LarKC

The Large Knowledge Collider:
a platform for large scale integrated reasoning and Web-search

FP7 – 215535

D6.2 – Templates of periodic report on data and performances

Coordinator: Kono Kim (Saltlux)
With contributions from: Irene Celino, Emanuele Della Valle, Daniele Dell'Aglio, CEFRIEL; Yi Huang, Siemens; Werner Hauptmann, Siemens

Document Identifier:	LarKC/2008/D6.2
Class Deliverable:	LarKC EU-IST-2008-215535
Version:	Version 1.0
Date:	December 22, 2008
State:	Final
Distribution:	Public

EXECUTIVE SUMMARY

This document is concerned with the template of the periodic report on data and performances. It will be used in order to provide regular report on the volume of data acquired and the performance of the LarKC platform as plotted against the target performance characteristics established by the requirements analysis.[1] This template is one of self assessment plan and major monitorizable indicators for urban computing use case. This report is supposed to provide measuring and quantifying progress and impact of LarKC project. By this template, LarKC consortium and other interest groups can recognize easily the status and achievement of this project.

For reporting on the data the template includes the data source metadata, the "semantics" of the data source, and the data source format. These template elements were carefully selected to describe data characteristics, usages, size and so on. For another issue, reporting on performances we selected the criteria which are used to indicate what the LarKC platform is able to achieve at the reporting time in terms of scalibility, heterogeneity support, resuablity, and statistical measurements.

DOCUMENT INFORMATION

IST Project Number	FP7 - 215535	**Acronym**	LarKC
Full Title	The Large Knowledge Collider: a platform for large scale integrated reasoning and Web-search		
Project URL	http://www.larkc.eu/		
Document URL			
EU Project Officer	Stefano Bertolo		

Deliverable	**Number**	6.2	**Title**	Templates of periodic report on data and performances
Work Package	**Number**	6	**Title**	Urban Computing: Real Time City

Date of Delivery	**Contractual**	M9	**Actual**	December 2008
Status	version 0.11		final ■	
Nature	prototype □ report ■ dissemination □			
Dissemination level	public ■ consortium □			

Authors (Partner)	Kono Kim (Saltlux), Irene Celino, Emanuele Della Valle and Daniele Dell'Aglio (CEFRIEL), Yi Huang and Werner Hauptmann (Siemens)			
Responsible Author	**Name**	Kono Kim	**E-mail**	kono@saltlux.com
	Partner	Saltlux	**Phone**	+82-10-8145-5386

Abstract (for dissemination)	The template of the periodic report on data and performances provide regular report on the volume of data acquired and the performance of the LarKC platform as plotted against the target performance characteristics established by the requirements analysis. This template is one of self accessment plan and major monitable indicators for urban computing use case. This report is supposed to provide measuring and quantifying progress and impact of LarKC project.
Keywords	performance, data, use case, template, urban computing, measure, periodic report

Version Log			
Issue Date	**Rev. No.**	**Author**	**Change**
November 19, 2008	0.1	Irene	First draft of the document
November 24, 2008	0.2	Kono	Further additions
November 25, 2008	0.3	Irene	Feedbacks and contributions
December 2, 2008	0.4	Kono	Major contribution
December 4, 2008	0.5	Yi	Comments and contributions
December 5, 2008	0.6	Werner	More comments
December 6, 2008	0.7	Irene	Revision and feedbacks
December 8, 2008	0.8	Kono	Revision for last partners check
December 9, 2008	0.9	Yi	Feedbacks and contributions
December 15, 2008	0.10	Kono	Submitted for internal QA
December 22, 2008	1.0	Kono	Revision with remarks from QA

