



LarKC

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

FP7 – 215535

---

## **D9.2 1<sup>st</sup> Report on Market Observation and Standard Assessment**

---

**Coordinator: Georgina Gallizo, HLRS**

**With contributions from: Emanuele Della Valle, Armando Beffani, Dario Cerizza, CEFRIEL; Yi Huang, SIEMENS; Bosse Anderson, AstraZeneca; Stefan Wesner, HLRS; Eyal Oren, VUA; Kono Kim, Saltlux; Atanas Kiryakov, Ontotext**

Document Identifier:	LarKC/2008/D9.2 /v1.0
Class Deliverable:	LarKC EU-IST-2008-215535
Version:	1.0
Date:	21.11.2008
State:	Final
Distribution:	Public



## EXECUTIVE SUMMARY

This deliverable aims to provide a first analysis of the technology topics related to the LarKC project, the current status of the related products in the market as well as an identification of the main standardisation bodies addressing them.

A preliminary analysis of the LarKC market has been performed, environment and context analysed and technology products and services identified. This is considered a first step in a deeper market analysis, based on “exploitable items”, to be performed in subsequent deliverables.

In a rapidly changing world, it is necessary for project researchers to be continuously aware of the status of the related technologies. Technologies used to build the project will be, as much as possible, based on mature and emerging standards. In order to satisfy this commitment, an initial identification of bodies and groups of interest, and of the involvement of the LarKC partners within them, has been performed and is delivered in this document. In subsequent deliverables, a detailed standardisation strategy will be established. This will include the LarKC consortium’s plans to influence current standards trends with concrete results of the project.



## DOCUMENT INFORMATION

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

<b>Deliverable</b>	<b>Number</b>	9.2	<b>Title</b>	1st Report on Market Observation and Standard Assessment
<b>Work Package</b>	<b>Number</b>	9	<b>Title</b>	Exploitation and Standardization

<b>Date of Delivery</b>	<b>Contractual</b>	M 06	<b>Actual</b>	21.11.2008
<b>Status</b>	version 1.0		final x	
<b>Nature</b>	prototype <input type="checkbox"/> report <input checked="" type="checkbox"/> dissemination <input type="checkbox"/>			
<b>Dissemination level</b>	public <input checked="" type="checkbox"/> consortium <input type="checkbox"/>			

<b>Authors (Partner)</b>	Georgina Gallizo, Stefan Wesner (HLRS); Emanuele Della Valle, Armando Beffani, Dario Cerizza (CEFRIEL); Yi Huang (SIEMENS); Bosse Anderson (AstraZeneca); Eyal Oren (VUA); Kono Kim (Saltlux); Atanas Kiryakov (Ontotext)			
<b>Responsible Author</b>	<b>Name</b>	Georgina Gallizo	<b>E-mail</b>	gallizo@hlrs.de
	<b>Partner</b>	HLRS	<b>Phone</b>	

<b>Abstract (for dissemination)</b>	<p>This deliverable aims to provide a first analysis of the technology topics related to the LarKC project, the current status of the related products in the market as well as an identification of the main standardisation bodies addressing them.</p> <p>A preliminary analysis of the LarKC market has been performed, environment and context analysed and technology products and services identified. This is considered a first step in a deeper market analysis, based on “exploitable items”, to be performed in subsequent deliverables.</p> <p>In a rapidly changing world, it is necessary for project researchers to be continuously aware of the status of the related technologies. Technologies used to build the project will be, as much as possible, based on mature and emerging standards. In order to satisfy this commitment, an initial identification of bodies and groups of interest, and of the involvement of the LarKC partners within them, has been performed and is delivered in this document. In subsequent deliverables, a detailed standardisation strategy will be established. This will include the LarKC consortium’s plans to influence current standards trends with concrete results of the project.</p>
<b>Keywords</b>	Market observation, Market analysis, SWOT analysis, Standardisation, Exploitation














<b>Version Log</b>			
<b>Issue Date</b>	<b>Rev. No.</b>	<b>Author</b>	<b>Change</b>
16.07.2008	0.1	Georgina Gallizo	First draft
28.08.2008	0.2	Georgina Gallizo	Input from CEFRIEL on Market Observation methodology
08.09.2008	0.3	Georgina Gallizo	Input from CEFRIEL chapter Market Observation. Assignment of responsible partners to sections.
09.09.2008	0.4	Georgina Gallizo	Input from CEFRIEL, section Analysis of the products and services (Information Retrieval). Input from WP9 telco.
15.09.2008	0.5	Georgina Gallizo	Input from VUA (Eyal Oren), section Analysis of the products and services (Reasoning Systems)
18.09.2008	0.6	Georgina Gallizo	Input from HLRS (Georgina Gallizo), section Analysis of the products and services (HPC and distributed systems); Acronyms table included
21.09.2008	0.7	Georgina Gallizo	Updated input from HLRS (Georgina Gallizo), section Analysis of the products and services (HPC and distributed systems); Updated input from Eyal Oren (VUA) on reasoning systems Input from Eyal Oren (VUA), Involvement in Standardisation bodies Input from Kono Kim (Saltlux) on Environment and Context section
06.10.2008	0.8	Georgina Gallizo	Updated input from HLRS (Georgina Gallizo), section Analysis of the products and services (HPC and distributed systems); Contribution from Cefriel (Dario) to SWOT analysis; Contribution to Kono to Context section; Input from WP9 telco.
12.10.2008	0.9	Georgina Gallizo	General formatting: List of figures, list of tables, tables names. Input from HLRS (Stefan Wesner) on HPC and distributed computing analysis of products and market analysis; Inputs from AstraZeneca (Bosse Andersson) and Siemens (Yi Huang) on Market analysis; Input from VUA (Eyal Oren) on






			stand. matrix.
16.10.2008	0.10	Georgina Gallizo	Update of chapter: Executive summary, conclusions, standardisation bodies Input from Ontotext (Atanas Kiryakov) on reasoning products and services and SWOT analysis. Version ready for quality assessment
29.10.2008	0.11	Michael Witbrock	Quality Assessment comments
30.10.2008	0.12	Georgina Gallizo	Address comments from Quality Assessor
12.11.2008	0.13	Georgina Gallizo	Address quality remarks from coordinator. Input on: <ul style="list-style-type: none"> <li>• Methodology (CEFRIEL)</li> <li>• Market Observation – Phase 1 (Astrazeneca)</li> <li>• Technology topics (HLRS, CEFRIEL, VUA, UIBK, Ontotext)</li> </ul>
15.11.2008	0.14	Georgina Gallizo	General sanity check, formatting,...
21.11.2008	1.0	Georgina Gallizo	Final version to be submitted to EC

## 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: <a href="mailto:dieter.fensel@sti-innsbruck.at">dieter.fensel@sti-innsbruck.at</a>
AstraZeneca AB		Bosse Andersson AstraZeneca Lund, Sweden Email: <a href="mailto:bo.h.andersson@astrazeneca.com">bo.h.andersson@astrazeneca.com</a>
CEFRIEL - SOCIETA CONSORTILE A RESPONSABILITA LIMITATA		Emanuele Della Valle, CEFRIEL - SOCIETA CONSORTILE A RESPONSABILITA LIMITATA, Milano, Italy, Email: <a href="mailto:emanuele.dellavalle@cefriel.it">emanuele.dellavalle@cefriel.it</a>
CYCROP, RAZISKOVANJE IN EKSPERIMENTALNI RAZVOJ D.O.O.		Dr. Michael Witbrock, CYCROP, RAZISKOVANJE IN EKSPERIMENTALNI RAZVOJ D.O.O., Ljubljana, Slovenia, Email: <a href="mailto:witbrock@cyc.com">witbrock@cyc.com</a>
Höchstleistungsrechenzentrum, Universitaet Stuttgart		Georgina Gallizo, Höchstleistungsrechenzentrum, Universitaet Stuttgart, Stuttgart, Germany, Email: <a href="mailto:gallizo@hlrs.de">gallizo@hlrs.de</a>
MAX-PLANCK GESELLSCHAFT ZUR FOERDERUNG DER WISSENSCHAFTEN E.V.		Dr. Lael Schooler Max-Planck-Institut für Bildungsforschung Berlin, Germany Email: <a href="mailto:schooler@mpib-berlin.mpg.de">schooler@mpib-berlin.mpg.de</a>
Ontotext Lab, Sirma Group Corp		Atanas Kiryakov, Ontotext Lab, Sofia, Bulgaria Email: <a href="mailto:atanas.kiryakov@sirma.bg">atanas.kiryakov@sirma.bg</a>
SALTLUX INC.		Kono Kim, SALTLUX INC, Seoul, Korea, Email: <a href="mailto:kono@saltlux.com">kono@saltlux.com</a>
SIEMENS AKTIENGESELLSCHAFT		Dr. Volker Tresp, SIEMENS AKTIENGESELLSCHAFT, Muenchen, Germany, E-mail: <a href="mailto:volker.tresp@siemens.com">volker.tresp@siemens.com</a>
THE UNIVERSITY OF SHEFFIELD		Prof. Dr. Hamish Cunningham, THE UNIVERSITY OF SHEFFIELD Sheffield, UK, Email: <a href="mailto:h.cunningham@dcs.shef.ac.uk">h.cunningham@dcs.shef.ac.uk</a>



<p>VRIJE UNIVERSITEIT AMSTERDAM</p>		<p>Prof. Dr. Frank van Harmelen,        VRIJE UNIVERSITEIT AMSTERDAM,        Amsterdam, Netherlands,        Email: Frank.van.Harmelen@cs.vu.nl</p>
<p>THE INTERNATIONAL WIC        INSTITUTE, BEIJING UNIVERSITY        OF TECHNOLOGY</p>		<p>Prof. Dr. Ning Zhong,        THE INTERNATIONAL WIC        INSTITUTE,        Mabeshi, Japan,        Email: zhong@maebashi-it.ac.jp</p>
<p>INTERNATIONAL AGENCY FOR        RESEARCH ON CANCER</p>	 <p>International Agency for Research on Cancer        Centre International de Recherche sur le Cancer</p>	<p>Dr. Paul Brennan,        INTERNATIONAL AGENCY FOR        RESEARCH ON CANCER,        Lyon, France,        Email: brennan@iarc.fr</p>



