providing the datasets as **dump files** that can be loaded from the project repository, HOBBIT will also provide a **queryable interface** that will serve all the dataset metadata. This interface will enable the platform users to run their own queries against the dataset metadata to obtain tailored datasets that fit exactly each user needs. While we initially planned to setup a SPARQL endpoint to host this metadata, we now opt for a **Triple Pattern Fragments interface**[1] that allows us to publish this data at a lower cost, while still enabling the metadata to be queried (using SPARQL queries).

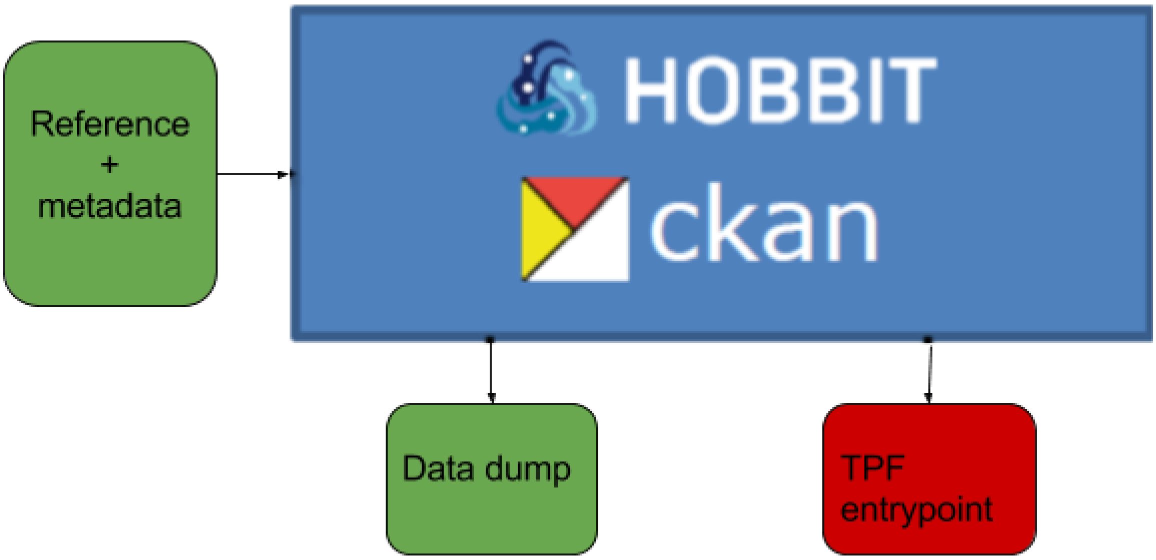


Figure 1. Data Management Lifecycle Overview

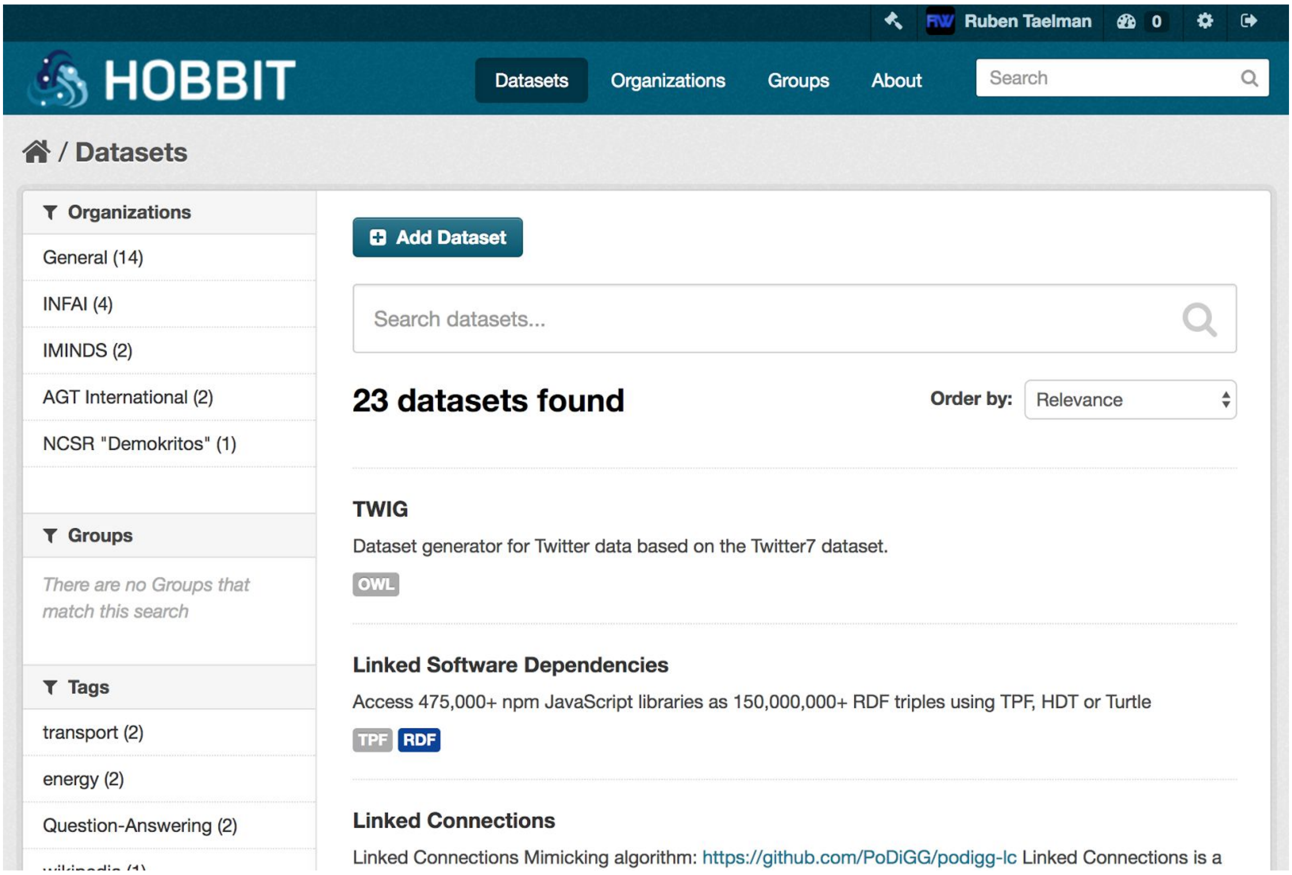
To **keep the dataset submission process manageable**, we host an instance of the CKAN open source data portal software, extended with custom metadata fields for the HOBBIT project. This instance is hosted at http://hobbit.iminds.be. Because the CKAN instance only stores *metadata* about the datasets, the datasets themselves need to be stored elsewhere, such as the HOBBIT FTP storage. Users who want to add a dataset of their own, first need to request[2] to be added to an organization on the CKAN instance, after which they can add datasets to this organization. If users have no storage available for their dataset, they can add their dataset to the HOBBIT FTP server after contacting us.

Datasets will be kept available on the HOBBIT platform for **at least the lifetime of the project**, unless they are removed by their owners. After the project, the HOBBIT platform will be

