HOBBIT

Collaborative Project

Holistic Benchmarking of Big Linked Data

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

Deliverable 1.3.1
Preliminary Association Mission Statement and Business Scenarios

Dissemination Level	Public
Due Date of Deliverable	Month 18, 31/05/2017
Actual Submission Date	Month 18, 31/05/2017
Work Package	WP1 - Requirements Elicitation and Community Building
Task	T1.3
Type	Report
Approval Status	Final
Version	1.0
Number of Pages	11

Abstract: This deliverable serves as a preliminary version of the mission statement for the HOBBIT association. In addition, this deliverable identifies a first set of business cases that have been pinpointed so far by the members of the consortium.

The information in this document reflects only the author's views and the European Commission 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
0.0	07/05/2017	First draft	Axel-Cyrille Ngonga Ngomo (InfAI)
0.2	12/05/2017	Revision of first draft	Giulio Napolitano (IAIS)
0.3	19/05/2017	Updates and corrections	Axel-Cyrille Ngonga Ngomo (InfAI)
1.0	26/05/2017	Quality checks	Eliza Shaw (InfAI)
1.0	31/05/2017	Final version submitted	Nadine Jochimsen (InfAI)

Author List

Organization	Name	Contact Information
InfAI	Axel-Cyrille Ngonga Ngomo	ngonga@infai.org

Executive Summary

This deliverable presents a preliminary version of the mission statement of the HOBBIT association. The association was designed to be as complementary as possible to existing organizations to ensure that synergies with other associations can be achieved. The four business cases presented were the result of reflections upon discussions with potential end users at events such as EDF, Apache Big Data and ESWC (see D8.2) as well as with technology vendors in and outside of the consortium.

Contents

List of Tables

1 Introduction

During the first half the HOBBIT project, the consortium has aimed to establish itself as the provider of a benchmarking platform for industry and academia with a focus of Big Linked Data technologies. By approaching both communities at more than 35 occasions (see D8.2), the consortium aimed to

- determine the need for a benchmarking platform for Big Linked Data in both academia and the industry. The idea here was to ensure that the generators of novel ideas (academia) and the customers in need (industry) both had a common point of reference pertaining to benchmarking Linked Data technologies;
- detect potential business use cases for this platform from the perspective of both potential categories of users and
- detect potential synergies with other benchmarking organizations.

The result of the analysis of

- the mission statements of other benchmarking organizations (TCP[1], the Benchmarking Network [2], LDBC[3], EUBA[4]) and
- the output of the interactions with industry and academia

led to the mission statement presented in Section 2.

The business use cases were mostly derived from discussions carried out with Big-Data-affine companies. Two events were of particular importance for this purpose:

- the EDF 2016 event organized with BigDataEurope and
- the discussion carried out at the Apache Big Data Europe event.

The results from these events were further refined through further interviews with companies in the consortium (TomTom, AGT, USU). The preliminary business use cases are elaborated upon in Section 3.

[1] http://tpc.org
[2] http://benchmarkingnetwork.com/
[3] http://ldbcouncil.org
[4] http://euba.com

2 Preliminary Mission Statement

In short, the aim of the HOBBIT association is to provide benchmarks, key performance indicators, benchmarking tools and benchmarking services for the *independent* and *repeatable* benchmarking of Big Data technologies which rely on Linked Data in any way. The HOBBIT association[5] was formed as a result of the HOBBIT project, which was led by the Institute for Applied Informatics e.V. with the following partners: FORTH, TomTom, USU, AGT, NSCR-D, Fraunhofer IAIS, iMEC and Open Link Sofware.

The HOBBIT association aims to facilitate the systematic evaluation, improvement and objective comparison of scalable Linked Data solutions. By these means, it aims to enable developers , vendors and users of such technologies to (1) characterize strengths and weaknesses of technologies of interest across several domains, (2) detect potential areas of improvements for existing solutions, (3) make informed decisions pertaining to the purchase and use (users) as well as the development and extension (developers and vendors) of existing technologies. By being a vendor-independent association based on open-source technologies, HOBBIT aims to ensure that these insights can be derived in an *unbiased manner*.

To achieve the goals aforementioned, HOBBIT develops and provides a series of software and data solutions.

1. *The universal benchmarking platform*: The HOBBIT platform is a scalable and distributed benchmarking platform for the *FAIR*[6] configuration, execution and analysis of benchmarks and their results along the whole of the Linked Data Lifecycle. This lifecycle includes data ingestion, storage, integration, analysis, versioning, curation, and usage in diverse domains. The association relies mostly on this platform to ensure that it generates comparable results across the domains and use cases it tackles.

2. *Benchmarks derived from real-world data*: The HOBBIT association currently provides 8 benchmarks based on a dozen of real-world industry-grade datasets including datasets from diverse domains. The benchmarks are designed to (1) reflect real loads from industry so as to generate realistic characteristic performance curves and (2) to push the limits of the systems across a series of choke points known to be current limitations of existing solutions. By maintaining and extending these benchmarks, the association pursues the goal of supporting the development of scalable Linked Data technologies by industry and academia.

