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 8.5.1
Initial Data Management Plan

Dissemination Level	Public
Due Date of Deliverable	Month 6, 31/05/2016
Actual Submission Date	Month 6, 31/05/2016
Work Package	WP 8
Task	T 8.1
Type	Report
Approval Status	Approved
Version	1.0
Number of Pages	9
Filename	D8.5.1_Initial_Data_Management_Plan.pdf

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

Project funded by the European Commission within the H2020 Framework Programme

History

Version	Date	Reason	Revised by
0.1	09/02/2016	First Draft	Tom De Nies
0.2	09/05/2016	Second Draft	Tom De Nies
0.3	13/05/2016	Review	Marco Huber
0.4	29/05/2016	Version after 1st Review	Tom De Nies
0.5	29/05/2016	2nd Review	Axel Ngonga
1.0	29/05/2016	Final Version after Review	Tom De Nies

Author List

Organisation	Name	Contact Information
iMinds	Tom De Nies	tom.denies@ugent.be
USU Software	Marco Huber	m.huber@usu-software.de
InfAI	Axel Ngonga	ngonga@informatik.uni-leipzig.de

Executive Summary

This report describes the initial data management plan for the project. This data management plan will be used as a guideline when handling the data submitted by members of the HOBBIT community to the benchmarks.

In this initial plan, we discuss the envisioned data management lifecycle (how can data be added to the platform, how can it be accessed, and how long will it be kept?), as well as the details of the data management plan as they have been agreed upon by the consortium at this time.

Table of Contents

List of Figures

1. Data Management Lifecycle

HOBBIT will continuously collect various datasets (i.e., not limited to specific domains) as the base for benchmarks. Those data will initially be provided by the project industrial partners, and later on by members of the HOBBIT community.

To make the data **discoverable** and **accessible**, besides providing the generated benchmarks as **dump files**[1] that can be loaded from the project repository, HOBBIT will also provide a **SPARQL endpoint** that will serve all the benchmark datasets. The HOBBIT SPARQL endpoint will enable the platform users to run their own queries against one or more benchmark(s) to obtain tailored benchmark(s) that fit exactly each user needs.

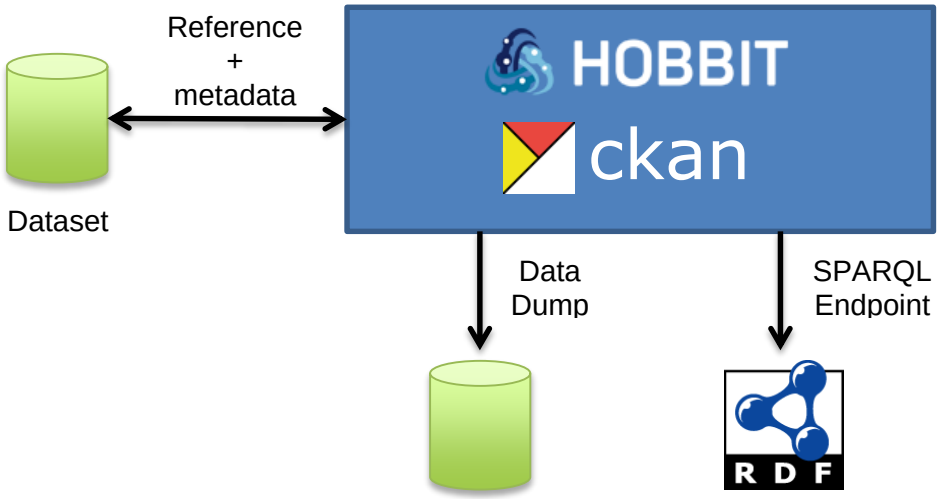

Figure 1. Data Management Lifecycle Overview

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. For the time being, this instance is hosted at http://hobbit.iminds.be. When the benchmarking platform itself goes online, the CKAN instance will be moved there, to accommodate more space for datasets. Users who want to add a dataset of their own, first need to request to be added to an organization on the CKAN instance, after which they can add datasets to this organization. http://project-hobbit.eu/contacts/

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

[1] The file format has not been decided upon yet, and will depend on the results of the first requirements investigation and survey of the stakeholders.