## TABLE OF CONTENTS

<b>LIST OF FIGURES.....</b>	<b>10</b>
<b>LIST OF TABLES.....</b>	<b>10</b>
<b>ACRONYMS.....</b>	<b>11</b>
<b>1. INTRODUCTION .....</b>	<b>12</b>
<b>2. KEY TECHNOLOGY TOPICS.....</b>	<b>12</b>
<b>3. THE METHODOLOGY FOR MARKET OBSERVATION .....</b>	<b>16</b>
3.1. PHASE 1 – PRELIMINARY ANALYSIS .....	17
3.2. PHASE 2 – MARKET ANALYSIS – AS IS .....	17
<i>Products and services analysis .....</i>	<i>18</i>
<i>Demand analysis – Market Segmentation.....</i>	<i>18</i>
<i>Supply and Industry analysis.....</i>	<i>19</i>
<i>Metrics and indexes.....</i>	<i>19</i>
3.3. PHASE 3 – TREND ANALYSIS – TO BE.....	19
<i>Trend Analysis.....</i>	<i>19</i>
<i>Scenario Analysis.....</i>	<i>20</i>
<i>Case Studies .....</i>	<i>20</i>
<b>4. MARKET OBSERVATION – PHASE 1: PRELIMINARY ANALYSIS .....</b>	<b>20</b>
4.1. INTRODUCTION .....	20
4.2. OBJECTIVES .....	20
4.3. OBJECT OF THE ANALYSIS.....	21
4.4. TIME PERIOD.....	23
4.5. ENVIRONMENT AND CONTEXT .....	23
<i>Political factors.....</i>	<i>23</i>
<i>Economic factors.....</i>	<i>23</i>
<i>Social factors.....</i>	<i>23</i>
<i>Technological factors.....</i>	<i>24</i>
<b>5. MARKET OBSERVATION – PHASE 2: MARKET ANALYSIS .....</b>	<b>24</b>
5.1. ANALYSIS OF THE PRODUCTS AND SERVICES .....	24
<i>Information Retrieval products and services .....</i>	<i>25</i>
<i>Reasoning products and services .....</i>	<i>26</i>
<i>HPC and Distributed Computing products and services .....</i>	<i>30</i>
5.2. SWOT ANALYSIS .....	38
<i>Information Retrieval products SWOT analysis.....</i>	<i>38</i>
<i>Reasoning products SWOT analysis.....</i>	<i>40</i>
<i>HPC and Distributed Computing products SWOT analysis.....</i>	<i>43</i>
<b>6. TARGET STANDARDISATION BODIES .....</b>	<b>44</b>
6.1. MPI FORUM.....	45
<i>Description.....</i>	<i>45</i>
<i>Relevance for LarKC.....</i>	<i>46</i>
<i>Current status.....</i>	<i>46</i>
<i>Monitoring .....</i>	<i>46</i>
<i>Participation and contributions .....</i>	<i>46</i>
<i>LarKC partners involvement.....</i>	<i>46</i>





6.2.	OASIS .....	46
	<i>Description</i> .....	46
	<i>Relevance for LarKC</i> .....	47
	<i>Current status</i> .....	47
	<i>Monitoring</i> .....	47
	<i>Participation and contributions</i> .....	48
	<i>LarKC partners involvement</i> .....	48
6.3.	OPEN GRID FORUM .....	48
	<i>Description</i> .....	48
	<i>Relevance for LarKC</i> .....	48
	<i>Current status</i> .....	49
	<i>Monitoring</i> .....	49
	<i>Participation and contributions</i> .....	50
	<i>LarKC partners involvement</i> .....	50
6.4.	W3C .....	50
	<i>Description</i> .....	50
	<i>Relevance for LarKC</i> .....	50
	<i>Current status</i> .....	50
	<i>Monitoring</i> .....	51
	<i>Participation and contributions</i> .....	51
	<i>LarKC partners involvement</i> .....	51
6.5.	OTHERS .....	51
	<i>Knowledge discovery standards</i> .....	51
7.	<b>CONCLUSIONS</b> .....	<b>52</b>
8.	<b>REFERENCES</b> .....	<b>52</b>



## List of Figures

Figure 1 LarKC Platform and plug-in model .....	21
Figure 2 Performance Pyramid for German HPC.....	31

## List of Tables

Table 1 Roles in the LarKC value network .....	21
Table 2 Information Retrieval (IR) Products and Services : the entries H, M, and L denote a high, medium, or low degree of innovation with respect to the listed technological focus.....	25
Table 3 Reasoning products and services analysis .....	28
Table 4 Reasoning products and services : RDF indexers .....	29
Table 5 HPC and Distributed Computing Solutions – Infrastructure Viewpoint.....	31
Table 6 HPC Systems classification.....	32
Table 7 HPC : Characterization of parallel programming models .....	33
Table 8 HPC : Comparison of parallel programming models .....	33
Table 9 Distributed computing products analysis .....	35
Table 10 Information Retrieval products SWOT analysis .....	38
Table 11 Reasoning products SWOT analysis .....	40
Table 12 HPC and Distributed Computing products SWOT analysis.....	44
Table 13 Standardisation matrix.....	45



## Acronyms

Acronym	Definition
DL	Description Logic
EC	European Commission
Flop/s	Floating point operations per second
GDP	Gross Domestic Product
GPU	Graphic Processing Units
HPC	High Performance Computing
ICT	Information and Communication Technologies
ISV	Independent Software Vendor
KR	Knowledge Representation
LarKC	The Large Knowledge Collider
MPI	Message Passing Interface
NG	Named Graphs
OECD	Organisation for Economic Co-operation and Development
OGSA	Open Grid Services Architecture
OWL	Ontology Web Language
P2P	Peer to Peer
PNRP	Peer Name Resolution Protocol
QoS	Quality of Service
RIF	Rule Interchange Format
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
RG	Research Group
SLA	Service Level Agreement
SOA	Service Oriented Architecture
SOAP	Simple Object Access Protocol
WG	Working Group
WP	Work Package
WSMO	Web Service Modelling Ontology



## 1. Introduction

This deliverable aims to be a first analysis of the technology environment of the LarKC project, including the current status of the related products in the market as well as an identification of the main standardisation bodies addressing them.

The document is structured in two main blocks, Market Observation and Standards Assessment.

Chapter 2 identifies the key technology topics being addressed in LarKC and serves as a basis to develop the two main parts of the document, Market Observation and Standard Assessment.

Chapter 3 introduces the methodology to be followed in performing the market observation of those technologies. In Chapter 4, the first phase of the market observation is performed under the title “Preliminary analysis”, and Chapter 5 introduces the first steps of the Market Analysis with regards to the identified relevant technology topics. In the following months of the project, a deeper Market Analysis will be performed. The expected results from the LarKC project will be mapped to potential products or services to be exploited (“exploitable items”). The methodology for market observation introduced in this document will be applied to those “exploitable items”, and that will form the basis for elaborating the exploitation strategy of the LarKC project. The result of those tasks will be reported in the subsequent releases of this deliverable, D9.3 2nd Report on Market Observation and Standard Assessment, due in M30 and D9.5 3rd Report on Market Observation and Standard Assessment, due in M42.

Chapter 6 aims to identify the main standardisation bodies applicable to the LarKC technology topics. First of all, a mapping is given between the identified technology topics and the corresponding standardisation bodies. For each of them, a brief description is given, the relevance for LarKC, possible ways of monitoring and contributing and finally the involvement of the LarKC partners is described. An intermediate release of the standardisation assessment is planned in M18, which will elaborate a concrete standardisation strategy for the LarKC project.

## 2. Key technology topics

This section aims to include a list of the key technology topics that are of interest for LarKC.

Throughout this document, three main groups of technology topics are addressed:

- **Information retrieval**

Information retrieval is the set of techniques aimed to extract specific information from various types of data such as documents, images, audio tracks, etc... Semantic technologies have been one of the most important technologies for providing solutions to information retrieval problems. Indeed, those technologies permit to extract metadata from sources and to leverage over domain models in order to solve information retrieval challenges.

The research activities performed in LarKC are promising to move the state of the art of semantic technologies a step toward the management of complex, numerous and distributed information. For this reason, information retrieval is one field where LarKC technologies may be used to solve problems so far unsolved. In this deliverable, we start observing the most prominent products and services in the information retrieval field.

- **Reasoning**

Reasoning can be understood as (the process of) inferring implicit information from a given set of data. In the context of LarKC, we are mostly interested in logical reasoning. The data over which we



reason will be mostly Semantic Web data: RDF data, using terms from corresponding RDFS and OWL schema's. We are also interested in (Horn-like) rule languages such as RIF.

Common reasoning tasks include verifying internal consistency of a dataset or knowledge base, verifying the consistency of a dataset to its corresponding schema(s), and inferring implicit facts by combining explicit facts and/or their corresponding schema(s). For example, if a knowledge base contains two facts, "a" and "not a", this knowledge base will be (in most logics) called 'inconsistent'; detecting such inconsistencies is an example reasoning task, and is useful when validating data acquisition phases. Similarly, data can be inconsistent with regards to an external definition, which could e.g. contain cardinality constraints, existential constraints, or functional dependencies. Again, validating the data against the schema can point to problems in data acquisition or interoperability problems between two data providers, but can also be used for augmenting data with values derived from the schema. In all examples, the data is enriched or validated with knowledge from a schema or ontology containing compact and explicit declarative knowledge.

All LarKC use-cases mention concrete examples of reasoning tasks and expected advantages; in the city traffic scenario, reasoning should be used to infer higher-level abstractions from low-level sensor data and to help decide on best routes for the users; in the medical use-cases, reasoning is used to infer as-yet-unknown relations between compounds, genes, pathways, etc. from a large body of literature and test data, but also to align data from different datasets (using eg mapping rules between genomic databases).

Given the goals of the LarKC project, reasoning technology is a major topic for LarKC. How to perform logical reasoning in the languages of our interest is relatively well-known, how to do so efficiently in the presence of huge volumes of data is still an open question. Ontotext, one of the leading commercial vendors of highly scalable reasoning systems for RDF and OWL is member of the LarKC consortium, and together with them the project monitors current suppliers of reasoning technology, the market potential and business value of scalable reasoners, and the ongoing efforts in standardization around reasoning languages and protocols.

