HOBBIT
Holistic Benchmarking
of Big Linked Data

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 8.5.3
Final Data Management Plan

Dissemination Level	Public
Due Date of Deliverable	Month 36, 30/11/2018
Actual Submission Date	Month 36, 27/11/2018
Work Package	WP8
Task	T8.1
Type	Report
Approval Status	Final
Version	1.0
Number of Pages	10
Filename	D8.5.3_Final_Data_Management_Plan.pdf

Abstract: This report describes the final 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.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688227.

History

Version	Date	Reason	Revised by
0.1	05/11/2018	Draft	Ruben Taelman
0.2	07/11/2018	Peer review	Carolin Walter
1.0	09/11/2018	Revisions included and final version created	Ruben Taelman

Author List

Organisation	Name	Contact Information
imec	Ruben Taelman	ruben.taelman@ugent.be
USU	Carolin Walter	c.walter@usu.de

Executive Summary

This report is a final version of D8.5.2, which described the intermediate data management plan. It describes the final data management plan. This plan summarizes the work done on 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 can 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:

- Refine actual work done;
- Removal of query interface;
- Current status update.

Table of Contents

List of Figures

1. Data Management Lifecycle

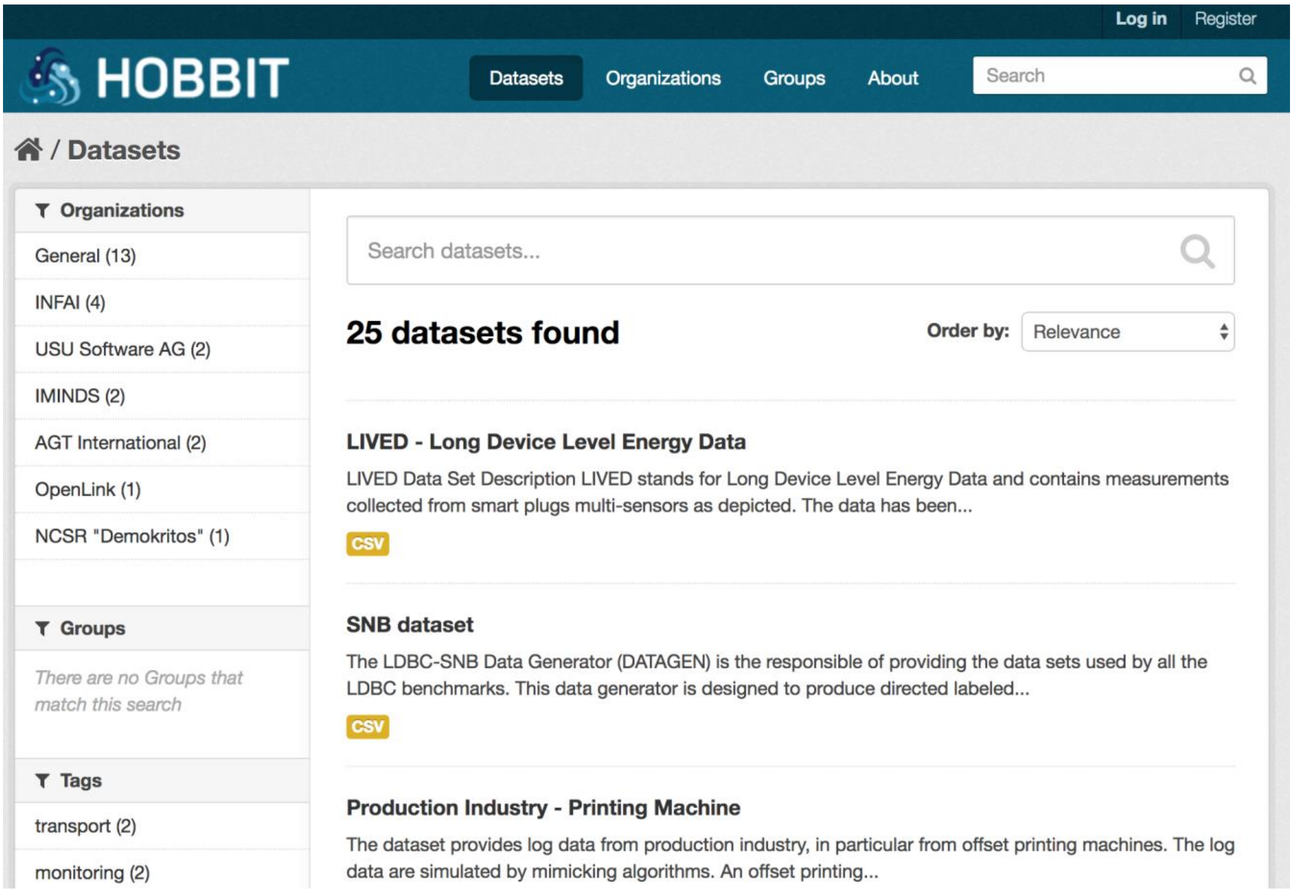
HOBBIT continuously collects datasets (i.e., not limited to specific domains) as the base for benchmarks. Those datasets are provided by both the project industrial partners and members of the HOBBIT community.