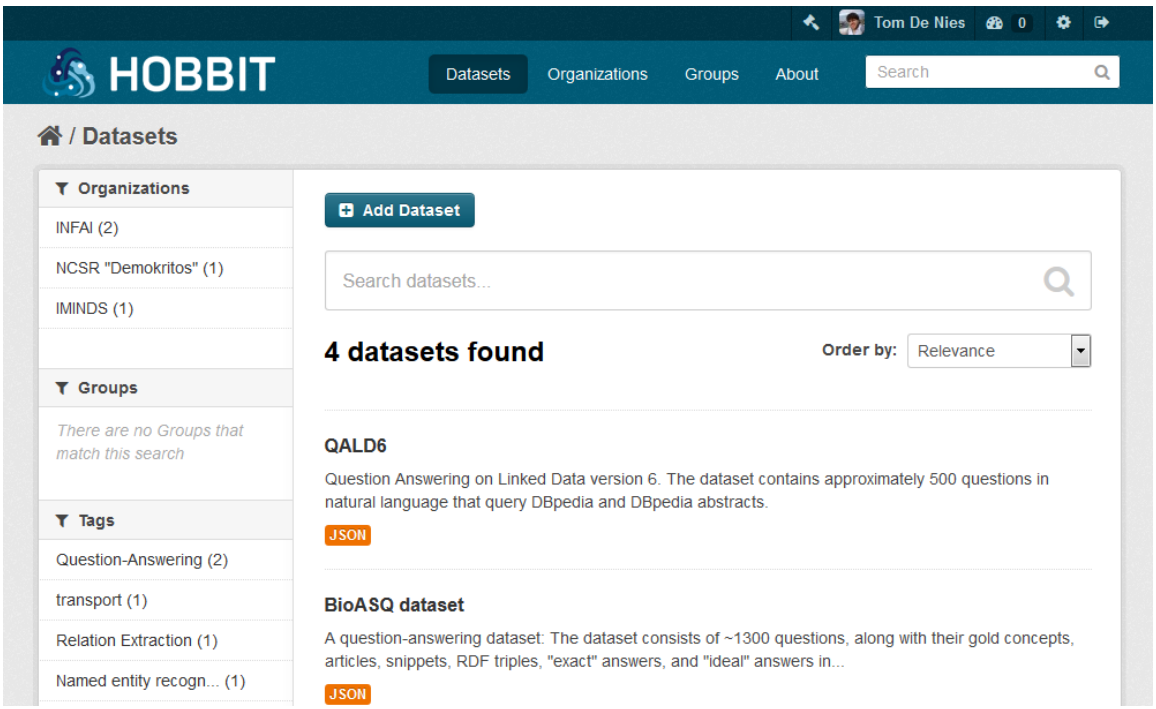

Figure 2. Screenshot of the current CKAN deployment.

2. Data Management Plan

In conformity with 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 references by using a URL. This URL can then be used to access the dataset (either through dump file or SPARQL endpoint), and also be used as an identifier to provide metadata.

2.2 Dataset 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

Since we are dealing with Linked Data sets, it makes sense to adhere to a Semantic Web context for the description of the datasets as well. Therefore, we will use W3C recommended vocabularies such as DCAT 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

Currently, this metadata is stored in the CKAN instance's database. However, the plan is to convert this information to DCAT and make it available for querying once the benchmarking platform is running.