- **High Performance Computing and Distributed Computing**

- High Performance Computing aims to improve the performance or solve complex computing problems. Although distributed computing can be considered a particular case, usually the term high performance computing refers to the use of supercomputers of computer clusters for the execution of complex processes, making use, normally, of parallel programming techniques. High performance computing is being considered within the LarKC project as a technology to improve the performance of certain plug-ins. For this purpose, parallelization programming models must be taken into consideration at development time.
- Distributed Computing is a computing model that involves multiple computers that may be remote from each other, in order to solve a single problem. There are many different types of distributed computing approaches (client-server, peer-to-peer, thinking@home, cloud computing, ...) and many challenges to overcome in successfully designing the appropriate solution. The main goal of a distributed computing system is to connect users and resources in a transparent, open and scalable way. LarKC will take advantage of distributed computing technologies in order to achieve an improved performance of certain plug-ins. Different approaches may be considered:
  - split the algorithm in tasks that can be executed in parallel, in remote locations, without the need of frequent communication between them, accessing the same data (that can be replicated locally or accessed remotely)
  - split the data in different chunks that are sent (or easily accessible) to remote locations, every of them running the same algorithm
  - combination of the two previous approaches



Orthogonal to these main groups there is another set of technologies to be considered within LarKC, which will be mapped at the end of the document onto relevant standardization bodies addressing its components:

- **Data formats and language profiles**

The data formats and data models specify the basic structuring of the data relevant to the reasoning tasks. Those determine many engineering aspects of the reasoning platforms and tools, most notably, the data structures used for representation, management, storage, and indexing. RDF is widely accepted as a data model in most the contemporary KR formalisms. Still, the semantic web community, driven to a major extent by the tool providers and the users, has recognized the need of extensions of the standard. The most notable extension are the so-called Named Graphs (NG), which allow for efficient management of provenance. NG made their way into the specification of the SPARQL query language. While NG appear relatively simple extension of the RDF model, they have considerable impact on the computational complexity (both with respect to time and space) of all sorts of data management and reasoning tasks. Further, it appears that the semantics of NG is still underspecified and further extensions are necessary for efficient dealing with task-specific metadata (e.g. such relevant to the reasoning process).

Although RDFS and OWL are widely accepted as schema and ontology modeling languages, building a properly structured, understandable, and manageable ecosystem for reasoning tools requires finer grained distinctions between sub-languages (called dialects, fragments, or profiles), as well, as proper bridges to rule languages and rule-based systems. Both OWL and OWL 2 provide definitions of such sub-languages, but those are not sufficient for proper management of the expectations of the customers with respect to tools that implement more specific language profiles due to the principle nature of the implemented reasoning approach or due to design decisions related to language features which are inappropriate for considerable groups of applications. Such examples are:

- Treating domains and ranges of properties as evidence for the type of the resources is inappropriate in many data management scenarios, where those have to be used for consistency checking. Inference based on domain and range is appropriate in the case of several RDF statements used for annotation of an HTML page – the basic scenario considered in the design of RDFS. However, there is a wide range of data integration scenarios, where RDF(S) and OWL are used for data integration in more controlled environments (e.g. integration of databases in the life sciences). In such cases domain and range need to be interpreted as consistency restrictions.
- The reflexivity of owl:sameAs and the fact that each URI should be formally treated as member of rdfs:Resource and owl:Top, forces the tools to “consider” three more statements for each node of an RDF graph. Considering could mean different things for different systems and approaches (e.g. forward- vs. backward chaining, fetching query results, etc.), but in all cases, compliance with this aspects of the standard imposes considerable performance overheads. While these features of the language make (some aspects of) the semantics of OWL look well-grounded and self-contained, the associated performance penalties are unjustified for wide range of applications, which do not count on inference with respect to this aspects of the semantics.

Both data models and language profiles are addressed in a greater detail in LarKC deliverable D1.1.3 *Initial Knowledge Representation Formalism* [46], in the context of the definition of the conceptual framework of the project. We are discussing them here, because the corresponding features of the reasoning tools and technologies have considerable impact on their performance and applicability for different types of applications. Thus, they are directly relevant to their positioning on the market, as well, as for the segmentation of the market itself. For the sake of example, a DL reasoner, dealing with standard RDF model, have very few application in common with a data management platform, which offers tractable reasoning on top of extended RDF data model. A comprehensive analysis of the market should not consider them as competitors and compare them directly.



- **Service description and invocation**

Semantic Web Services are an extension of the Semantic Web, in a similar way as ordinary Web services are an extension of traditional Web resources, that go beyond static documents by allowing to invoke some action with an effect on the world and provide a distinct functionality. The key difference is that Semantic Web Services are not only described at a purely syntactic layer but are also annotated with semantic which makes data machine-interpretable.

The value of adding semantic descriptions to a service is that many of the tasks that are usually performed manually should be performed automatically, i.e. location, invocation and composition of a Web service.

In general two different levels can be identified in stack of a Web service description language, namely a semantic and a non-semantic level.

Core points that are usually addressed in the description of a Web service are:

- The data model used for input and output.
- Functional and Non-functional descriptions.
- Behavioural Descriptions of the Web service.

At the non-semantic level several standards cover these concerns. WSDL is used to describe the interface of a Web service and its operations (functional description), while a set of specifications such as WS-Security, WS-Reliability, ... are concerned with the non-functional description of a service. The common data model/exchange format employed in these standards is XML schema along with SOAP, HTTP, ... as underlying communication protocol.

Considering semantic descriptions of Web service currently two of the main approaches are WSMO and OWL-S. Both of them form comprehensive frameworks which model services in a top-down fashion, which means that usually first services are modelled by specifying their semantics to cover the points above, and then grounding them in concrete invocation and communication technologies (see above).

A recent, comparatively light-weight approach to add arbitrary semantic annotations to existing WSDL descriptions is SAWSDL (Semantic Annotations for WSDL and XML Schema). WSMO-Lite in turn is a concrete proposal to provide these semantic descriptions layered on top of SAWSDL and thus can be used to model service in a bottom-up fashion, starting from existing WSDL interfaces.

The relevance of these technologies for the LarKC platform is actually two-fold:

- 1.) Plug-ins on the platform share many characteristics with Web services, and could in fact be implemented as such. This implies that they also need to be described appropriately in a very similar fashion in order to construct a concrete LarKC pipeline that satisfies certain requirements.
- 2.) Plug-ins need to be invoked within the LarKC platform, and as well for this tasks existing standards (protocols, serialization formats, ...) need to be considered.

- **Distributed service architectures**

A distributed service architecture consists on a number of loosely coupled services (each of them including self-contained functionalities) that, regardless of the geographical location (and ideally regardless of the platform, programming language and other used technologies), can interoperate seamlessly. This allows, from a collection of more or less simple services, to compose more complex services. Besides reusability of the independent services, it is important to consider other issues such as interoperability, communication and synchronization issues.

SOA is a paradigm for the design and realization of a distributed architecture. The concrete SOA solution must be designed for every concrete situation, considering the specific context and requirements. Other technologies to design and deploy a distributed service architecture are for example P2P, thinking@home, etc

A distributed service architecture is being considered within LarKC for the deployment of geographically distributed plug-ins, developed with different technologies, that interoperate between them and with the LarKC Collider Platform, for the execution of the pipeline that will provide the



end user with the answer to his query. As stated above, one possibility for the implementation of the LarKC plug-ins in a SOA is using Semantic Web Services technologies.

- **Resource description and invocation**

Analogous to the service description, resources can be also described with semantic technologies. Within the Grid (and distributed computing) community, an effort is ongoing in modeling Grid resources, as part of the work performed in the OGF standardization body. Resources in a distributed architecture need to advertise their capabilities and the consumer services, or jobs, according to Grid terminology, need to know these capabilities in order to manage and use them in an appropriate way. At the same time, the jobs need to express a set of requirements, in terms of resources, necessary for their correct execution. Therefore, a job needs to run where its resource requirements are satisfied or can be provisioned. Jobs may be, for example, parallelized applications running on multiple nodes in a cluster. So requirements may be expressed, for example, as job A needs a minimum of three processors with n CPU seconds or m seconds of wall clock time available [43].

LarKC will take example of this modeling and will follow the work in OGF for describing the resources requirements of the LarKC plug-ins, as if they were jobs in a grid. In LarKC terminology, plug-ins will be modeled as services in a distributed architecture. Therefore, every plug-in will be characterized by a set of functional parameters, describing its functionality, plus a set of QoS parameters/requirements, such as minimum/maximum number of cluster nodes, minimum/maximum memory requirements, performance function (e.g. identified resources/second). This work is currently being performed in LarKC WP1 and WP5.

- **Parallel programming models**

For tightly coupled parallel computations, the use of parallel programming models (as described in 5.1) developed within the HPC research community is generally the right solution. The different programming models offer a combined method for distributing the work and data to the processors and the memory, and for synchronization between the processes and for communication of data between the processes. Parallel programming models are of relevance for LarKC for the development of certain plug-in algorithms. The choice of the parallel programming model and the way this is applied to the development of the algorithm will have impact in the plug-in performance. Before choosing the programming model, it must be analysed the suitability of the algorithm to be parallelized. It may happen that the performance of the algorithm is not improved through parallelization. Algorithms programmed to be split in parallel tasks may be executed either in high performance computing environments (such as a cluster) or in distributed computing environments, depending on different factors, such as the degree of coupling of the parallel tasks, location of the data, etc.

### 3. The Methodology for Market observation

Before starting the analysis, it is necessary to define the methodology adopted to support the Market Observation process. In our study, we use a step-by-step approach. After stating the objectives of the analysis, we will detail what we want to analyze and describe it using several indices; those indices will then be assessed over an identified period of time. Our methodology consists of three main phases:

- **Preliminary Analysis:** in order to set up the analysis, this analysis details and contextualizes what is to be evaluated.
- **Market Analysis (or AS IS Analysis):** assesses the object of the analysis at the beginning of the evaluation period and describes it using several metrics and indices.





- **Trends and Scenarios Analysis (or TO BE Analysis):** describes the evolution of the object of analysis over the evaluation period, measuring and assessing over time the metrics and indices defined in the previous phase, understanding how certain actions could modify the course of this evolution, and proposing corresponding scenarios.

In this deliverable, we will develop the contents of the Preliminary Analysis and some aspects of the Market Analysis with respect to the identified relevant technology topics.

