HOBBIT

Collaborative Project

Holistic Benchmarking of Big Linked Data

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

Deliverable 1.1.1
Preliminary Community Member List, Use Cases, and Datasets

Dissemination Level	Public
Due Date of Deliverable	Month 12, 30/11/2016
Actual Submission Date	Month 14, 31/01/2017
Work Package	WP1 - Requirements Elicitation and Community Building
Task	T1.1
Type	Report
Approval Status	Final
Version	1.0
Number of Pages	14

Abstract: This deliverable serves as an overview of the work carried out so far towards gathering datasets and use cases from the community. Through this outreach, we aimed to make an increasing number of legal entities interested in the HOBBIT project. Here, we focus mainly on the assets (datasets, use cases) we were able to gather so far. A companion deliverable, D1.4, explains our outreach strategy plan and how we aim to build the HOBBIT community further.

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
0.0	14/10/2016	First draft created	Frank Salliau (iMEC)
0.1	28/10/2016	Draft revised	Axel-Cyrille Ngonga Ngomo (InfAI)
0.2	20/01/2017	Final version created	Axel-Cyrille Ngonga Ngomo (InfAI)
0.3	27/01/2017	Peer reviewed	Nadine Jochimsen (InfAI)
1.0	27/01/2017	Updates and corrections	Axel-Cyrille Ngonga Ngomo (InfAI)
1.0	31/01/2017	Final version submitted	Nadine Jochimsen (InfAI)

Author List

Organization	Name	Contact Information
iMec	Frank Salliau	frank.salliau@ugent.be
InfAI	Axel-Cyrille Ngonga Ngomo	ngonga@informatik.uni-leipzig.de

Executive Summary

This document details the preliminary state of the HOBBIT community. It includes an overview of the number of contacts within the different levels of engagement of the project. Currently, the project has been able to gather 228 relevant contacts and were already able to interact with 135 of these. This interaction has led to the gathering of 19 datasets that are now available to the consortium. A summary of these datasets (including a short description of their content and purpose, size and expected growth) is given herein. The interaction with experts and other research projects has also led to the preliminary definition of use cases within which benchmarking as offered by HOBBIT could be of central importance. These use cases and the relevant benchmarks and datasets are also detailed within this deliverable. Overall, the deliverable shows that the outreach and community building within HOBBIT has already led to fruitful interaction and to gathering relevant datasets and use cases. The usage of these interactions for the sake of community building is detailed in deliverable D1.4, the companion deliverable to D1.1.1.

Contents

List of Tables

List of Figures

1 Introduction

In its first year, HOBBIT has aimed to establish itself as the provider of a benchmarking platform for industry and academia with a focus of Big Linked Data technologies. One of the key steps towards achieving this goal was to build up a community of interested parties around the project. As shown in Figure 1, the idea behind this community is to

1. gather supplementary datasets relevant to the project,
2. gather KPIs for the evaluation of the frameworks,
3. gather solutions to benchmark and
4. collect potential members of the HOBBIT association.

