WP1-D1.3
Case Study Evaluation Benchmark

Project full title:	Knowledge Driven Data Exploitation
Project acronym:	K-Drive
Grant agreement no.:	286348
Project instrument:	EU FP7/Maria-Curie IAPP/PEOPLE WP 2011
Document type:	D (deliverable)
Nature of document:	R (report)
Dissemination level:	PU (public)
Document number:	ISOCO/WP1-D1.3/D/PU/a1
Responsible editors:	Panos Alexopoulos
Reviewers:	Yuan Ren
Contributing participants:	ISOCO, UNIABDN
Contributing workpackages:	WP1
Contractual date of deliverable:	31 December 2012
Actual submission date:	31 December 2012

Abstract
In this deliverable, we specify an evaluation framework for the assessment of the effectiveness and efficiency of the K-Drive system, in the context of the use case scenario described and specified in deliverables D1.1 and D1.2. The framework focuses on the scenario's two main tasks, namely semantic question answering and semantic data summarisation, and defines appropriate procedures and metrics for measuring the final system's effectiveness in dealing with them.

Keyword List
Semantic Question Answering, Semantic Data Summarisation, Evaluation Framework

Project funded by the European Commission within the 7th Framework Programme/Maria Curie Industry-Academia Partnerships and Pathways schema/PEOPLE Work Programme 2011.

Case Study Evaluation Benchmark

Panos Alexopoulos[1], Andrew Walker[2] and Jose-Manuel Gomez-Perez[1]

[1] iSOCO, Spain
Email: {palexopoulos, jmgomez}@isoco.com

[2] Department of Computing Science, Aberdeen University, UK
Email: andrew.walker.05@abdn.ac.uk

31 December 2012

Abstract
In this deliverable, we specify an evaluation framework for the assessment of the effectiveness and efficiency of the K-Drive system, in the context of the use case scenario described and specified in deliverables D1.1 and D1.2. The framework focuses on the scenario's two main tasks, namely semantic question answering and semantic data summarisation, and defines appropriate procedures and metrics for measuring the final system's effectiveness in dealing with them.

Keyword List
Semantic Question Answering, Semantic Data Summarisation, Evaluation Framework

Contents

1 Introduction

In deliverables D1.1 and D1.2 we specified a comprehensive use case scenario for an intended K-Drive system involving the provision of advanced methods, techniques and tools for reuse and access of semantic data that are publicly available. In particular, we specified a set of functional requirements covering two main tasks:

1. **The user-driven and scenario-based semantic question answering**, namely the transformation of natural language questions to semantic queries and their execution against one or more datasets in order to derive answers.

2. **The user-driven and scenario-based semantic data summarisation**, namely the facilitation of better understanding and evaluation of a given dataset, through the provision of a customized view of it in an intuitive and non-technical way.

In this deliverable, in turn, we focus on specifying an appropriate framework for evaluating the K-Drive system against the requirements of the above two tasks and measuring its effectiveness and efficiency. This framework, which is to be applied in the later stages of the projects after the development of the system, consists of a set of metrics and processes for assessing specific aspects and dimensions of the system's functionality, including performance, effectiveness, usability and utility.

Give that, the structure of the rest of this document is as follows. In the next section we consider the task of semantic question answering, we provide a brief overview of the typical ways employed for its evaluation in the literature and we define a specific evaluation strategy for its evaluation within K-Drive. The same is done in section **??** for the semantic data summarisation task.

2 Evaluation of Semantic Question Answering

The challenge of question answering over linked data is an intersection of various previously-established topics of research such as Information Retrieval and Natural Language Processing. Linked data on the web has seen exponential growth since its inception to an estimated 31 billion RDF triples in September 2011. Significant progress has been made on question answering from textual data [Strzalkowski and Harabagiu, 2007] and in Natural Language Interfaces to relational databases [Minock, 2010] while, in the intersection, we have seen [Cimiano et al., 2008]. [Bernstein et al., 2005], [Lopez et al., 2007], [Lopez et al., 2009] and [Tran et al., 2009].