### 3.1. Phase 1 – Preliminary analysis

This phase can be considered as a “requirement analysis” and it has been implemented as follows:

- **Objectives:** in this step we try to answer to the questions “Why are we doing the analysis?” and “What are the objectives we would like to reach?”. It is important to clearly state these objectives since our work will be evaluated against them. Furthermore, these objectives establish a baseline for the definition of representative metrics and indices quantifying and describing the phenomena for assessment. A technique that can be used to derive the right indices from the formulated objectives is the Goal Question Metrics (GQM) technique. These objectives could focus on economic, social, political, technological, cultural and other factors.
- **Perimeter or object:** the perimeter of the analysis indicates what we do, and what we do not, have to analyze; in this step we answer the questions “For which geographical area should we set up the analysis?”, “Which markets should we consider?” “Which characteristics of the object to be analyzed are we interested in?” “Who are the users?” “Who are the stakeholders?”. The definition of the object of analysis can also indicate inputs, outputs, users, activities, conditions, *etc.* and can be described by a model.
- **Time period:** besides understanding “What to analyze”, we should understand “When to set up the analysis”. This means identifying the starting and the final date of the time period or the time periods over which we want to evaluate the object.
- **Environment or context:** there are many factors outside the perimeter of the analysis which can affect the object of the analysis; these factors can be, for instance, external stakeholders, market conditions, laws, policies, regulations, constraints, assumptions etc. It is important to identify relationships and dependencies between what is inside the perimeter and what is outside, in order to better understand the root causes of phenomena and their evolution over time. A useful technique to be used in order to describe the environment or context is the **PEST analysis** [1].

### 3.2. Phase 2 – Market Analysis – AS IS

After detailing the object of the analysis, we can start the market analysis. In marketing, the market refers to a group of the consumers or organizations that are interested in some products or services, that have the resources to acquire these products or services, and that are permitted by law and other regulations to acquire them. We can approach the analysis of a market by separately considering the following factors:

- the main characteristics of the products or services,
- the demand for such products or services, and
- the supply of such products and services.



In this phase, the market analysis should be done at the **beginning of the chosen time period (AS-IS)**, while in the following phase we will consider its evolution over time.

### Products and services analysis

The object of the analysis has been introduced and described in the preliminary analysis. In this paragraph, according to a market perspective, we detail some characteristics of the object to be analyzed, as the products and services offered to prospects and clients. The steps to be followed are the following:

- Group the products and services of the object of the analysis by type or common characteristics.
- Detail the characteristics of such products and services:
  - Elaborate a **Value Proposition** for the products and services classified in the previous step.  
A value proposition is a statement that summarizes why consumers and organizations should buy or use such products or services.
  - Implement a **SWOT analysis [2]**

### Demand analysis – Market Segmentation

The demand is the amount of the products and services requested by organizations (business demand) or individuals (consumer demand) at a given price. Demand for a product or service is determined by many different factors beyond its price, such as the prices of substitute goods and complementary goods.

In order to execute an effective analysis of the demand we can use the market segmentation approach, which is the process of dividing a total market into market groups consisting of organizations or individuals who have relatively similar needs and behaviours.

A market can be segmented on more than one basis, and industrial markets are segmented somewhat differently from consumer markets as described below:

- **Consumer market segmentation:** a basis for segmentation is a factor that varies among groups within a market, but that is consistent within groups. One can identify four primary bases on which to segment a consumer market:
  - **Geographic segmentation** is based on regional variables such as region climate, population density and population growth rate.
  - **Demographic segmentation** is based on variables such as age, gender, ethnicity, occupation, income and family status.
  - **Psychographic segmentation** is based on variables such as values, attitudes and lifestyle.
  - **Behavioural segmentation** is based on variables such as usage rate and patterns, price sensitivity, brand loyalty and benefits sought.
- **Business market segmentation:** while many of the consumer market segmentation bases can be applied to businesses and organizations, the different nature of business markets often leads to segmentation on the following bases:
  - **Geographic segmentation** is based on regional variables such as customer concentration, regional industrial growth rate and international macroeconomic factors
  - **Customer type** is based on factors such as the size of the organization, its industry, position in the value chain, or its sales.
  - **Buyer behaviour** is based on factors such as loyalty to suppliers, usage patterns and other size.



## Supply and Industry analysis

The supply is the quantity of products and services that available to meet a demand, or that are available for purchase at a given price. Supply analysis means understanding which the providers of such products and services are and assessing the characteristics of their industries.

According to Michael Porter's five forces model [5] the elements to be considered in the analysis are the following:

- **Entry barriers:** these are the “are obstacles in the path of a firm which wants to enter a given market such as economies of scale, brand identity, switching costs, capital requirements, access to distribution, proprietary learning curve, government policy, etc.
- **Bargaining power of suppliers:** it is the suppliers' ability to influence the prices of supplies and it is determined mainly by the followings factors: differentiation of inputs, switching costs of suppliers and firms in the industry, presence of substitute inputs, supplier concentration, importance of volume to the supplier, cost relative to total purchases in the industry, impact of inputs on cost or differentiation, threat of forward integration relative to threat of backward integration by firms in the industry.
- **Bargaining power of buyers:** it is the buyers' ability to influence the prices of final products and it is determined mainly by the followings factors: buyer concentration versus firm concentration, buyer volume, buyer switching costs relative to firm switching costs, buyer information, ability to backward integrate, substitute products.
- **Threat of substitute products:** it is the availability of substitutive products which can replace ours and it is mainly determined by the following factors: relative price performance of substitutes, switching costs and buyer propensity to substitute.
- **Rivalry:** it is represented by the competition with other organization and it's mainly determined by the following factors: industry growth, fixed (or storage) costs/value added, product differences, brand identity, switching costs, concentration, diversity of competitors, exit barriers.

## Metrics and indexes

In the previous paragraphs, we examined the characteristics of products and services offered to the market, the characteristics of the demand of such products and services and the characteristics of the supply of such products and services. The aim of this paragraph is to identify some metrics and indices that can represent the main characteristics of what we want to analyze. Metrics and indices should:

- represent the objectives of the analysis,
- consider the main characteristics of the scope of the analysis,
- be simple to use, and
- be small in number in order to be handled effectively.

These indices should represent all the aspects considered in the objectives (economic factors, social factors, political factors, technological factors, etc...). Example indices include:

- economic indices: market share of a product, sales per year, etc,
- social indices: number of new users per year,
- political indices: number of new standards per year, and
- technological indices: number of innovative products per year.

### 3.3. Phase 3 – Trend Analysis – TO BE

#### Trend Analysis

In Phase 2, we have analyzed the market at the beginning of the selected time period. Now, we introduce the variable “time” in order to understand the market' evolution. Using the year as a basis of analysis, we can



make hypotheses about future years to be analyzed and estimate the possible outcomes. The detailed approach to be followed is:

- Apply a “root causes analysis”; this is a class of problem solving methods aimed at identifying the root causes of problems or events. This allows identification of the relationships between the object and internal and external factors. A possible technique to be used is the Ishikawa diagram [25].
- Make some assumptions about the internal and external factors which can influence the evolution of the identified indices.
- Estimate the expected value of identified indices over the chosen time period.

### **Scenario Analysis**

In the previous paragraph, we have described the implementation of a root causes analysis and some assumptions which can refer to a “standard” or “average” scenario. Now, we identify alternative scenarios for changes in the assumptions. The steps to be followed are:

- identify a limited number of variations of the assumptions,
- determine the effects using the results of root causes analysis, and
- describe the alternative scenarios, assessing the identified indices over the chosen time period.

At the end of the scenario analysis, it is possible to assess the accuracy of the results by developing and applying a sensitivity analysis on the input parameters.

### **Case Studies**

In the previous paragraphs we focused on the assessment of the state of the art and evolution of some particular groups and typologies of products and services. The aim of this section is to present some case studies which can detail the concepts previously formulated. These case studies could be significant and representative of the corresponding groups and typologies. The number of these samples depends on the level of representativeness of the single case studies: if one or two of them can express all the characteristics of their group, they can be sufficient; otherwise we should search for other examples.

The analysis of these cases should detail the results previously developed according to the following scheme:

- Presenting the particular product or service.
- Positioning of the product or service in the market.
- Understanding its evolution over time, considering possible scenarios according to the developments of the reference market and industry.

## **4. Market Observation – Phase 1: Preliminary analysis**

### **4.1. Introduction**

The preliminary analysis can be considered as the “requirements analysis” where objectives, objects, time period and context are defined. The market for LarKC, “an integrated platform for semantic computing on a scale well beyond what is currently possible” is obviously not mature. Therefore this preliminary analysis will be revisited during the project’s life cycle and updated as the market evolves.

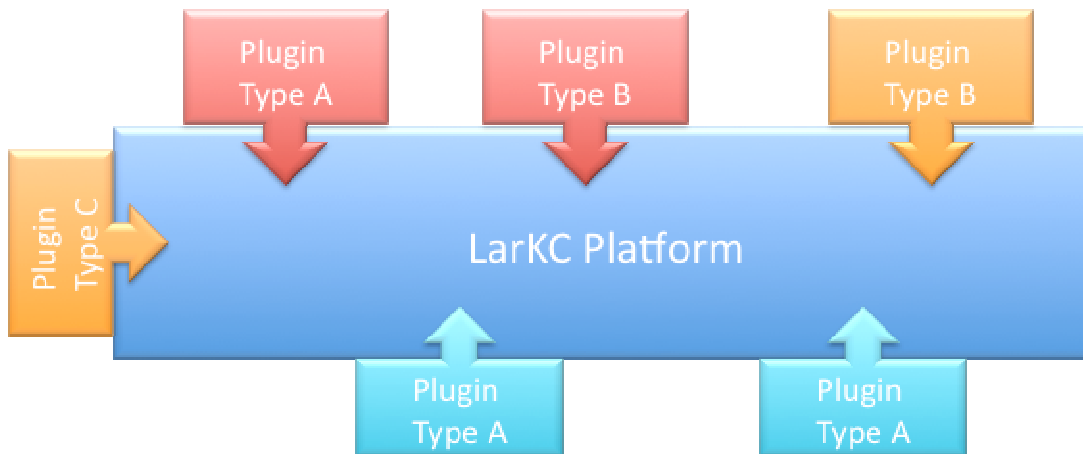
### **4.2. Objectives**

The objective with the market observation is to establish measures that make possible for the EC and the consortium partners to evaluate the outcome of LarKC.

The EC invests €1.8 billion via its Information Society Technology theme for strategic research in priority areas in information and communication technologies (ICT) to promote innovation and technical leadership. LarKC is one project funded, and the market analysis will provide measures to measure LarKC's success.

### 4.3. Object of the Analysis

The approach applied for the market analysis discussed below assumes a simplified view on the LarKC platform as shown in Figure 1. The assumptions taken are that the platform allows plug-ins of several types to be added to the platform including several (potentially competing) offers for the same type.