[1] http://linkeddatafragments.org/in-depth/#tpf

[2] http://project-hobbit.eu/contacts/

maintained by the HOBBIT Association, and so will the datasets. **Owners may add or remove** a dataset at any time.



23 datasets found

TWIG

Dataset generator for Twitter data based on the Twitter7 dataset.

OWL

Linked Software Dependencies

Access 475,000+ npm JavaScript libraries as 150,000,000+ RDF triples using TPF, HDT or Turtle

TPF RDF

Linked Connections

Linked Connections Mimicking algorithm: https://github.com/PoDiGG/podigg-lc Linked Connections is a

Figure 2. Screenshot of the current CKAN deployment.

2. Data Management Plan

Conform to the guidelines of the Commission, we will provide the following information for every dataset submitted to the project. This information will be obtained either through automatically generating it (e.g., for the identifier), or by asking whoever provides the dataset upon submission.

2.1. Dataset reference and name

The datasets submitted will be identified and referenced by using a URL. This URL can then be used to access the dataset (either through dump file, TPF entrypoint or SPARQL endpoint), and also be used as an identifier to provide metadata.

2.2. Data set description

The submitter will be asked to provide a short textual, human-interpretable description of the dataset, at least in English, and optionally in other languages as well. Additionally, a machine-interpretable description will also be provided (see 2.3 Standards and metadata).