Nevertheless, in contrast to the Information Retrieval community, where evaluation using standardized techniques, such as those used for the annual TREC[1] conferences, has been common for decades, in the Semantic Web community there are not yet adopted standard evaluation benchmarks for semantic question answering systems over real-world datasets.

A recent attempt to create such standards by developing datasets needed to formally judge the quality of ontology-based question answering approaches, has been the QALD challenge, realized by two relevant workshops so far[2] [3]. This challenge practically provides tests consisting of a variety of questions of different complexity, designed to represent questions that real end

[1] http://trec.nist.gov/

[2] http://www.sc.cit-ec.uni-bielefeld.de/qald-1

[3] http://greententacle.techfak.uni-bielefeld.de/ cunger/qald/index.php?x=challenge&q=2

users would ask. For example a particular challenge involved answering 100 questions based on the DBPedia and MusicBrainz (populated) ontologies (50 each).

Results were measured by precision, recall and f-measure, both for each individual question and each set of 50 questions together. More specifically, for individual questions the effectiveness measures were the following:

$$\text{Recall} = \frac{\text{number of correct system answers}}{\text{number of gold standard answers}}$$

$$\text{Precision} = \frac{\text{number of correct system answers}}{\text{number of system answers}}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Similarly, for each set of questions the respective measures were:

$$\text{Recall} = \frac{\text{number of correctly answered questions}}{\text{number of questions}}$$

$$\text{Precision} = \frac{\text{number of correct answered questions}}{\text{number of answered questions}}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

For the evaluation of the K-Drive semantic question answering system we will utilize the same measures as above. Moreover, the datasets to be considered will be those of QALD (for comparison purposes with other systems) as well as custom ones to be derived from the ones specified in deliverable D1.1.

In the latter case particular focus will be given to queries that highlight the need for custom interpretation based on the characteristics of the application scenario. More specifically, as already explained in D1.2, an important requirement for the K-Drive system is the ability to select and use in a custom way those semantic data that are most appropriate for the given application scenario. That is because there are often situations where users ask questions with some goal in mind (e.g. to educate themselves on an issue, to verify a fact, to take a decision etc.) and perhaps with some special focus (e.g. a film critic might be asking questions about films and genres). In such cases, some datasets (or some parts of them) may be more effective in answering these questions than others (e.g. greater subject coverage, availability of particular relations or concepts etc.) and therefore it is important that the former are identified and used instead of the latter.

In this context, it is important to evaluate the K-Drive system against queries that are better answered when only a particular dataset (or subset of it) is actually considered, and to verify the followings:

1. If the system manages to effectively detect and consider the correct data subset which is able to provide the best answers to the user queries.

2. If by following this approach the overall query answering effectiveness is increased.

3 Evaluation of Semantic Data Summarisation

Evaluation approaches of data summarisation techniques and frameworks fall generally into two categories, namely intrinsic and extrinsic ones. An intrinsic evaluation tests the summarisation system in itself by assessing mainly the coherence and informativeness of summaries. A drawback of this is that human judgement often has wide variance on what is considered a "good" summary, which means that automatic evaluation process is particularly difficult.

An extrinsic evaluation, on the other hand, tests the summarisation based on how it affects the completion of some other task. That means that summaries are evaluated in a concrete task seeking to verify if the summaries are instruments which could be used instead of the full data in specific situations. Variables measured can be accuracy in performing a task or time to complete the task. Nevertheless, while extrinsic evaluation is very attractive from the point of view of information access, it is also very time-consuming and costly, so its application has been limited.

In the area of semantic data and ontologies, evaluation of summarisation approaches has been primarily done in an intrinsic way. For example, in the system of [d'Aquin and Motta, 2011], which summarizes semantic datasets by generating key queries they may answer, the authors asked 12 users with various degrees of familiarity with semantic technologies to inspect a reasonably large dataset and express up to 5 questions they believed to be interesting on this. These questions were then compared against those generated by the system.