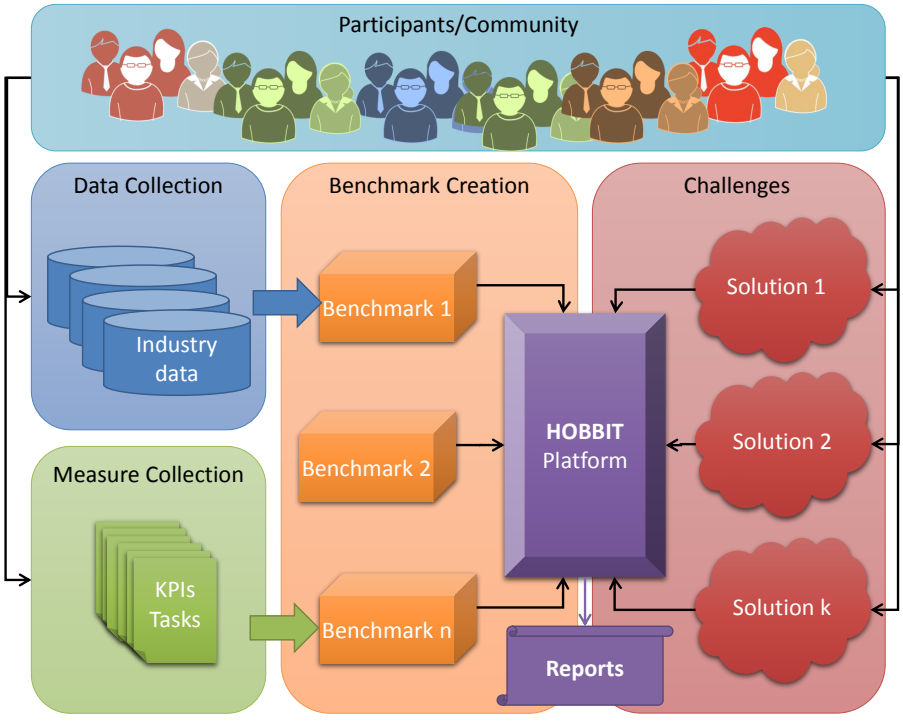

Figure 1: Overview of HOBBIT

This document details the preliminary state of the community, listing numbers of parties involved and entities that have expressed interest. It categorizes possible use cases, mentioning the interested parties. Use cases are linked to a set of datasets, of which the content, purpose, size, and expected growth are also described. Within the consortium, this document serves as foundation for an objective assessment of the efforts that were undertaken so far as well as for further discussion and growth of the community.

2 Preliminary State of the Community

2.1 Community Building Channels

To gather contacts for community building, we used a multi-channel strategy as described in Table 1. The results achieved by using this strategy are monitored continuously by the HOBBIT consortium (especially by the dissemination and outreach group). The outcomes of this monitoring are the subject of deliverable D1.4 of HOBBIT.[1]

Channel	Description
Mailing list	Subscriptions to the HOBBIT mailing list
Survey	Respondents to the survey sent out for requirements gathering
Flyers	Distribution of flyers at different events
Talks	Presentations of the HOBBIT project
Workshops	Organization of workshop at major conferences and events
Cooperations	Cooperation with relevant H2020 and national projects
Challenges	Organization of challenges at major conferences (ISWC, DEBS, ESWC)
Publications	Scientific publications about the core technologies of HOBBIT. Upcoming are publications which use the HOBBIT platform.

Table 1: Dissemination channels of HOBBIT

2.2 Current State of the Community

Over the last year, HOBBIT was disseminated in manifold ways with the aim of building up a community around the project. For example, the project was disseminated at more than 35 events (see Table 2 for an excerpt), within which we also aimed to get interested parties to join HOBBIT even at the lowest level of engagement possible. We also interacted through social media, for example by generating tweet content on a daily basis (see Figure 2). The parties we interacted with across our multi-channel outreach and dissemination strategy (see subsection 2.1) were asked to join the HOBBIT community or to provide us with contact data for further reference.

We gathered the following qualitative information on contacts:

- Email
- Full name, first name and last name
- Role and role type

[1] Available at https://project-hobbit.eu/about/deliverables/.