PROJECT CONSORTIUM INFORMATION

Participant's name	Partner	Contact
Semantic Technology Institute Innsbruck, Universitaet Innsbruck		Prof. Dr. Dieter Fensel, Semantic Technology Institute (STI), universitaet Innsbruck, Innsbruck, Austria, E-mail: dieter.fensel@sti-innsbruck.at
AstraZeneca AB		Bosse Andersson AstraZeneca Lund, Sweden Email: bo.h.andersson@astrazeneca.com
CEFRIEL - SOCIETA CONSORTILE A RESPONSABILITA LIMITATA		Emanuele Della Valle, CEFRIEL - SOCIETA CONSORTILE A RESPONSABILITA LIMITATA, Milano, Italy, Email: emanuele.dellavalle@cefriel.it
CYCORP, RAZISKOVANJE IN EKSPERIMENTALNI RAZVOJ D.O.O.		Michael Witbrock, CYCORP, RAZISKOVANJE IN EKSPERIMENTALNI RAZVOJ D.O.O., Ljubljana, Slovenia, Email: witbrock@cyc.com
Höchstleistungsrechenzentrum, Universitaet Stuttgart		Georgina Gallizo, Höchstleistungsrechenzentrum, Universitaet Stuttgart, Stuttgart, Germany, Email: gallizo@hlrs.de
MAX-PLANCK GESELLSCHAFT ZUR FOERDERUNG DER WISSENSCHAFTEN E.V.		Dr. Lael Schooler Max-Planck-Institut für Bildungsforschung Berlin, Germany Email: schooler@mpib-berlin.mpg.de
Ontotext Lab, Sirma Group Corp		Atanas Kiryakov, Ontotext Lab, Sofia, Bulgaria Email: atanas.kiryakov@sirma.bg
SALTLUX INC.		Tony Lee, SALTLUX INC, Seoul, Korea, Email: tony@saltlux.com
SIEMENS AKTIENGESELLSCHAFT		Dr. Volker Tresp, SIEMENS AKTIENGESELLSCHAFT, Muenchen, Germany, E-mail: volker.tresp@siemens.com
THE UNIVERSITY OF SHEFFIELD		Prof. Dr. Hamish Cunningham, THE UNIVERSITY OF SHEFFIELD Sheffield, UK, Email: h.cunningham@dcs.shef.ac.uk

VRIJE UNIVERSITEIT AMSTERDAM		Prof. Dr. Frank van Harmelen, VRIJE UNIVERSITEIT AMSTERDAM, Amsterdam, Netherlands, Email: Frank.van.Harmelen@cs.vu.nl
THE INTERNATIONAL WIC INSTITUTE, BEIJING UNIVERSITY OF TECHNOLOGY		Prof. Dr. Ning Zhong, THE INTERNATIONAL WIC INSTITUTE, Mabeshi, Japan, Email: zhong@maebashi-it.ac.jp
INTERNATIONAL AGENCY FOR RESEARCH ON CANCER		Dr. Paul Brennan, INTERNATIONAL AGENCY FOR RESEARCH ON CANCER, Lyon, France, Email: brennan@iarc.fr

TABLE OF CONTENTS

1. Introduction

This deliverable aims to provide statistical and quantitative and qualitative indicators which can be used for measuring and quantifying progress and impact of LarKC project[2]. By using this template, LarKC consortium and other interest groups can recognize easily the status and achievement of this project. This template of the periodic report on data and performances plays a role not just measuring and evaluating this project but also providing and communicating the whole LarKC deliverables to our final users.

The goal of LarKC is to provide an infrastructure that scales up to realistic semantic computing applications. Its success much depends on data scale and reasoning performance. That is the reason why we have to define a template for the periodic report on data and performances. In this scope, we considered several aspects of measurement criteria for data and performance. On one side, a small amount of criteria could be misread by themself. And on the other side, too many criteria are also too hard to capture and unnecessary while indicating unclear conclusions from the progress point of view.

During three months from M6 we have been carefully selecting criteria as template elements that indicate the status of LarKC project effectively. In results, the selected template elements consist of measurable values and clear descriptions, so that the final user can easily get access to the achievement of the Urban Computing[3] use case. For reporting on the data we considered the template elements like the data source metadata, the "semantics" of the data source, and the data source format. These template elements answer questions such as "How can LarKC exploit those data?", "What kind of data can be processed with LarKC?", "What is the size of data used in the urban computing use case?", and so on.
For reporting on performances, we selected the template elements which are used to indicate what the LarKC platform is able to achieve at the reporting time in terms of scalibility, heterogeneity support, resuable functionality and statistiscal measurements. These template elements can answer questions like "How much time we need to get the results?", "What kinds of functionalities LarKC have?", "How many users can work with?" and so on.

Using the template defined in this deliverable we will periodically report on data and performances and show the progress of the LarKC project from the point of view of the Urban Computing use case at M12, M18, M26, M33, and M42. (The corresponding deliverables are D6.4, D6.6, D6.7, D6.8, and D6.11 respectively.) Both data and performances play an important role for the goals mentioned above in the first paragraph, since our use case requires an integration of massive heterogeneous information sources such as telecommunication data, transportation data, and traffic data and a web-scale reasoning.[1]

2. Template for periodic report on Data

One of the objectives of the Urban Computing use case is to collect data and services that provide data about different activities that happen in an urban environment. To this end, since the first months of the project, we have started a gathering activity aimed at collecting those data and services.
In this section we provide some insights on the way we can structure the results of this activity, namely how we can describe the data and services we are gathering in terms of quantity, quality, usefulness and "linkability", i.e. the possibility to interlink different sources to exploit the advantages of their integration and not only of the single source.