Similarly, in [Li and Motta, 2010] the setting of the evaluation involved eight people, each with good experience on ontology engineering, who were asked, for a number of ontologies, to manually extract up to 20 key concepts they considered the most representative for summarizing the contents of each ontology. The concepts that were chosen by at least 50% of the experts form a reference summary, referred to as "ground truth" summary.

In K-Drive the main goal of the summarisation module, as specified in D1.2, is to allow its users to define and generate custom data summaries, by selecting, parameterizing and executing particular summarisation subtasks that are predefined and designed to produce particular types of summaries. These custom summaries are to satisfy particular data description needs, typically associated to a task that the user wants to perform or an ultimate goal he/she wants to achieve. In this context, the key dimensions of the summarisation system that need to be evaluated include:

1. **Accuracy:** This is the dimension mentioned above and involves users judging whether a particular summary is accurate or not. This dimension is not applicable to all types of summaries but only to those who are somehow vague (e.g. "key" concepts, "representative" queries etc).

2. **Usefulness:** The practical usefulness of the overall system, as perceived and rated by the its users. Some key metrics for measuring this are the following:

   (a) **Number of summary types:** A too large number of summary types may cause information overload for the users while a too small one may compromise the system's usefulness.

(b) **Comprehensibility of summary types:** The intended meaning of the information a summary type is supposed to convey needs to be clear for the user. For example, a summary type named *"Dataset vertical completeness" i*s less comprehensible than one named *"Range of entities the dataset covers"*.

(c) **Usefulness of summary types:** Typically, summary types will be defined with a set of potential usages in mind. Yet, it might be that these usages are not really of interest to the users of the system, thus making the particular types redundant.

(d) **Completeness of summary types:** Important summary types that are useful for particular tasks may not be supported by the system

3. **Recommendation Effectiveness:** As in all kinds of recommender systems, the effectiveness of this system in recommending useful complementary summary types to users is important.

4. **System Usability:** This refers to the easiness and intuitiveness with which the users are able to use the system for defining, managing, executing and consuming custom data summaries.

5. **System Performance:** Some summary types (e.g. important queries) may require substantial time to be generated and that might not be tolerable by users. For such cases, performance needs to evaluated.

To evaluate the above dimensions, we shall conduct an empirical evaluation of the system through action research. The latter is a social science research approach for qualitative evaluation of theory application to solve practical problems [Baskerville, 1999]. It has been originally used disciplines such as education, psychology and health care and it can generally be applied in field settings where more traditional experimental or quasi-experimental methods cannot easily be applied. One of its major advantages, which is important in our case, is that it can help to overcome the problem of persuading practitioners to adopt new techniques.

In practice, action research is usually carried out in a number of discrete cycles each of which consists of the following steps:

1. **Planning:** This step involves developing a plan to improve current practice. In our case, this practice is the way users typically evaluate and select semantic data and the plan is to change this through the use of K-Drive's summarisation system.

2. **Acting:** This step involves having a set of participants acting to implement the plan which, in our case, is translated into having users actually use the system and produce summaries for specific tasks.

3. **Observing:** In this step the user's actions are observed to collect evidence for the thorough evaluation of the outcomes. For the evaluation of the summarisation system, the users are expected to participate directly in the measurement of the above metrics through structured questionnaires and interviews.

4. **Reflecting:** In this last step participants reflect on their overall experience, discuss pros and cons of the system and suggest points of improvement.

The ultimate goal of the above process is to prove, in an empirical way, the practical utility of the system as a tool for better understanding and evaluation of semantic data as well as its superiority over existing similar systems and frameworks. For the latter, we will conduct, through the above process and metrics, comparative evaluations of existing systems like, for example, semantic web browsers[4] [Dzbor et al., 2007] [Berners-Lee et al., 2008] or other semantic data visualization environments [Chen et al., 2006] [Hervás and Bravo, 2011].

4 Conclusion

This deliverable describes the process we are going to follow in order to evaluate the effectiveness and efficiency of the K-Drive system, in the context of two tasks described and specified in deliverables D1.1 and D1.2, namely semantic question answering and semantic data summarisation.