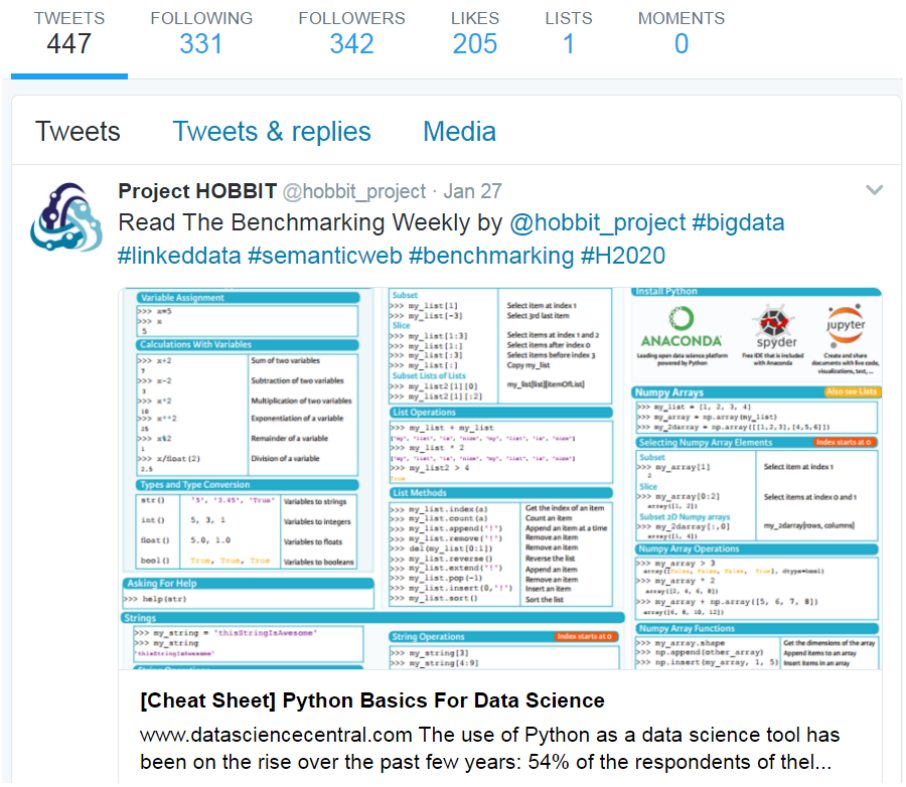

Figure 2: Snapshot of HOBBIT's Twitter account

- Company
- Country
- LinkedIn
- Comment
- Source
- Project Contact

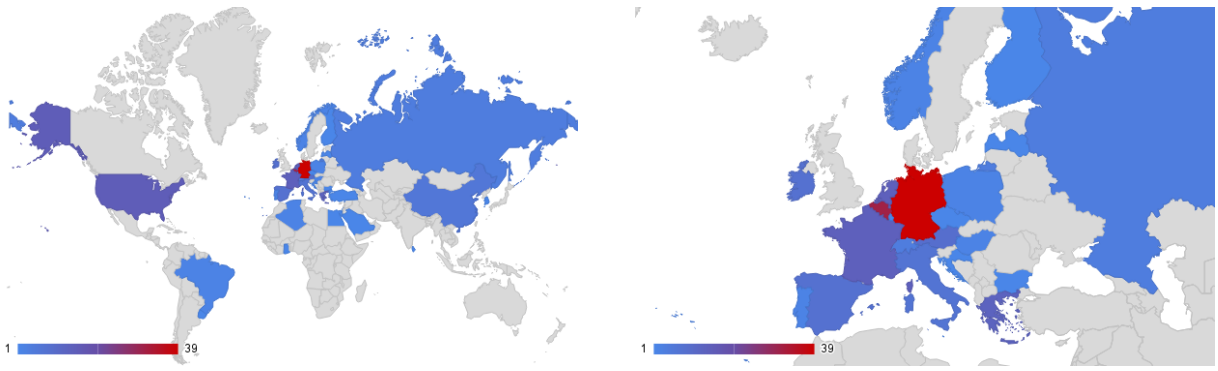

Figure 3: Distribution of HOBBIT contacts in the world (left) and in Europe (right)

So far, 228 contacts were established and registered in the project contact database. We mainly focused on attracting the attention of companies to the project. In particular, 29.2% of the members

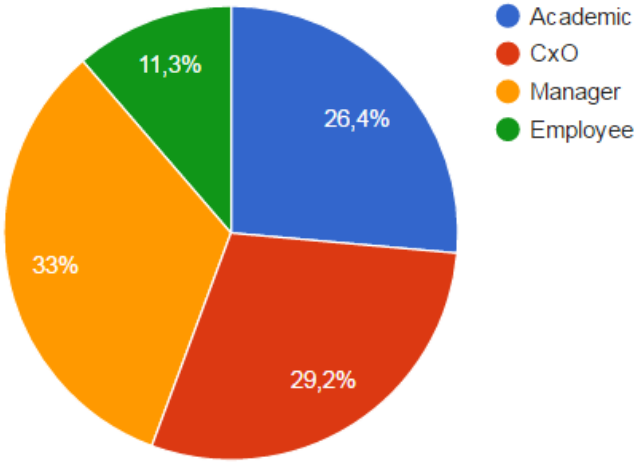

Figure 4: Distribution of roles of HOBBIT contacts

of the contact list are CxOs (e.g.,COO, CEO, CTO) while 33% are managers. The rest are academics (professors, researchers, etc.) and company employees (11.3%, see Figure 4). Of these 228 contacts, 135 contacts were subject to further interactions, while 93 are still to be contacted. We hence consider the 135 as being meaningful contacts in the sense of the description of work, meaning that we have already achieved 59% of the target of 250 meaningful contacts by the end of Year 1. While the contact data cannot be published in this deliverable for reasons of privacy, Figure 3 gives an overview of the geospatial distribution of the community so far. Most of our contacts are European, with 39 contacts from Germany and 30 from Belgium. We have however also aimed to reach out beyond Europe to get a glimpse of the current ideas, trends and use cases that could benefit the HOBBIT association later. For example, 15 of our contacts are from the US whereas 2 are from Brazil.