3. *Real-world datasets*: The HOBBIT association has been entrusted with a dozen of real-world datasets upon which it based its benchmark development. The association sees its mission in the collection of further datasets for the creation of more realistic benchmarks that will cater for the needs of the increasing number of Linked Data technology providers and users.

4. *Reference implementations of benchmarks and key performance indicators*: The HOBBIT association aims to address the constant challenge of making results achieved by different platforms comparable. To address this goal, the association provides reference implementations for benchmarks and indicators and corresponding tests to ensure that the results reported by companies and academics can be compared and allow for an objective assessment of the performance of solutions across domains.

[5] The association will have its own dedicated subpage under http://project-hobbit.eu and the URI will point to the association after the end of the project.

[6] https://www.force11.org/group/fairgroup/fairprinciples

The association is a non-profit organization and welcomes financial and material support towards the achievement of its goal.

3 Preliminary Business Cases

Over the last 18 months, the HOBBIT consortium has gathered potential business use cases from events (see D8.2), its contact list (see D1.1.1, D1.1.2) and from within the consortium. Currently, four main business cases are foreseen, of which an overview of given in Table 1.

Use Case	Strengths	Weaknesses	Opportunities
Smoke Tests	Easy to implement; portable; cookie-cutter (can be employed on any application with standard I/O, e.g., triple stores); preliminary benchmark results for functionality tests	Need for known brand; preliminary investments	Currently not available (open market); Easy to deploy as service even for remote applications
On-premise benchmarking; Development of domain-specific benchmarks	Tailored tests for dedicated application; benchmarking on premise ensure ensures security of data and is thus commonly accepted by system developers; companies have full control over the setup; high remuneration potential; rare expertise available to the consortium; access to datasets ensures access to novel markets	Potentially small number of customers; need for HOBBIT association to be known	Few services of the sort available at the moment; Tedious
Challenge organization	Existing business model (Kaggle); cookie-cutter through platform, only need to gather relevant data and tests from companies; potentially high remuneration; access to experts in different domains and opening for new business models (e.g., expert search)	Potential competition if successful (Kaggle, etc.)	Access to novel markets; access to supplementary datasets
Benchmarking output	Already available through challenges; derivation of novel insights across domains; consultancy + access to insights derived from raw challenge data	Need for critical mass of users	Interest from other communities

Table 1: Summary of preliminary business cases

3.1 Smoke Tests

Smoke tests (also known as sanity tests) [1] are commonly defined as simple tests designed to reveal failures of such severity that they can impede a piece of software from running accurately. A common example for the need of such tests is result completeness. A large number of tools guarantee that they return complete results w.r.t. to the input of the user (e.g., a SPARQL query). However, the completeness claims of many tools has been shown to be wrong (see, e.g., [2]). The idea behind this use case is to define a series of smoke tests that can be used along the Linked Data lifecycle and lead to a software certification. With a "Checked by HOBBIT" logo, software applications could guarantee that they pass all the tests in a relevant suite. Establishing such certifications would allow promoting the HOBBIT brand at low costs and attract a large number of interested parties to see the other offers of HOBBIT. We are considering whether such a certification could be made available against a nominal fee used to maintain the tests and update them regularly. The regular updates are deemed important, as they would prevent engines from displaying a behavior tailored towards the smoke tests (see Volkswagen incident). HOBBIT's large number of datasets will be critical in this sense to ensure that the smoke tests present the tools to be evaluated with a variety of tasks.

The **strengths** of this business case lie at hand. While the W3C and a few code bases provide simple tests for some of the Semantic Web functionality (e.g., checking the output of RDF libraries for whether it conforms to specifications), an exhaustive library of smoke tests for the Linked Data lifecycle does not exist. However, such a library would clearly be of benefit for application developers as it would allow certifying the tools they make available on the market. The library could be easily built to be one-size-fits-all and to rely exclusively on standards (SPARQL, HTTP, XML, etc.). Therewith, it would be a "cookie-cutter"-type product. Finally, the smoke tests could be derived directly from the HOBBIT benchmarks, making the product easy to build.

The **weaknesses** of this business case lie in its commercial viability. Given that one would basically sell a brand, it would be important for the HOBBIT branding to be known and accepted. Hence, the acceptance of the branding would depend on the efforts made during and after the project to popularize the HOBBIT brand. This will most probably be combined with financial investments that will have to be carried before revenue can be made. Still, the **opportunites** linked to this case are clear: Implementing this business case would establish HOBBIT across the whole Linked Data community as the reference organization for checking the viability of implementations. Hence, it would allow securing a large number of members, who would want to influence the smoke tests through their membership. This large number of members would ensure that even more relevant smoke tests are carried out and that corner cases are not forgotten, hence leading to a self-funded ecosystem of use cases and HOBBIT members.