**Figure 1 LarKC Platform and plug-in model**

The different plug-ins are provided by different organizations and have different capabilities even if they are from the same type.

In such a model different roles in the value network might be present. Selected roles (e.g. all roles related support or training are neglected) are described in the table below:

**Table 1 Roles in the LarKC value network<sup>1</sup>**

<i>Role</i>	<i>Description</i>
<b>Platform Operator (Service)</b>	The platform operator hosts the core LarKC infrastructure services on its own or on resources provided by Commodity Resource or Cloud Providers. These services include the possibility to register plug-in providers and allow consumer to search for them. Indicative Business Models could be to request fees from plug-ins provided through the platform or by advertisements on/subscriptions to the entry point to the platform e.g a web based portal.
<b>Platform Developer</b>	The platform developer own parts of the software necessary to operate the platform or has provided commercial implementations with better quality (e.g. increased performance or more sophisticated Graphical User Interface). A possible model is to license these components to Platform Operators.
<b>Plug-in Developer</b>	Plug-in Developers have either developed a plug-in from scratch or have improved an existing plug-in for the LarKC platform. The

<sup>1</sup> At the time of writing this deliverable, the LarKC value network is under discussion within WP9. Therefore, this table may be modified in future versions of WP9 deliverables.



Role	Description
<b>Commodity Resource Provider</b>	<p>anticipated model is two-fold. Either the software will be licensed to the Plug-in Service Provider or is assumed to be used internally by being at the same time in the role of the Plug-in Service Provider.</p> <p>A commodity resource provider is offering simple fixed server infrastructure as typical for web hosting providers. The provided resources are off-the-shelf and mid-range performance. The offered resources are rather fixed and cannot easily scaled up or down.</p>
<b>Cloud Resource Provider</b>	<p>In contrast to the Commodity Resource Provider the Cloud Resource Provider offers low-end to mid-range computing or storage capacity that is able to be scaled up or down as needed within given limits. The dynamic provision model lead to a higher costs and is not competitive if the needed computing/data is of rather fixed static nature.</p>
<b>Donation Resource Provider</b>	<p>Similar to models as used on BOINC a donation resource provider is offering its (typically) low end resources at no cost as a donation.</p>
<b>HPC Resource Providers</b>	<p>As for certain plug-ins or for specific use cases e.g. particular large data sets the performance offered by the three resource providers above might not be sufficient. In such cases the precious infrastructure offered by HPC Resource Providers is offering advantages such as reduced round trip time or the capability to solve the problem at all.</p> <p>The special nature of the resources (e.g. high end network interconnect, very high memory bandwidth,...) makes them not competitive in price to commodity resources if this special properties are not needed.</p>
<b>Plug-in Service Provider</b>	<p>A plug-in service provider is using the software components (plug-ins) developed by the plug-in developer, is using the features of the platform and is using internal or external resource providers of the different types as outlined above.</p>
<b>Application Developer and Application Provider</b>	<p>A diverse range of applications may be developed and provided (by the Application Developer and the Application Provider, respectively), using a combination of different Plug-in services, platform and resources. Some examples are the following:</p> <ul style="list-style-type: none"> <li>• <b>Real time city:</b> An application meeting the requirements of the use case, “Real Time City” in WP6.</li> <li>• <b>Semantic Integration for Early Clinical Development:</b> An application meeting the requirements of the use case WP7a.</li> <li>• <b>Carcinogenesis Reference Production:</b> An application meeting the requirements of the use case WP7b</li> <li>• <b>Other applications in different market sectors, that will be explored in future Market Observation deliverables</b></li> </ul>



#### **4.4. Time period**

The market for LarKC is not mature. Therefore we will continuously evolve analysis methods, measures etc. in parallel with project development. The final data collection and analysis will be executed and delivered at project closure. The time period to be considered in the analysis is **2008-2015**.

#### **4.5. Environment and Context**

For the market observation we have identified a number of important environmental factors that will be considered during the analysis. We are basing our analysis on the Political, Economic, Social, and Technological (PEST [1]) analysis concept stratifying macro-environmental factors of importance for LarKC into the four categories.

At this stage we have identified some key macro-environmental factors, however we will continue to refine/identify the most important factors throughout the project lifetime.

##### **Political factors**

Political factors of relevance for the LarKC market:

- The European Commission efforts to promote research and initiatives in the fields of Web 2.0 and Semantic Web. This encourages the international collaborative activities of researchers and industry, including standardisation and integration of data collected in different countries.
- Seamless national borders (including Treaty of Lisbon [17]) need integration of heterogeneous data environment for data exchange under semantic enabled manner.
- Many scientific fields address questions on a global scale and need institutional arrangements for the integration and management of information coming from different organizational sources.
- Environmental regulations are imposing more and more restrictions in all industry sectors. The intelligent use of technology, such as the semantic technologies, is supporting industry on this matter. An example of application is the real time monitoring and alerting to take proper actions against the increase of pollution.

##### **Economic factors**

Economic factors of relevance for the LarKC market:

- The increasing investment in research and innovation is accelerating the progress of technology.
- Globalisation of world economy.
- Explosion in costs for health care and life extension due to an aging population.
- Open source policies versus proprietary licenses is an ongoing discussion, affecting the development of technologies and the benefits of the industry.

##### **Social factors**

Social factors of relevance for the LarKC market:

- In today's information society it becomes more and more necessary the development of solutions to manage the high amount of available information, coming from heterogeneous and distributed sources.
- Increase of metropolitan population, globalization over all the societies.
- The development of the mobile communications and computation industry, especially the growth in cellular phone use, and the resulting change of life style requires the intelligent technology including mass data processing.
- The trends of aging in the developed world and the increasing numbers of youth in the developing world may be creating an economic and social time bomb.



- Increasing needs for health care and life extension for an aging citizenry.
- Biomedical advances, such as the DNA revolution, require the use of more advanced technologies to process all available data.
- Increase in environment regulation according to climate change. Strengthening of international environment regulation such as Kyoto Protocol and Ozone Agreement is requiring conducting the real time monitoring for the deflation of CO<sub>2</sub> and air pollution
  - It increases the needs for solutions such as sensor-networks and context-aware technology-based urban computing.
  - City planning and development pay much more attention to designing and operating for low energy consumption and a pollution free environment.
- There is an increasing need for European-wide (and international-wide) data, interoperability of data and languages to harmonisation of data access policies, the standardisation of digitation processes as well as interoperability between the humanities and the social sciences in general [28].

### **Technological factors**

Technological factors of relevance for the LarKC market:

- Lots of research efforts and industrial investment are focusing on knowledge processing and knowledge-enabled systems which are currently expected to improve human life and society.
- Recently the volume of data and information circulated and accumulated through various networks including web and mobile-data is explosively increasing, and the real time processing of mass and heterogeneous data will present huge challenge in the near future.
- Artificial intelligence is emerging which, when combined with biotechnology and nanotechnology, may very well transform the concept of what it means to be human [15].
  - These technological needs are lead to high expectations for semantic technology and massive-scale reasoning.
  - Interdisciplinary research is required to develop artificial intelligence and meaning-based computing systems.
- The progress of technology is also affecting positively scalable reasoning. According to the results presented in [47], while the best scalability achievements in the end of year 2007 were in the range up to 1 billion RDF statements, now we have results from several tools that approach the 10 billion statements threshold.
- The availability of high performance and distributed computing and data networks adds to the capability of widening the extent and diversity of the data collection and data elaboration [28].

## **5. Market Observation – Phase 2: Market analysis**

In this chapter a first approach to the market analysis is performed. As this is the first report on Market Observation out of a planned series of three in the LarKC project, a high level market analysis is performed, as a first step to approach the market status of the LarKC related products and services. A more in-depth analysis as well as a Trend Analysis (the third phase of our Market Observation methodology) will be performed in deliverables D9.3 and D9.5 (the 2<sup>nd</sup> and 3<sup>rd</sup> Reports on Market Observation and Standard Assessment).

### **5.1. Analysis of the products and services**

The object of the analysis can be detailed by clustering it by types of products and services.





**Information Retrieval products and services**

The scope of this section is analysis of current innovation trends in Information Retrieval products and services. In particular we focus on Web Search Engines as the predominant class of product and service in the Information Retrieval field. As we shall see, innovation within Information Retrieval does not necessary mean introducing new technologies for crawling, indexing, searching and ranking. Indeed, several innovative Web Search Engines are reaching a higher level of effectiveness (c.f.. more traditional solutions) by providing enhanced visualization and user involvement.

Hereafter we shall characterize four evolutionary trends that can be used as a way to cluster innovative Web Search Engines before performing a SWOT analysis.

- **Query pre-processing:** when the user provides a search query the search engine analyzes it and alters it. A basic, well-known technique in this category is the use of stop-words (such as “the”, “a”, “of”, etc.): words that are deleted from the query because they would match numerous resources of little interest for the user. A slightly more advanced technique involves expanding the query by adding alternative terms that are similar to those requested by the user; basic techniques of this kind apply various transformations such as lowercasing, removing plurals, or stemming. More innovative techniques apply semantic analysis to the search query and try to determine what the use “means”.
- **Specific Media Focus:** Search Engines that fall into this category let the user know in advance that the result will be images or movies or blogs or people’s profiles or something else of a predefined type. When the media is known in advance, special techniques, which only work for the specific source, can be applied. For instance, if media is “image”, image processing techniques can be used both for indexing and for searching the images.
- **Core algorithm improvement:** Search Engines whose developers are working on improving crawling, indexing and searching techniques belong to this category. Collaborative filtering and other similar techniques that try to harness “collective intelligence” are well-known cases of this kind. Another innovative approach is the employment of semantic technologies to realize focussed crawlers, conceptual indexers and semantic search functionalities.
- **Post-processing and results visualization:** several innovative Search Engines are introducing techniques that mainly focus on post-processing the search results and facilitating the user’s inspection of the result set. Clustering and facet browsing are two examples of such techniques that are increasingly used.