Event name	**Attendees/readers (estimates)**	**Dissemination and Outreach action**
European Data Forum	300+	Presentation in collaboration with BigDataEurope[2]
BITKOM Big Data Summit	600+	Pitch of HOBBIT
CEBIT	300+	Pitch of HOBBIT
ESWC 2016	300+	HOBBIT event
ISWC 2016	300+	BLINK workshop
ISWC 2016	300+	Link Discovery Tutorial
ECAI 2016	300+	Paper presentation
Diverse project meetings	100+	HOBBIT pitch, liaison
Interviews	10,000+	HOBBIT pitch

Table 2: Excerpt of dissemination and outreach events in which HOBBIT participated. ESWC stands for Extended Semantic Web Conference. ISWC is the International Semantic Web Conference. ECAI is the European Conference on Artificial Intelligence.

3 Datasets

During year 1, 19 datasets were gathered by the consortium and in a CKAN repository, accessible through the URL http://hobbit.iminds.be. These datasets are listed in the tables below.

Dataset	Description	Size (approximation)	Growth (expected)
Medical Subject Headings (MeSH)	Public RDF Datasets of Medical Subject Headings (MeSH) controlled vocabulary	27,883 descriptors in 2016 MeSH; 87,000 entry terms, 232,000 Supplementary Concept Records (SCRs)	Approximately 2% per year
LinkedSpending	LinkedSpending contains government spendings from all over the world as Linked Data. LinkedSpending uses the information collected by the OpenSpending project and makes it available as data cube	2 million financial transactions	7% per year
DBpedia	DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web. DBpedia allows answering complex questions using the W3C standard SPARQL.	3 billion facts, 125 languages, 38.3 entities	10-20% per year
CER Smart Metering Project	The Smart Metering Electricity Customer Behaviour Trials (CBTs) took place during 2009 and 2010 with over 5,000 Irish homes and businesses participating.	5,375 homes, 780 businesses	Static
Next Bike	Live information of GPS position of around 20,000 bicycles in about 70 cities (http://www.nextbike.net/)	Live stream of 3,000 bike positions, 70 cities	Unclear

Table 3: Excerpt of the datasets available to the HOBBIT project. A complete list can be found at http://hobbit.iminds.be. Generators can create datasets of any size. Hence size and expected growth cannot be stated.

Dataset	Description	Size (approximation)	Growth (expected)
BioASQ	Dataset underlying the question answering challenge of the same name. The challenges focuses on large-scale biomedical semantic indexing and question answering	800 questions	20,00%
Energy Map Germany	CSV data of development of solar energy within Germany with installation date, location, nominal capacity, GPS information	1.5 million entries	1-2%
LDBC	The LDBC-SNB Data Generator (DATAGEN) is the responsible of providing the data sets used by all the LDBC benchmarks.	Generator	Generator
LinkedGeoData	LinkedGeoData is an effort to add a spatial dimension to the Web of Data / Semantic Web.	30 billion facts	5-10% per year
TLC Trip Record Data	This dataset includes trip records from all trips completed in yellow and green taxis in NYC in 2014 and selected months of 2015.	1.1 billion taxi trips	10-20% per year
GitHub Data	GitHub is how people build software and is home to the largest community of open source developers in the world, with over 12 million people contributing to 31 million projects.	31 million projects,12 million users	5-10% per year
TWIG Ontology	The ontology for the synthetic version of Twitter based on the Twitter7 dataset.	Generator	Generator
QALD6	Question Answering on Linked Data version 6. The dataset contains approximately questions in natural language as well as the corresponding SPARQL queries and keyword queries to gather information from DBpedia, DBpedia abstracts and related datasets.	500 questions	10%

Table 4: Excerpt of the datasets available to the HOBBIT project. A complete list can be found at http://hobbit.iminds.be. Generators can create datasets of any size. Hence size and expected growth cannot be stated.