To **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 https://hobbit.ilabt.imec.be/. Figure 1 shows an screenshot of this CKAN instance, where several datasets are listed. 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[1] 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. Because of this, storage requirements in this CKAN instance are limited, which is why no data deletion strategy is needed.

Datasets will be kept available on the HOBBIT platform for **at least the lifetime of the server**, unless they are removed by their owners. After the project, the HOBBIT platform will be maintained by the HOBBIT Association, and so will the datasets. **Owners may add or remove** a dataset at any time.

In the previous version of this deliverable, we described a query interface that was to be setup over the metadata of this CKAN instance. As there was no need for such a query interface, both inside and outside of the project, and the setup would be non-trivial, we removed this interface.

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



25 datasets found

LIVED - Long Device Level Energy Data

LIVED Data Set Description LIVED stands for Long Device Level Energy Data and contains measurements collected from smart plugs multi-sensors as depicted. The data has been...

CSV

SNB dataset

The LDBC-SNB Data Generator (DATAGEN) is the responsible of providing the data sets used by all the LDBC benchmarks. This data generator is designed to produce directed labeled...

CSV

Production Industry - Printing Machine

The dataset provides log data from production industry, in particular from offset printing machines. The log data are simulated by mimicking algorithms. An offset printing...

Figure 1: 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 it can 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).