**Table 2 Information Retrieval (IR) Products and Services : the entries H, M, and L denote a high, medium, or low degree of innovation with respect to the listed technological focus.**

Product/Service name	URL	Query pre-processing	Specific Media Focus	Core algorithm improvement	Post-processing and results visualization
HAKIA	<a href="http://www.hakia.com/">http://www.hakia.com/</a>	H		M	L
ASK	<a href="http://www.askx.com/">http://www.askx.com/</a>	H			L
ANSWERBUS	<a href="http://www.answerbus.com/">http://www.answerbus.com/</a>	H			



<b>COLLARITY</b>	<a href="http://www.collarity.com/">http://www.collarity.com/</a>	<b>M</b>		<b>M</b>	<b>H</b>
<b>SEARCHMASH</b>	<a href="http://www.searchmash.com/">http://www.searchmash.com/</a>		<b>M</b>		<b>H</b>
<b>TECHNORATI</b>	<a href="http://technorati.com/">http://technorati.com/</a>		<b>H</b>		<b>H</b>
<b>RIYA</b>	<a href="http://www.riya.com/">http://www.riya.com/</a> <a href="http://www.like.com/">http://www.like.com/</a>		<b>H</b>	<b>H</b>	
<b>CHACHA</b>	<a href="http://www.chacha.com/">http://www.chacha.com/</a>	<b>L</b>		<b>H</b>	<b>M</b>
<b>SPOCK</b>	<a href="http://www.spock.com/">http://www.spock.com/</a>		<b>H</b>	<b>M</b>	<b>L</b>
<b>HEALIA</b>	<a href="http://www.healia.com/">http://www.healia.com/</a>			<b>M</b>	<b>H</b>
<b>QUINTURA</b>	<a href="http://www.quintura.com/">http://www.quintura.com/</a>			<b>L</b>	<b>H</b>
<b>CLUSTY</b>	<a href="http://clusty.com/">http://clusty.com/</a>			<b>H</b>	<b>H</b>
<b>KARTOO</b>	<a href="http://www.kartoo.com/">http://www.kartoo.com/</a>			<b>H</b>	<b>H</b>

### Reasoning products and services

We analyse current reasoning systems for Semantic Web data, including research projects, prototypes and commercial and industrial offerings. Since the field is extremely innovative and dynamic, without clear stabilisation, characterising the products is not straightforward. On the one hand, it is unclear which product features the market is interested in, on the other hand, the technology is not mature enough that abstract features can be easily recognised.

Below, we have listed an overview of some of the most prominent RDF stores and reasoners in the field. The terminological distinction between stores and reasoners is common, but somewhat arbitrary; generally speaking "RDF stores" focus on data storage and retrieval with very limited reasoning (RDF and RDFS, often though hard-coded support for specific entailments), while "reasoners" focus less on scalability and more on higher-level semantics (OWL). Semantic repositories fit in the middle, as a special sort of RDF stores, offering both scalability and light-weight reasoning. RDF stores can replace DBMS in many applications, while on the other hand their functionality can be matched by RDF "wrappers" of DBMS. These stores and reasoners are relevant to LarKC as potential components (wrapped to support the LarKC API) for storing and reasoning with RDF data, but also as competitors in handling large amounts of RDF data.

In the overview, we have characterised each of the stores and reasoners by some common properties. First, the level of semantics supported by the stores (i.e., the semantics under which the data are interpreted during query answering), which ranges from pure RDF (interpreting the triples almost as just the triples themselves), to RDFS (interpreting the RDF Schema information as defined by the RDFS standard), to the various layers of OWL. The term RDFS++, is used by some of the vendors without a formal definition, to denote support for RDFS plus partial support for some few OWL primitives (usually, owl:inverseFunctionalProperties and owl:sameAs). The term OWL-Horst refers to the fragment of OWL identified by ter Horst (Journal of Web Semantics, 2005), commonly chosen by the developers of semantic repositories for its good computational properties; there are also several extensions of this fragment, adding support for extra entailments without substantially changing its complexity (e.g. OWL Prime in ORACLE 11g and OWL-Max in OWLIM). These systems are all based on R-entailment – a simple rule language for RDFS which resembles horn clauses. A more elaborate discussion of the supported language fragments as well as extensive analysis of the performance and scalability of the state-of-the-art reasoning-related products can be found in [42].

We indicate the owners (companies or research institutions) and license terms of each product, the query language supported (almost all stores now support SPARQL), and whether they support full-text indexing of terms in literals (relevant for information retrieval scenarios). We list whether they have built-in support for navigating (browsing) through the contents of their store.



Finally, we also list some prominent services that index RDF data on the Web, and allow locating that data by keyword, entity, triple pattern, or full SPARQL query. Where known, we indicate the amount of data currently stored by these indexers at the time of writing (Sep. 2008), in number of sources (RDF documents) indexed. These indexers are relevant to LarkC since they may act as 'identify' components in the LarkC pipeline: they help locate existing RDF data.



Table 3 Reasoning products and services analysis

	product	Link	owner	semantics	queries	license	full-text	naviga-tion
<b>RDF Stores</b>	Jena	<a href="http://jena.sourceforge.net/">http://jena.sourceforge.net/</a>	HP	rules	SPARQL	open-source	no	no
	Mulgara	<a href="http://www.mulgara.org/">http://www.mulgara.org/</a>	private	rules	SPARQL, TQL	open-source	no	no
	Sesame	<a href="http://www.openrdf.org">http://www.openrdf.org</a>	Aduna	RDFS / R-entailment	SPARQL, SeRQL, RDQL	open-source	no	no
	Talis	<a href="http://www.talis.com/platform/index.shtml">http://www.talis.com/platform/index.shtml</a>	Talis	RDF	SPARQL	commercial	yes	yes
	Virtuoso	<a href="http://www.openlinksw.com/virtuoso/">http://www.openlinksw.com/virtuoso/</a>	OpenLink	RDFS++	SPARQL	<i>both</i>	yes	yes
	AllegroGraph	<a href="http://agraph.franz.com/">http://agraph.franz.com/</a>	Franz Inc.	RDFS++	SPARQL	commercial	yes	no
	YARS	<a href="http://sw.deri.org/2004/06/yars/">http://sw.deri.org/2004/06/yars/</a>	DERI	RDF	SPARQL	open-source	yes	no
<b>Semantic Repositories</b>	Oracle 11g	<a href="http://www.oracle.com/technology/products/database/oracle11g/index.html">http://www.oracle.com/technology/products/database/oracle11g/index.html</a>	Oracle	OWL-Horst	custom SQL	commercial	yes	no
	DAML DB/ASIO	<a href="http://www.bbn.com/technology/data_indexing_and_mining/asio_parliament/">http://www.bbn.com/technology/data_indexing_and_mining/asio_parliament/</a>	BBN	OWL-Horst / part. SWRL	SPARQL (Jena, Sesame)			
	SWI-Prolog	<a href="http://www.swi-prolog.org/packages/Triple20/">http://www.swi-prolog.org/packages/Triple20/</a>	Univ. of Amsterdam	RDFS++	SPARQL	open-source	yes	no
	OWLIM	<a href="http://www.ontotext.com/owlim/">http://www.ontotext.com/owlim/</a>	Ontotext	OWL-Horst / R-Entailment	SPARQL (Sesame)	commercial	yes	no
<b>Reasoners</b>	Fact++	<a href="http://owl.man.ac.uk/factplusplus/">http://owl.man.ac.uk/factplusplus/</a>	Univ. Oxford	OWL-DL	DIG	open-source	no	no
	Ontobroker	<a href="http://www.ontoprise.de/">http://www.ontoprise.de/</a>	Ontoprise	OWL-DL, F-Logic	SPARQL	commercial	no	no
	OpenCyc	<a href="http://www.opencyc.org/">http://www.opencyc.org/</a>	Cycorp	OWL, CycL	SPARQL	Open-source	no	no
	Pellet	<a href="http://pellet.owlld.com/">http://pellet.owlld.com/</a>	Clark & Parsia	OWL-DL	SPARQL	open-source	no	no
	RacerPro	<a href="http://www.racer-systems.com/">http://www.racer-systems.com/</a>	Racer Systems	OWL-DL	SPARQL	open-source	no	no



**Table 4 Reasoning products and services : RDF indexers**

	<b>product</b>	<b>Link</b>	<b>owner</b>	<b>semantics</b>	<b>queries</b>	<b>full-text</b>	<b>size</b>
<b>RDF indexers</b>	Falcon-S	<a href="http://iws.seu.edu.cn/services/falcons/">http://iws.seu.edu.cn/services/falcons/</a>	Southeast University, China	RDF	entities	yes	12M sources, 600M triples
	Swoogle	<a href="http://swoogle.umbc.edu/">http://swoogle.umbc.edu/</a>	Univ. of Maryland	RDF	entities	yes	3M sources, 700M triples
	Sindice	<a href="http://sindice.com">http://sindice.com</a>	DERI	RDF	triple pattern	yes	38M sources, 600M triples
	SWSE	<a href="http://swse.deri.org/">http://swse.deri.org/</a>	DERI	RDF	SPARQL	yes	400K sources
	Watson	<a href="http://watson.kmi.open.ac.uk/">http://watson.kmi.open.ac.uk/</a>	Open University	RDFS	SPARQL	yes	??



## **HPC and Distributed Computing products and services**

Here we analyse current high performance computing and distributed computing solutions that may be of interest to support the implementation and deployment of the LarKC platform.

High performance computing Products and Services have to be analysed from several view-points. The first part of this analysis concerns the current trend in infrastructure and hardware driven both by Flop/s<sup>2</sup> per Watt considerations and by the availability of computing systems with sufficient capacity to consider replacing experiments performed in the production design lifecycle with computational engineering.

A second viewpoint for analysis is the availability of applications within the various layers on the so called “performance pyramid” introduced in the context of the PRACE (Partnership for Advanced Computing in Europe) project [26].

A third aspect of the analysis concerns the availability of toolkits and software packages for the realisation and provision of HPC based services such as Programming Models and Grid or Cloud computing software.

### **Infrastructure Viewpoint**

Parts of the LarKC platform will have a reasonably low level of coupling making them potentially executable across a distributed infrastructure, whereas other parts can only be efficiently operated in a tightly coupled environment with large storage capacities and high bandwidth connections such as that realised in compute clusters.

While several BOINC [18] based projects such as Seti@Home or Folding@Home have proved successfully that a widely distributed set of compute nodes can perform large computational tasks, their approach is not the most cost effective solution from a global viewpoint. The benefit of distribution originates largely from the fact that the costs for the computation (e.g. power costs) are donated by the owner of the PC system running the BOINC clients. Additionally the costs for the data transfer for the clients are not considered. Modern tightly integrated computing systems, in particular if they use new chip architectures such as the IBM Cell Processor used for the current fastest Supercomputer in the world, the IBM roadrunner system, have a much better ratio of Flop/s per Watt. As power costs (and the more or less linearly related cooling costs) are the major cost element the solutions propagated by BOINC or similar projects are only efficient if the compute time is donated. That this might not be possible in the long run can be seen by trends toward using BOINC for non community services such as Eternity2 or the consideration of Sony to provide the PS3 users running a folding client on BOINC basis with benefits for their online shops.