3.2 On-Premise Benchmarking

Most companies have no interest in making their data and software tools available for benchmarking. Still, the interest in objective performance measurements grows constantly. This business use case targets this constellation by offering "on-premise" benchmarking services. The potential customers are two-fold:

1. data owners interested in the right tool for their particular data processing task and
2. solution providers interested in knowing how well they perform when compared with the competition.

For the first customer segment, the service offers would include

1. going to the premises of the customers,
2. installing the benchmarking platform and the corresponding benchmarks and available reference software,
3. implementing the necessary system adapters for the platforms of their choice to run with HOBBIT,
4. running experiments and
5. generating reports (possible with suggestions for improvement).

The procedure for the second customer segment would be similar, with the slight difference that they would most probably be interested in the baselines (as reference data) and their own piece of software.

In all cases, the **strenghts** of this business case is that it requires a significant amount of rare expertise that would only be available to the consortium. This makes the use case at hand difficult to implement for potential competitors at the same costs as for the HOBBIT association. The need for specific functionality (the system adapters) makes the use case more attractive from a financial standpoint. However, it also means that a significant amount of work would have to be performed for each business case of the type. This can still be regarded as an advantage, as a series of such cases would significantly improve the ecosystem around the HOBBIT platform and many adapters could be reused.

The **weaknesses** of the use case at hand are linked to the need for tailored solutions for each of the instantiations of the use case. Still, the **opportunities** clearly make this business case viable, as it would support the growth of the HOBBIT ecosystem across different systems. Moreover, attractive deals linked to HOBBIT memberships would support the extension of the member base for the association.

3.3 Challenge Organization

Organizations such as Kaggle have shown that the organization of challenges can be a viable and attractive business model. In this business use case, we build upon the idea behind Kaggle by supporting customers of the association in the creation of business-relevant challenges: Data-driven companies are constantly faced with challenges that are difficult to solve in-house or where even slight performance improvements could be turned into significant financial gains (e.g., a reduction of the query runtimes of distributed query engine can save space, time and energy). Such organizations often have to buy experts knowledge without any guarantees of results. HOBBIT could support companies facing this need by offering to support the organization, development, deployment and management of challenges. For the organizations, it would mean the provision of datasets and of benchmark specifications to be used. Baselines could also be provided. The HOBBIT consortium would manage the process from the formulation of the tasks to the selection of winners, including all intermediary steps.

The **strengths** of this business case lie in the business model behind it being known to be viable. Organizations such as Kaggle have shown that supporting the benchmark process can be turned into a profitable endeavor. The platform and the benchmarks build a solid foundation for showcasing the abilities and strengths of the association to potential customers. Like in the first use case, this however demands some investments into the HOBBIT brand, which is one of its **weaknesses**. The **opportunities** connected with this business case are however immense as establishing HOBBIT as the reference organization for the design and deployment of challenges should be able to sustain the

organization financially. Moreover, the number of members could be increased by offering reduced fees to association members, a feature that has been foreseen since the beginning of the project.

3.4 Benchmarking Output

Consultants such as Gartner[7] have built an important part of their business on the distribution of documents pertaining to the performance of frameworks and corresponding prognoses. While the HOBBIT association does not plan to reach the business volume of such organizations, the uptake of Linked Data in industry makes the provision of reports on system performances, current bottlenecks and recent advances a potentially viable pillar for the association. This is especially of interest because a significant portion of these reports can be generated automatically but still demands the special expertise of the consortium to be assigned the right interpretation.

The **strenghts** of this business case lie in its being one-size-fits-all for each step of the Linked Data lifecycle. Moreover, a large proportion of the generation process can be automated, making it a low-cost product for the association. The necessary expertise to assign the results generated a correct meaning also makes this particular case difficult to implement for other organizations, thus making the potential competition limited. The **weaknesses** of the business case are again related to branding. The association must be well known amongst practitioners for the reports to attain a high monetary value and be bought regularly. The **opportunity** behind this business case is still clear as there are no associations that (1) focus on the whole of the Linked Data lifecycle while (2) aiming to provide insights on the performance of a large number of tools and solutions.

4 Conclusion

In this document, we presented a preliminary version of the mission statement of the HOBBIT organisation. In addition, we presented a brief and preliminary compilation of potential business cases for the HOBBIT association. Throughout the business use cases, the need for HOBBIT to be established as a reference platform became clear. Hence, over the next 18 months, we will intensify our efforts to make the value of the HOBBIT platform and of the HOBBIT results clear to industry and academia. The association will be built up with the aim of gathering members that are not in the consortium but are willing to push the development of high-performance Linked Data driven solutions.

[7] http://gartner.com

References

[1] Steve McConnell. Daily build and smoke test. *IEEE software*, 13(4):144, 1996.

[2] Muhammad Saleem, Yasar Khan, Ali Hasnain, Ivan Ermilov, and Axel-Cyrille Ngonga Ngomo. A fine-grained evaluation of SPARQL endpoint federation systems. *Semantic Web*, 7(5):493–518, 2016.