Dataset	Description	Size (approximation)	Growth (expected)
BENGAL	This family of datasets for named entity recognition, entity disambiguation and relation extraction are generated automatically out of RDF data using natural language generation.	Generator	Generator
LIVED	The "Long Device Level Energy Data" (LIVED) dataset and contains measurements collected from smart plugs multi-sensors as depicted.	2.5 billion measurements	Static
Linked Connections	Linked Connections is a method for generating publishing transit data using a low-cost API. It does this by exposing data in JSON(-LD).	Generator	Generator

Table 5: Excerpt of the datasets available to the HOBBIT project. A complete list can be found at http://hobbit.iminds.be. Generators can create datasets of any size. Hence size and expected growth cannot be stated.

4 Use Cases

The use cases of interest to the HOBBIT community and contacts vary significantly and are still being collected. So far, we were able to gather preliminary descriptions within

1. dissemination events,
2. interviews,
3. collaborations with other projects and
4. in deliverables of other projects.

This data collection process returns use cases hint at applications in the following domains (note that the names and contacts from which the information was gathered are partly omitted on purpose for the sake of privacy):

- **Industry 4.0**: The use of semantics in the industry 4.0 is of central importance for the creation of machines that can justify their behavior and interact with their users. Amongst other activities, we gathered information from the experts in the SAKE[3] and STEP[4] projects, who expressed interest in benchmarking link discovery, storage, machine learning and visualisation. Datasets such as the CER Smart Metering, LIVED and Weidmüller are of interest.
- **Geospatial data analysis**: Geospatial datasets belong to the largest and most used datasets on the planet. Contacts with experts from related projects (GeoKnow,[5] GEISER,[6] SmartRegio,[7] STEP, SLIPO, SAGE) revealed that these experts are interested in HOBBIT datasets related to geospatial entities and points of interest (LinkedGeoData, Energy Map Germany, LinkedConnections, TLC Record Trip). The benchmarks of interest here are related to knowledge extraction from structured and unstructured data, storage, versioning and machine learning and visualisation.
- **Smart Energy**: Devising a machinery that can use energy data to provide customers with intelligent energy services ranging from the automatic selection of energy providers to the detection of unwanted states (machinery on during the weekend, open fridge doors, etc.) is regarded as an innovative goal worthy of pursuit. Benchmarking how well such systems perform demands benchmarks in data acquisition, storage, versioning. Relevant datasets include the LIVED, Weidmüller and CER Smart Metering datasets.
- **Weather Data Analysis**: The increasing amount of streaming data from weather sensors demands novel techniques for the semantic analysis of streaming data. The area of continuous queries was regarded as one of the key areas for which benchmarking methodologies and unified semantics still need to be dealt with. Here, Smart metering data (LIVED, Weidmüller, CER) are regarded as being of significance, while storage and acquisition benchmarks are key.
- **Human Resource Management**: A rather surprising use case for the HOBBIT datasets, generators and benchmarks for the sake of finding good candidates for job offers. Novel applications

[3] http://sake-projekt.de
[4] https://www.projekt-step.de/
[5] http://geoknow.eu/
[6] http://www.projekt-geiser.de/
[7] http://www.smartregio.org/

for this purpose demand efficient entity recognition, entity linking and relation extraction, which are the area targeted by the knowledge extraction benchmark of HOBBIT. Relevant datasets here include the TWIG and the BENGAL datasets.

- **Enterprise Search**: Searching through streams of ever changing data is of central importance for data-driven companies. The use cases here include federated search across several datasets (see projects DIESEL[8] and WDAqua[9]) to search on mobile devices (e.g., project QAMEL[10]). The QALD 6, DBpedia, BioASQ, MESH and BENGAL datasets are here the most related while the knowledge acquisition benchmarks are the most important.
- **European societal challenges**: Through our collaboration with BigDataEurope, we were able to gather use cases for HOBBIT for seven of the societal challenges formulated by the European Union (i.e., health, food and agriculture, energy, transport, climate, social sciences and security). Given the diversity of the challenges, virtually all datasets and benchmarks provided by HOBBIT are relevant for at least one of the challenges or for the technical solutions underlying these challenges. For example, the CER Smart Metering data and the data storage and knowledge benchmarks are of central importance for the energy domain while LinkedConnections and all other transport datasets are relevant for the transport societal challenge.

Minor use cases include works on linguas francas for storage, morphology analysis as well as indexing for storage and question answering.

[8] https://diesel-project.eu/
[9] http://wdaqua.eu/
[10] https://qamel.eu/