The market observation is that one can see a re-consolidation of server infrastructure in big computing centres based on the Green IT consideration of energy efficient computing. This applies not only for the HPC domain but to the server market at large.

The next table shows a classification of HPC and Distributed Computing solutions from the infrastructure viewpoint:

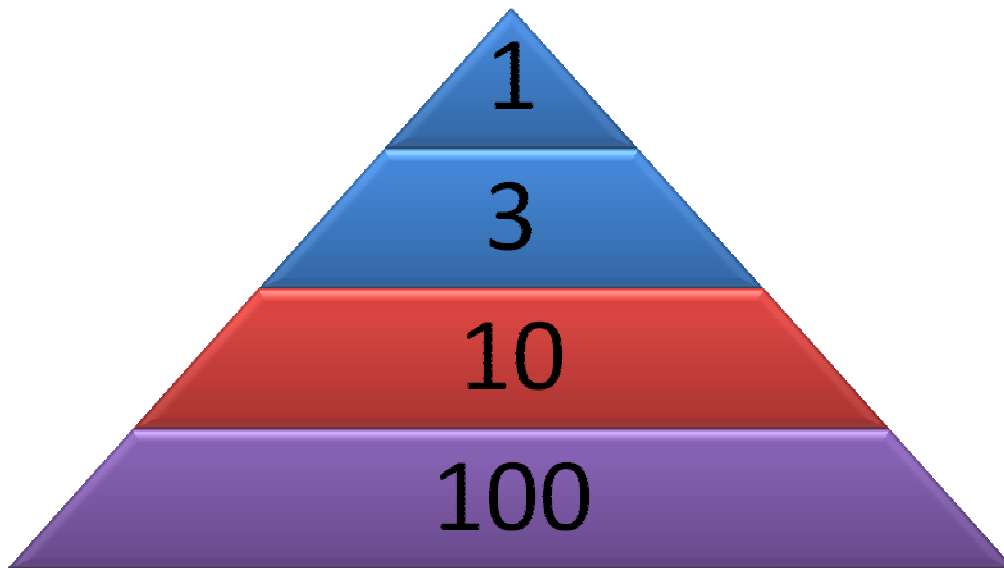
---

<sup>2</sup> Flop/s stands for Floating point operations per second

**Table 5 HPC and Distributed Computing Solutions – Infrastructure Viewpoint**

Solution	CPU types	Interconnect Bandwidth	Cost/Energy Efficiency
Highly Distributed (such as BOINC)	Variety of CPU Generation and types	Internet Speed	Low
Centralized but low to medium performance range (such as Amazon Elastic Computing)	Standard scalar server processors	10 – 100 Mbit/s	Average
Centralized but high performance	From standard scalar server processors up to highly specialised processors	$\geq 10$ Gbit/s	Up to 350 MFlops/W
Centralized, high performance hybrid computing system	Tight coupling of different processor types from Scalar, to Accelerators (e.g. Cell), and up to highly specialised processors	1 Gbit/s – 10 Gbit/s	Special Applications utilizing the characteristics of the different CPUs have decreased CPU time and Round Trip times and an even better energy efficiency

A result from the table above is that centralized solutions have advantages from an energy efficiency viewpoint. However the most efficient systems are based on specialised processor types. This means that the development of application exploiting the features of the special processors and their fast interconnecting network requires time and specific skills from the programmer. The problem that novel architectures and high end computing systems cannot easily be exploited by a large number of users is not limited to the application domain of LarKC but is generally true.



**Figure 2 Performance Pyramid for German HPC**



In the figure above the performance pyramid model introduced in the ESFRI Roadmap [28] and within the PRACE project [26] is shown with numbers for Germany. In Germany there are currently 3 national computing centres offering high end computing systems of different architectures for advanced users and specialised applications at the frontier of HPC research. HLRS is one of these 3 centres on tier-1. On the layer below (tier-2) around 10 regional computing centres provide typically general purpose server systems for a broader demand for experienced users that as of now cannot make effective use of high end computing systems. The lowest layer in this pyramid is built from off-the-shelf server systems typically offered at universities for all kind of compute jobs. This model is supposed to be extended with a European wide Computing System likely to be offered by bi-yearly assigning the task to successive national computing centres in EU member countries. While from this viewpoint it may seem that the relevance of high end computing systems to LarKC is low it must be noted that systems with equivalent capabilities to those of current tier-0/1 systems have reached tier-2 within 3-4 years and tier-3 some years later. In order to have a competitive advantage, an investment in exploiting new processor architectures by developing appropriate algorithms and applications cannot be started when they are already mainstream (tier-3). As of now (late 2008) the currently most significant change in the computing architectures are the move to increasingly multi-core processors, the utilization of the specialised processors in Graphic Processing Units (GPUs) for calculation and the application specialised processors such as the Cell processor. It is not clear which of these several trends, and which aspects of the architecture of high end computing systems will mature into tier-2 and tier-3 in the next few years. The investment solely in a single architecture therefore constitutes a significant risk.

**Table 6 HPC Systems classification**

Tier	Characteristic	Availability of ISV <sup>3</sup> Codes	Industrial Usage
0	Specialised highly efficient computing system for addressing grand challenges	Very Low	Very small number of applications (e.g. detection of oil reservoirs)
1	Special efficient computing systems for high end users and development platform for setting the basis for advancing applications to tier-0	Very Low	Industrial branches with high computing needs for specific problems (e.g. fluid dynamics simulation for aeroplane or car design, molecular dynamics simulation for drug design).
2	Compute intensive common tasks and development of new applications in preparation for tier-1	Good	Integrated in common workflows e.g. as part of the design process for gaining competitive advantage e.g. via shorter time to market
3	General purpose server system	High	Common use.

<sup>3</sup> Independent Software Vendor





## Software Viewpoint

Various approaches for achieving types of parallelisation are considered here. For tightly coupled parallel computations, the use of parallel programming models (as described below) developed within the HPC research community is generally the right solution for the development of the algorithms. Approaches for more loosely coupled systems such as Grid Toolkits, P2P solutions are relevant as well.

Three major parallel programming models within HPC can be distinguished:

- HPF (High Performance Fortran) [8]
- OpenMP (Open Multi-Processing) [9]
- MPI (Message Passing Interface) [10]

In the following table we have characterised each of the programming models by some common properties and whether they are implicit to the model, they must be done manually (by the user) or they are not necessary/not applicable:

- Distributing work to processors: in this case work decomposition is based on loop decomposition
- Distributing data to the memory: this case is applicable if the memory is distributed
- Synchronization
- Communication

**Table 7 HPC : Characterization of parallel programming models**

	HPF	OpenMP	MPI
Work Distribution	Implicit	Specified by user	Specified by user
Data distribution	Specified by user	NA	Specified by user
Synchronization	Implicit (by compiler)	Implicit	Implicit
Communication	Implicit (by compiler)	NA	Specified by user

The next table shows a deeper analysis of the three identified programming models, according to [7]:

**Table 8 HPC : Comparison of parallel programming models**

	HPF	OpenMP	MPI
Maturity of programming model	+	++	++
Maturity of standardisation	+	+	++
Migration of serial programs	0	++	--
Ease of programming	+	++	-
Correctness of parallelization	++	-	--
Portability to any HW architecture	++	-	++
Availability of implementations of the standard	+	+	++
Availability of parallel libraries	0	0	0



	HPF	OpenMP	MPI
Scalability to hundreds / thousands of processors	0	--	++
Efficiency	0	-	++
Flexibility – dynamic program structures	-	-	++
Flexibility – irregular grids, triangles, tetrahedrons, load balancing, redistribution	-	-	++

Following analysis of the previous table, we can conclude that MPI has the greatest number of advantages, except with respect to programming the algorithms for their parallel execution, as everything must be prepared and specified by the developer.

With regards to more loosely coupled systems, there are many different types of **distributed computing** approaches/architectures, as described in [6], among them:

- client-server architecture
- peer to peer (P2P) /
- thinking@home (SETI@home like approach)
- Cloud Computing
- SOA

From the distributed computing systems viewpoint, different kinds of distribution solutions/middleware are characterized in the following table, with regard to their degree of centralization, programming language and platform dependency.

Regarding the degree of centralization, the boundaries are not clear and there are a number of factors that can determine degree of centralization of a system. Broadly speaking, there are three main areas that determine whether a system is centralized or decentralized<sup>4</sup> [11]:

- Resource Discovery
- Resource Availability
- Resource Communication

---

<sup>4</sup> When we say a resource is centralized, we do not mean to imply that there is only one server serving the information, rather, we mean that there are a fixed number of servers (possibly one) providing the information, and that this provision does not scale proportionately with the size of the network [11].