Acknowledgement

This research has been funded by the European Commission within the 7th Framework Programme/Maria Curie Industry-Academia Partnerships and Pathways schema/PEOPLE Work Programme 2011 project K-Drive number 286348 (cf. `http://www.kdrive-project.eu`).

References

[Baskerville, 1999] Baskerville, R. L. (1999). Investigating information systems with action research. *Commun. AIS*, 2(3es).

[Berners-Lee et al., 2008] Berners-Lee, T., Hollenbach, J., Lu, K., Presbrey, J., Prud'hommeaux, E., and Schraefel, M. M. C. (2008). Tabulator redux: Browsing and writing linked data. In *LDOW*.

[Bernstein et al., 2005] Bernstein, A., Kaufmann, E., Göhring, A., and Kiefer, C. (2005). Querying ontologies: A controlled english interface for end-users. In Gil, Y., Motta, E., Benjamins, V. R., and Musen, M. A., editors, *International Semantic Web Conference*, volume 3729 of *Lecture Notes in Computer Science*, pages 112–126. Springer.

[Chen et al., 2006] Chen, I.-X., Yang, C.-Z., and Hsu, T.-L. (2006). Design and evaluation of a panoramic visualization environment on semantic web. *Inf. Softw. Technol.*, 48(6):402–409.

[Cimiano et al., 2008] Cimiano, P., Haase, P., Heizmann, J., Mantel, M., and Studer, R. (2008). Towards portable natural language interfaces to knowledge bases - the case of the orakel system. *Data Knowl. Eng.*, 65(2):325–354.

[d'Aquin and Motta, 2011] d'Aquin, M. and Motta, E. (2011). Extracting relevant questions to an rdf dataset using formal concept analysis. In Musen, M. A. and Corcho, s., editors, *K-CAP*, pages 121–128. ACM.

[4] http://browse.semanticweb.org/

[Dzbor et al., 2007] Dzbor, M., Motta, E., and Domingue, J. (2007). Magpie: Experiences in supporting semantic web browsing. *Web Semant.*, 5(3):204–222.

[Hervás and Bravo, 2011] Hervás, R. and Bravo, J. (2011). Towards the ubiquitous visualization: Adaptive user-interfaces based on the semantic web. *Interact. Comput.*, 23(1):40–56.

[Li and Motta, 2010] Li, N. and Motta, E. (2010). Evaluations of user-driven ontology summarization. In *Proceedings of the 17th international conference on Knowledge engineering and management by the masses*, EKAW'10, pages 544–553, Berlin, Heidelberg. Springer-Verlag.

[Lopez et al., 2007] Lopez, V., Uren, V., Motta, E., and Pasin, M. (2007). Aqualog: An ontology-driven question answering system for organizational semantic intranets. *Web Semant.*, 5(2):72–105.

[Lopez et al., 2009] Lopez, V., Uren, V., Sabou, M. R., and Motta, E. (2009). Cross ontology query answering on the semantic web: an initial evaluation. In *Proceedings of the fifth international conference on Knowledge capture*, K-CAP '09, pages 17–24, New York, NY, USA. ACM.

[Minock, 2010] Minock, M. (2010). C-phrase: A system for building robust natural language interfaces to databases. *Data Knowl. Eng.*, 69(3):290–302.

[Strzalkowski and Harabagiu, 2007] Strzalkowski, T. and Harabagiu, S., editors (2007). *Advances in Open Domain Question Answering (SUNY Albany and University of Texas at Dallas) Springer (Text, speech and language technology series, edited by Nancy Ide and Jean Véronis, volume 32), 2006, xxvi+566 pp; hardbound, ISBN 978-1-4020-4744-2.*, volume 33.

[Tran et al., 2009] Tran, T., Wang, H., Rudolph, S., and Cimiano, P. (2009). Top-k exploration of query candidates for efficient keyword search on graph-shaped (rdf) data. In *Proceedings of the 2009 IEEE International Conference on Data Engineering*, ICDE '09, pages 405–416, Washington, DC, USA. IEEE Computer Society.