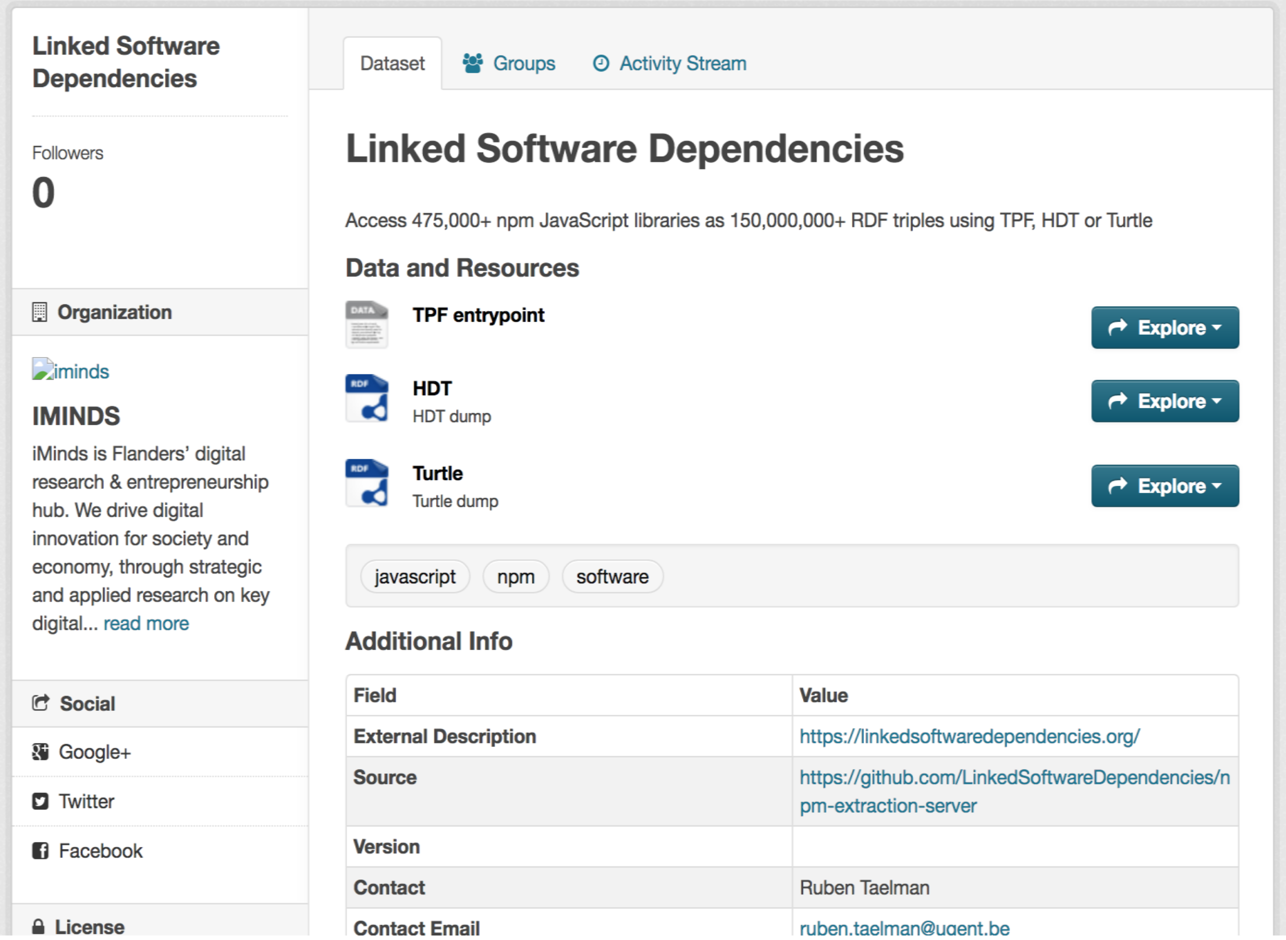
2.3. Standards and Metadata Publication

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[2]

This metadata is stored in the CKAN instance's database, and can be view on the dataset overview page, as shown in Figure 2.

[2] Part of the custom metadata



Linked Software Dependencies

Followers

0

Organization

iminds

IMINDS

iMinds is Flanders' digital research & entrepreneurship hub. We drive digital innovation for society and economy, through strategic and applied research on key digital... read more

Social

Google+

Twitter

Facebook

License

Linked Software Dependencies

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

Data and Resources

TPF entrypoint

HDT
HDT dump

Turtle
Turtle dump

javascript npm software

Additional Info

Field	Value
External Description	https://linkedsoftwaredependencies.org/
Source	https://github.com/LinkedSoftwareDependencies/npm-extraction-server
Version	
Contact	Ruben Taelman
Contact Email	ruben.taelman@ugent.be

Figure 2: Screenshot of a dataset overview page, with the collected metadata.

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, several mechanisms are put in place.

As per the original proposal, HOBBIT deploys a standard data management plan that includes (1) employing **mimicking algorithms** that compute and reproduce variables that characterize the structure of company-data, (2) feeding these characteristics into **generators that are able to generate data similar to real company data** without having to make the real company data available to the public. The mimicking algorithms are 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 circumvents 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 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 are **protected** (i.e., visible only to specific user groups) on request of the data owner, and in agreement with the HOBBIT platform administrators.

2.5. Current Status

The domain name has been changed to https://hobbit.ilabt.imec.be/, due to internal organization changes in imec. As described in the intermediate data management plan, all organizations are available on the CKAN instance: https://hobbit.ilabt.imec.be/organization

Each **organization** made their datasets available, either publicly, or only with the consortium for sensitive data. The number of datasets has been increased to 25 datasets, of which half are RDF datasets. 23 of those datasets are publicly available under an open license. The server behind this CKAN instance will remain active for at least one year after the project ends. In this period, ownership will be transitioned to the HOBBIT Association.