2.1. Data sources description

In our data gathering activities, we are identifying a number of data sources that can be useful for our use case and the application we can envision to use those data.

Therefore, in our reports, we will first of all describe each data source with a number of information detailing the **data source metadata**:

- **Name**: an identifier of the data source;
- **Description**: a short overview of the data source;
- **Producer/Owner**: contact information of the producer, owner or distributor of the data;
- **Namespace/Web address**: if any, a namespace identifying the source or the Web address where more information indicating the location of the original data can be found about the source;
- **Availability:** details about the access restrictions to the data; e.g., the data are public, private, the owner requires a registration before letting download/access the source, etc.;
- **Download/Upload/Acquisition date:** last date in which the data source was accessed or checked.
- **Version:** to identify updates to the same dataset.
- **Physical size:** KB, MB, GB or TB in raw or compressed mode.
- **Nature of data type:** static, dynamical data, data stream, update rate for streaming data (e.g. every minute, hour, day, week, etc.)
- **Quality of the data source:** how accurate is it, how reliable is the information of the data source, how current / up-to-date is it (e.g., weather report data can be updated every day, every hour, etc.). If available/appropriate, some confidence measure will be included.

Other important information regards the **"semantics" of the data source**, because this kind of metadata represents a valuable clue on the use we can make of the source in our use case scenarios:

- **Typology of data:** the nature of the data source in terms of the domain or field covered; we do not aim at building a complete taxonomy of the data types, however some examples of them are: traffic information, events, geo-locations and maps, weather updates, etc.;
- **Geographic coverage of data**: because of the very nature of the Urban Computing scenario, it is of utmost importance to know what the data coverage is in terms of geographic area (e.g., country, region, city, neighbourhood, street, etc.);
- **Applied Systems**: the types of systems (or actors) that are foreseen to use this data source (e.g., transportation system, bioinformatics, tourist guide application, etc.);
- **Existence of a schema/ontology**: if any, information about the schema of the data provided (and its "translatability" in a formal model like an ontology); if an ontology is available, what kind of language and expressivity (e.g., rules and axioms) it contains;

- **Existing links with other data-sources**: if any, the existing connections or cross-references between this data source and other archives and provenance of information. This information includes compliance to standardization efforts such as Geographic Data Files (GDF) standard;
- **Possible linkage to other data-sources**: if any and if foreseeable, connections or cross-references that could be established between this data source and other archives and provenance of information;
- **Scale of data**: no. of concepts, of relations between them, and of individuals;

Finally, since we aim at using and exploiting those data within the LarKC project, in software applications built upon the LarKC platform and with the project technologies, fundamental importance must be given to the **data source format**. To this end, we envision to characterize the data source with the following metadata:

- **Format of data**: in what kind of format the data source is available; (e.g., native RDF, relational database, XML files, GIS-specific formats, REST/SOAP services, CSV export, etc.); in particular, we highlight the fact that a source is in RDF or if it can be RDF-ized by a Transform plug-in;
- **Generation method**: how the data was produced by its owner; e.g. automatic, man-made, machine-generated, etc.;
- If an RDF-based version is available or can be built:
  - **Support query language**: e.g. SPARQL, RDQL, etc.;
  - **Support triple type**: RDF/XML, N3, Turtle, etc.;
  - **No. of explicit statements**: number of triples contained in the data source;
  - **Total no. of statements**: it represents an index about the cardinality of the source; it corresponds to the number of explicit statements plus the number of inferred statements on the basis of the ontological schema.;
  - **Noise, Uncertainty and inconsistence of data**: information for the necessity of pre-processing, missing data, logical self-contradictory status, unbalanced, etc. (e.g., in an FOAF data one knows all and all others know nobody.);
- If no RDF-based version is available, some information about scale and dimension of the correspondent dataset will be made available. (e.g. proprietary query language, proprietary data type, No. of explicit lines, etc.)

2.2. Report Template for Data

The template for periodic report on data is summarized in the following table.