2.3. Standards and metadata publication

Since we are dealing with Linked Datasets, it makes sense to adhere to a Semantic Web context for the description of the datasets as well. Therefore, in line with the application profile for metadata catalogues in the EU, DCAT-AP, we will use W3C recommended vocabularies such as DCAT and Dublin Core to provide metadata about each dataset. The metadata that is currently associated with the datasets includes:

- Title
- URL
- Description
- External Description
- Tags
- License
- Organization
- Visibility
- Source
- Version
- Contact
- Contact Email
- Applicable Benchmark[3]

Currently, this metadata is stored in the CKAN instance's database. However, the plan is to convert this information to RDF and make it available for querying using a publication pipeline.

[3] Part of the custom metadata

This pipeline, which will be run daily, consists of the following steps:

1. Using a CKAN plugin[4], we export the CKAN data to RDF. For instance, in the Turtle format. The CKAN plugin has built-in support to serialize the data following the EU DCAT-AP profile[5]. If needed, profiles can be added or adjusted[6], which may be required for our custom metadata.
2. Converting the raw RDF dump to HDT[7], which reduces storage requirements for the CKAN data.
3. Publishing the HDT file through a TPF interface, allowing us to publish the data at a low cost.

In case we notice that the raw RDF dump is relatively small, i.e. compressing it won't achieve significant storage reductions, we can skip the HDT conversion and directly expose the raw dump using TPF.

2.4. Data Sharing

Industrial companies are normally unwilling to make their internal data available for competitions because this could reduce the competitiveness of these companies significantly. However, HOBBIT aims to pursue a policy of making data **open, as much as possible**. Therefore, a number of mechanisms are put in place.

As per the original proposal, HOBBIT will deploy a standard data management plan that includes (1) employing **mimicking algorithms** that will compute and reproduce variables that characterize the structure of company-data, (2) feeding these characteristics into **generators that will be able to generate data similar to real company data** without having to make the real company data available to the public. The mimicking algorithms will be implemented in such a way that can be used within companies and simply return parameters that can be used to feed the generators. This preserves Intellectual Property Rights (IPR) and will circumvent the hurdle of making real industrial data public by allow configuring deterministic synthetic data generators so as to compute data streams that display the same variables as industry data while being fully open and available for evaluation without restrictions.

Since we will provide a mimicked version of the original dataset in our benchmarks, **open access will be the default behaviour**. However, on a case-by-case basis, datasets might be **protected** (i.e., visible only to specific user groups) on request of the data owner, and in agreement with the HOBBIT platform administrators.

[4] https://github.com/ckan/ckanext-dcat

[5] The properties that relevant for CKAN are documented by the plugin. There is a default mapping: suggestion https://github.com/ckan/ckanext-dcat#rdf-dcat-to-ckan-dataset-mapping. Depending on the profile different sets can be serialized. In this case, we opt for the EU DCAT-AP profile.

[6] https://github.com/ckan/ckanext-dcat#writing-custom-profiles

[7] HDT is a compressed RDF format. http://www.rdfhdt.org/what-is-hdt/

2.5. Archiving and preservation (including storage and backup)

HOBBIT will also support the functionality of accessing and querying past versions of previous dataset metadata, where all different dataset metadata versions will be publically available as frequent[8] dump files (Turtle[9] and HDT). These daily dumps will be kept for at least until the maximum amount of storage is reached. After that, for the oldest daily dumps, monthly dumps are provided. At least from the last month daily dumps will remain available. The data will be stored on the CKAN server(s), at least for the duration of the project. After the project, this responsibility is transferred to the HOBBIT Association, who will be tasked with the long term preservation of the datasets.

2.6. Current Status

As described in the initial data management plan, all organizations are available on the CKAN instance: https://hobbit.iminds.be/organization
Each **organization** has made some of their datasets available, either publicly, or only with the consortium for sensitive data.

Furthermore, we have created a **"general"** group[10] on CKAN in which we added several datasets that are useful for benchmarking but are not owned by partners of the HOBBIT project.
The number of datasets has been increased to 23 datasets, of which half are RDF datasets. 21 of those datasets are publicly available under an open license.

The implementation of the publication pipeline of dataset metadata and its configuration are a work in progress and will be finished at the latest by M36, i.e., the final version of the data management plan.

[8] Either daily, or upon changes in the repository
[9] Can be compressed using gzip
[10] https://hobbit.iminds.be/organization/general