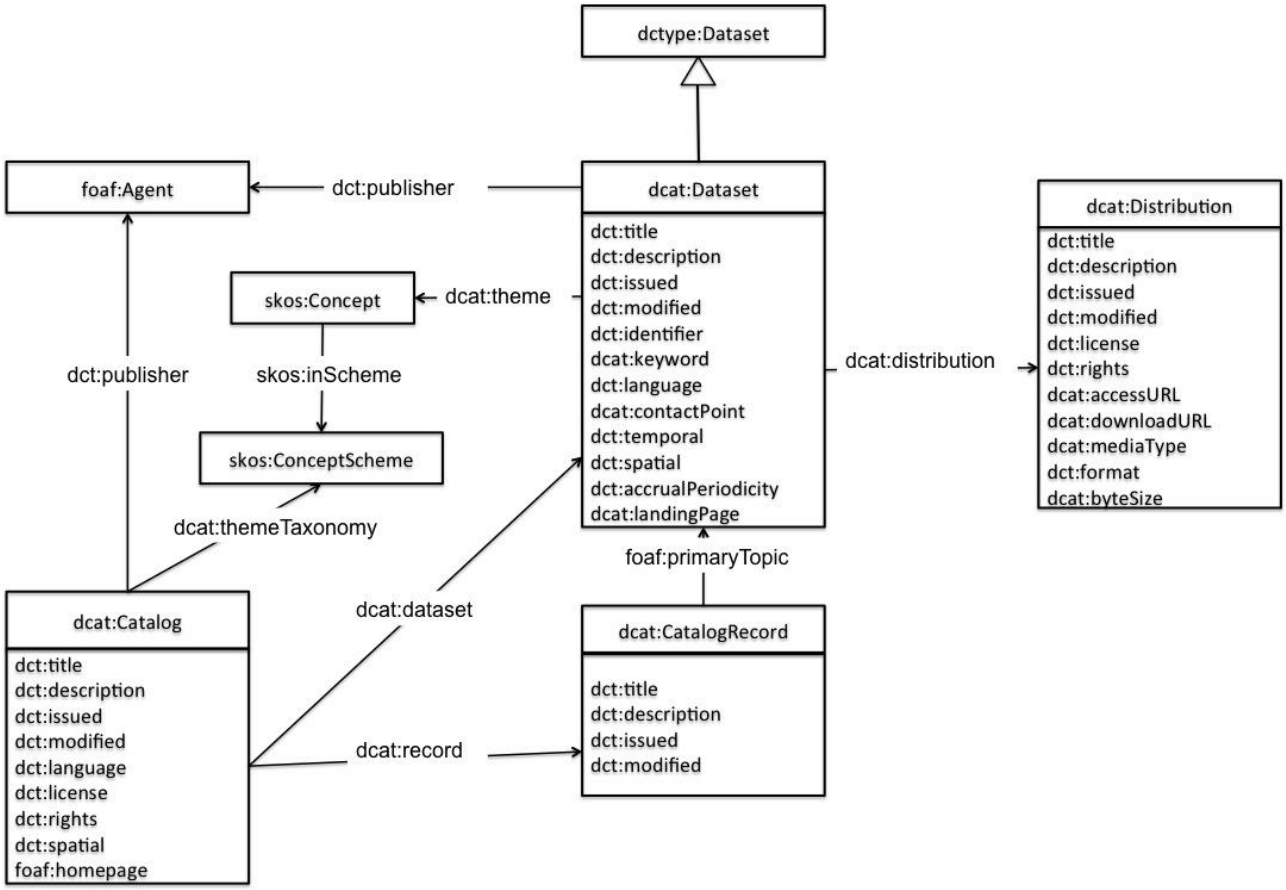

Figure 3. DCAT ontology overview (source: https://www.w3.org/TR/vocab-dcat/)

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.

2.5 Archiving and Preservation (Including Storage and Backup)

HOBBIT will also support the functionality of accessing and querying past versions of an evolving dataset, where all different benchmark versions will be publically available as dump file as well as from the project SPARQL endpoint. The data will be stored on the benchmarking platform 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.