Table 1 Template for periodic report on data for Urban Computing use case using LarKC platform

Data Source Description			
Report ID			
Section 1	**Data source metadata**		
Name		**Producer/Owner**	
Description			
Namespace/Web Address			
Availability			
Download/Upload/Acquisition date			
Version			
Physical size			
Nature of data type			
Quality of the data source			
Section 2	**"semantics" of the data source**		
Typology of data			
Geographic coverage of data			
Applied systems			
Existence of schema/ontology			
Existing links with other data-sources			
Possible linkage to other data-sources			
Scale of data			
Section 3	**Data source format**		
Format of data			
Generation method			
Support query language			
Support triple type			
No. of explicit statements		**Total no. of statements**	
Noise, Uncertainty and inconsistence of data			
Remarks			

3. Template for periodic report on Performances

When building Urban Computing application over the collected data, we try to address specific needs of the citizens and the civil servants, i.e. the final users of the system. To this end, another important factor to evaluate the work we are carrying on in this use case relies on the performances the LarKC platform can show when addressing the requests from the users.
Apart from the application-specific information (which is not interesting for the project to be reported because we will not focus on the application itself but concentrate the request), in this section we summarize the performance requests coming from the Urban Computing use case for the LarKC platform. Some of those parameters come from the collaborative discussion with WP5[4], WP7a and WP7b partners.

3.1. Performances description

Urban Computing requires much sensor information and the "semantics" of sensor information in terms of quantity and quality which is very significant to operate whole city infrastructure while the operation cost of urban computing infrastructure keeps reasonable with efficiency[5]. After discussion within WP6 and other WP partners, the performances of Urban Computing application are screened and classified by four different ways: scale, heterogeneity, reusability and statistical measurements.

1 **Scale**: Indicators for building a platform for massive distributed incomplete reasoning at a scale well beyond what is currently possible. Information coming from multiple sensors (traffic detectors, public transportation, pollution monitors, etc.) as well as from citizens' observation (black points, commercial activities' ratings, events organization, etc.) would be used for the use case.

2 **Heterogeneity**: Urban Computing may need different reasoners for temporal reasoning, spatial reasoning, and causal reasoning. However, it does not necessarily mean that we have to develop a single but powerful reasoner which can cover all of those reasoning tasks. A system which supports reasoning heterogeneity would find a way to allow multiple single-paradigm-based reasoners to achieve the result of reasoning heterogeneity.

3 **Reusability**: The platform will not be a single turnkey one-size-fits-all solution, but will instead be a pluggable architecture usable by researchers and practitioners to develop and deploy their own reasoning components, and to combine these with components by others. Flexible and open architectures are one of major key success factors to achieve cost efficiency for urban computing application.

4 **Statistical measurements**: would be used to evaluate for example predictions of the traffic flows of a certain road of interest during the next 30 minutes or predictions how long a traffic jam will take. For such forecasts we needs to train "good" predictive models based on past observations, so called training data. A common case is that we hide a part of the observed data and use the remaining data to train models and parameters as well. After that we should judge the ability of the learnt models by applying them for forecasting the hidden data. Moreover, we would like to compare several methods in terms of those measures to figure out which are most appropriate to the given tasks.

3.2. Performances measures

Therefore, in our reports, we will first of all describe each performance factors with a number of information detailing the **scale**:

- **Materialization –** whether and to what extent forward-chaining is performed (e.g. materialized in memory only, no materialization, materialized in disk, etc). Usually materialization requires many disk spaces;
- **Complexity of the data model –** some Semantic Repositories employ extended RDF data models, e.g. including named graphs. Richer data-models are more "expensive";

- **Data size –** the size of the repository. It could be a number of triples contained in the data source or total no. of statements: it represents an index about the cardinality of the source;
- **Max time for final answer –** maximum time for final answer given a query;
- **Max time for first answer –** maximum time for first answer given a query;
- **Number of results per time unit –** number of results per time unit which will be set according to the purpose of a application;
- **Transformation time with source and destination format** – time for transforming some piece of data from one representation to another.(e.g. a plug-in would take a Natural Language Document as input and would extract a number of triples from the document to create an RDF graph, which it would return as output.) This transformation time is not restrict to natural Language doc to RDF graphs;

Secondly, other important performance factor for **heterogeneity** comprises:

- **Query complexity –** the number of the constraints (e.g. triple-pattern joins), the semantics of the query (e.g. negation-related clauses), the usage of operators which are tough to support through indexing (e.g. LIKE).
- **Reasoning complexity –** the complexity of the ontology/logical language, the specific ontology and dataset; (e.g. SHIQ, SHOIQ, etc)
- **Attempt completeness –** whether completeness is attempted(yes/no)
- **Attempt soundness –** whether soundness is attempted(yes/no)
- **Want justifications –** whether justification is needed(yes/no)
- **Knowledge representation –** how to use a symbol system to represent "a domain of discourse" (e.g. Closed World Assumption or Open World Assumption)

Thirdly, the **reusability** can be measured by the following template elements:

- **Number of plug-ins –** LarKC platform is supposed to develop several plus-ins. Urban Computing use case will utilize these deliverables. This refer to the number of plug-ins that used for this use case including third parties;
- **Number of parallel users –** Max number of supported multiple connections for the client applications;
- **Max number of supported distributed and parallel index system –** Max number of supported distributed and parallel systems for constantly growing amount of data about urban environments. This refer to the number of systems;
- **Supports concurrency –** write, read, and query triples to the data storage at the same time. (Yes/No)

Finally, we itemize some common **statistical measurements** which can be used to evaluate the performance of predictive models:

- **Accuracy –** the degree of closeness of a measured or calculated quantity to its actual (true) value.[6] **(**True Positive + True Negative) / all
- **Recall –** a measure of completeness.[7] True Positive **/ (**True Positive + False Positive)
- **Precision –** a measure of exactness or fidelity.[8] True Positive **/ (**True Positive + False Negative)
- **F-measure –** the weighted harmonic mean of precision and recall.[9] $(1+beta^2)$ precision recall / ($beta^2$precision + recall) (e.g., beta = 1)
- **ROC curve** – Receiver Operating Characteristic is a graphical plot of the sensitivity vs. (1 - specificity) [10]
- **AUC** – the area under the ROC curve;

Note that the training process is usually repeated independently on several different data sets in order to express the generality of the models. Thus, the evaluation results are represented via a mean value and a standard variance / standard error.[11]

3.3. Report Template for Performances

The template for periodic report on performances is summarized in the following table.

Table 2 Template for periodic report on performances for Urban Computing use case using LarKC platform

Performance Description			
Report ID			
Item	**Results**	**Comparisons**	**Remarks**
Section 1	**Scale**		
Materialization			
Complexity of the data model			
Data size			
Max time for final answer			
Max time for first answer			
Number of results per time unit			
Transformation time with source and destination format			
Section 2	**Heterogeneity**		
Query complexity			
Reasoning complexity			
Attempt completeness			
Attempt soundness			
Want justifications			
Knowledge representation			
Section 3	**Reusability**		
Number of plug-ins			
Number of parallel users			
Max number of supported distributed & parallel index system			
Support concurrency			
Section 4	**Statistical measurements**		
Accuracy			
Recall			
Precision			
F-measure			
ROC curve			
AUC			
Remarks			

4. Conclusions

In this document we described the purpose of the templates for the periodic report on data and performances and explained the detailed template elements for measuring the progress of the LarKC project and providing an easy access for LarKC partners and other interest groups to the status and achievement of the project. This template will be used to periodically report on data and performances at M12, M18, M26, M33, and M42. (The corresponding deliverables are D6.4, D6.6, D6.7, D6.8, and D6.11 respectively.)

The goal of these templates is to access our use case and make use of this template as a tool to achieve the followings:

1. To check the design of an integrated pluggable platform for large-scale semantic computing
2. To construct a reference implementation for an integrated platform for large-scale semantic computing
3. To demonstrate the effectiveness of the Urban Computing use case

To fulfil the above goals, we will collaborate with WP2, 3, 4, and WP5 for the validation of the use case and in order to get the performance data within the LarKC environment. The other use cases of LarKC, WP7a, WP7b will be also our close collaboration partners in order to identify commonalities and align results and performances.

What we reported in this version of the document can be updated while we find better criteria and template elements.

5. References

[1] E. Della Valle, I. Celino, D. Dell'Aglio, K. Kim, Z. Huang, V. Tresp, W. Hauptmann, Y. Huang, R. Grothmann, "Urban Computing: a challenging problem for Semantic Technologies", NeFoRS 2008, Bangkok, Thailand
[2] http://www.larkc.eu
[3] The definition of Urban Computing (A.K.A. Urban Atmosphere) can be found at http://www.urban-atmospheres.net/index.htm.
[4] We referenced WP5 deliverables D5.5.1; Definition of validation goals for the prototyping phase
[5] E. Della Valle, I. Celino, K. Kim, Z. Huang, V. Tresp, W. Hauptmann, Y. Huang, "Challenging the Internet of the Future with Urban Computing", OneSpace 2008(First International Workshop on Blending Physical and Digital Spaces on the Internet, colocated with FIS 2008), Vienna, Austria
[6] http://en.wikipedia.org/wiki/Accuracy
[7] http://en.wikipedia.org/wiki/Recall_(information_retrieval)
[8] http://en.wikipedia.org/wiki/Precision_(information_retrieval)
[9] http://en.wikipedia.org/wiki/Precision_and_recall#F-measure
[10] http://en.wikipedia.org/wiki/Receiver_operating_characteristic
[11] http://en.wikipedia.org/wiki/Standard_error_(statistics)