**Table 9 Distributed computing products analysis**

	<b>Resource Discovery</b>	<b>Resource Availability</b>	<b>Resource Communication</b>	<b>Programming Language</b>	<b>System Platform (MSWindows, Linux,...)</b>	<b>Distributed Architecture</b>
<b>BOINC</b>	Centralized	Centralized	Centralized	API available in: C, C++. For Java clients, wrapper available	Independent	Thinking@home like
<b>JXTA</b>	Decentralized	Decentralized	Decentralized	Independent	Independent	P2P
<b>WCF</b>	Decentralized/Centralized	Decentralized/Centralized	Decentralized/Centralized	All .NET languages	MSWindows	Client-Server/SOA/P2P
<b>WSRF.NET</b>	Centralized	Half-Centralized	Decentralized	C#	MSWindows	SOA (Web Services implementation)
<b>GT4</b>	Centralized	Half-Centralized	Decentralized	Java, C	Linux	SOA (Web Services implementation)
<b>Unicore</b>	Centralized	Half-Centralized	Decentralized	Java	Independent (runs on Linux, Mac or Windows)	SOA (Web Services implementation)
<b>gLite</b>	Centralized	Half-Centralized	Decentralized	C, Java	Scientific Linux	SOA (Web Services implementation)
<b>IBIS</b>	Decentralized	Centralized	Decentralized	Java	Independent	P2P
<b>Amazon Elastic Compute Cloud (EC2)</b>	Centralized	Centralized	Centralized	Independent (API provided in different languages)	Independent	Cloud computing
<b>Google App Engine</b>	Centralized	Centralized	Centralized	Python. Other being considered for	Independent	Cloud computing



				future releases			
<b>IBM's Cloud</b>	<b>Blue</b>	Centralized	Centralized	Centralized	na <sup>5</sup>	na <sup>6</sup>	Cloud computing

---

<sup>5</sup> No information available

<sup>6</sup> No information available



- BOINC [18]: Users get both the code and the data from a centralized server, when they are available to process. Therefore, the discovery is centralized (DNS) and the communication is centralized to the central server. Resource availability is also centralized because without the availability of the server, the many BOINC nodes cannot do anything since they need to access this server to download the next chunk of data [11]
- JXTA: Project JXTA [12] defines a set of protocols that can be used to construct peer-to-peer systems using any of the centralized, brokered and decentralized approaches but its main aim is to facilitate the creation of decentralized systems. Jxta peers can be located in a decentralized fashion; they have much redundancy in their availability and their communication is point to point and therefore no central control authority is needed for their operation [11]. Programming language and platform independence are achieved through the use of the Jxta protocols (not standardized) represented in a textual representation (e.g. XML). There are implementations of the protocols written in Java, C, Perl, and others
- WCF [22][29] unifies the various communications programming models supported in .NET 2.0, into a single model. Released in November 2005, .NET 2.0 provided separate APIs for SOAP-based communications (Web Services), binary-optimized communications between applications running on Windows machines (.NET Remoting), transactional communications (Distributed Transactions), and asynchronous communications (Message Queues). WCF unifies the capabilities from these mechanisms into a single, common, general Service-oriented programming model for communications. Furthermore, WCF offers a special binding to develop P2P services, based on PNRP (Peer Name Resolution Protocol).
- In WSRF.NET [20], GT4 [19], gLite and UNICORE [21] availability is centralized in the sense that a central registry is needed to look for the available services. However, once a client discovers a service, the client and service can continue communication without the availability of the service registry. Therefore, we say that the availability is Half-Centralized, as the availability is better than a strict centralized system.
- IBIS [13] is an abstraction layer on top of grids and computing clusters, providing a homogeneous interface for resource management, job deployment, execution and management, with various parallel programming models, and an efficient synchronous and asynchronous communication layer. Resource availability is centralized, having one to many servers that manage pools of computing resources; resource discovery is then performed on one of these servers. After discovery, nodes can communicate with each other directly, using point-to-point communication. IBIS runs on top of grid middleware such as Globus, but also on self-organising peer-to-peer networks.
- The Cloud Computing solutions are considered fully centralized, as the resources are centralized in the cloud.



## 5.2. SWOT analysis

The products analysed in the previous sections can be further analyzed by a **SWOT analysis** [2], as showed in the following sections:

### Information Retrieval products SWOT analysis

Table 10 Information Retrieval products SWOT analysis

Info Retrieval - Product types		
Query pre-processing	<b>Strengths</b>	<b>Weaknesses</b>
	Increase the recall by refining the query with stemming techniques, and by expanding it with synonymies, meronymies, hyperonym (or any other type of relation between terms).  Increase the precision by disambiguating the terms in the query and returning only the results associated to the specific meaning.	Need to be designed for a specific language (English, Italian, French, ...) or vocabulary.  Need to have previous knowledge about the context domain (e.g. which are the terms searchable and their relations, and which meanings are associated with the terms).
	<b>Opportunities</b>	<b>Threats</b>
	Deployable in those domains with high intrinsic complexity (e.g. healthcare).  Provide support to users that aren't expert of the domain in which they are searching.	May require a large effort to take into consideration language specific characteristics or to be applied in un-structured domains.
Specific Media Focus	<b>Strengths</b>	<b>Weaknesses</b>
	Enable to use techniques to extract characteristics and contents embedded in media files (e.g. text in a scanned page, colors and forms in a picture, some objects in a movie).	Difficult to reach a good level of precision in the content extracted from the media files.
	<b>Opportunities</b>	<b>Threats</b>
	Enable user in searching for content embedded in media files that is otherwise inaccessible or should require a manual extraction.	Cannot be applied in those market swhere it's too difficult to extract useful characteristics from media files (e.g. extract lyrics or pentagrams from a song).



<b>Core algorithm improvement</b>	<b>Strengths</b>	<b>Weaknesses</b>
	Result personalization based on the user profile (e.g. considering query history, order refinement using drag & drop).  Improve ranking algorithm based on collaborative filtering and communities feedbacks.	Semantic-based indexer requires to annotate correctly the content to be searched by the users.
	<b>Opportunities</b>	<b>Threats</b>
	Can improve the precision and recall in un-structured domains.  Can exploit human assistance at indexing or searching time (e.g. chacha.com).	Content annotation may require manual operation that increases the cost and time.
<b>Post-processing and results visualization</b>	<b>Strengths</b>	<b>Weaknesses</b>
	Innovative and fascinating graphic interface.  Organize results in clusters based on their analogies.  Enable new interaction and usability patterns to refine the query and to navigate across results.	Server and client overhead in dealing with complex graphic.  Difficult to have high precision in clusterizing results.
	<b>Opportunities</b>	<b>Threats</b>
	Shorten the time to inspect the results, since the relevant results appear more evident.  The graphic interface attracts users.	Too many impressive graphic displays may confuse and distract users.  Excessively low quality in clustering results may discourage users.



**Reasoning products SWOT analysis**

**Table 11 Reasoning products SWOT analysis**

Reasoning - Product types		
<p><b>RDF Stores</b></p>	<p><b>Strengths</b></p>	<p><b>Weaknesses</b></p>
	<p>Efficient management of sparse and heterogeneously structured data wrt to multiple and evolving schemata.</p> <p>Based on standard data definition, schema, and query languages (this is not the case with several other alternatives of the RDBMS, e.g. the so-called column stores)</p>	<p>Less mature then RDBMS and other DBMS. Follows a list of immature or missing functionalities: transaction support; data modification language; data aggregation functions</p> <p>As any new technology, RDF stores and frameworks are still not very well integrated within existing IT environments (e.g. programming languages, application servers, web UI frameworks).</p> <p>Usually cannot match the performance of RDBMS on datasets with stable schemata and low sparsity.</p>
	<p><b>Opportunities</b></p>	<p><b>Threats</b></p>
	<p>To become a standard solution for integration and federation of structured data from multiple sources. Apart from RDBMSs, RDF is more or less the only standard approach for this. RDF stores have the potential to lower the cost of data integration, however they still need to match the maturity of the RDBMS. (In fact, RDF repositories are already used for this purposes by the life science community)</p> <p>If the vision for the Semantic Web as development of the so-called “linked data”, gains speed, RDF repositories will be as important for Web 3.0 as the HTTP servers for the original WWW.</p>	<p>Negative brand image if the Semantic Web goes “out of fashion” in the industry (as happened to AI in the 90s).</p> <p>To suffer from sub-optimal development of the RDF specification and other related standards. Inappropriate extensions of the RDFS standards can have negative implications on the acceptance of the RDF stores. Conservation of the standards can also cause trouble (e.g. if extensions like the “named graphs” did not get standardized soon)</p> <p>The market can be overtaken by RDBMS plug-ins (in the same</p>





	To be integrated within existing systems (e.g. CMS, CRM, and BI).	way in which the major RDBMS released some XML support and native XML databases have never gained substantial market).
<b>Reasoners</b>	<b>Strengths</b>	<b>Weaknesses</b>
	<p>Reasoners allow for operational usage of declarative semantics. Generally, this allows for more efficient encoding and usage of the abstract model of the world that each piece of software works with. Currently, this model is either not formally specified or bits and pieces of its semantics are split and scattered into various schemata (e.g. DB and OO).</p> <p>Formal declarative semantics, specified at a single place and interpreted by an inference engine, allow for development of both cheaper and more intelligent programs.</p>	<p>The computational complexity/cost of the reasoning even with respect to relatively simple logical fragments results in performance and scalability which are far below the minimal requirements for many applications.</p> <p>Today's approach for KR (ontology modelling) have proven to be barriers to popularisation in many environments. Many engineers and architects, who can develop relational schema and manage RDBMS, find ontology development too complex, cannot debug/tune their ontology and cannot develop applications based on it.</p>
	<b>Opportunities</b>	<b>Threats</b>
	<p>To become an important part of the next generations of many information systems and applications, providing clear advantages in terms of functionality (e.g. more intelligent systems) and lower total cost of ownership.</p> <p>The opportunities for the reasoners are not related to replacement of the DBMS in existing environments, but rather with the discovery of appropriate usage patterns and positioning as a new component.</p> <p>Reasoners can be used to improve or make cheaper and more advanced various components and activities: UI (human-machine interaction), validation of data schemata, validation and optimisation of configurations and solutions,</p>	<p>To fail to bridge the gap to the wide IT audience, due to the perception of unmanageable complexity.</p> <p>Finding use cases, practices, and approaches which allow beneficial, cost efficient and manageable usage of reasoning is crucial. Unless such cases and practices are found over the next couple of years, it is likely that reasoning will lack the industrial interest associated with the expectations currently building around the Semantic Web.</p> <p>Formal symbolic reasoning (based on mathematical logic) fights for attention with a wide range of statistical methods which can deliver a different type of</p>



	decision making, etc.	intelligence to the applications. Unless useful synergies are found, symbolic reasoning can be marginalized (wrt real-world applications).
<b>Semantic Repositories</b>	<b>Strengths</b>	<b>Weaknesses</b>
	<p>Those derive the RDF stores. Further, the light-weight reasoning supported by these engines allows:</p> <ul style="list-style-type: none"> <li>- efficient support for interlinking data encoded WRT different schemata;</li> <li>- query answering based on automated data interpretation;</li> <li>- more flexible and intelligent mapping of information needs (queries) to the formalization of the data.</li> </ul> <p>For example, a semantic repository can return John, on query “select X where X relativeOf Marry”, based on assertion “Marry motherOf John” and a ontology where relativeOf is a symmetric relation, more general than motherOf.</p> <p>Semantic repositories allow even better cost efficiency, lowering further the cost of data integration and retrieval. Semantics that is otherwise hard-coded at multiple places in the applications or injected in the queries (thus adding unnecessary complexity) can be encoded at a single place (in the ontology) and handled by the repository.</p>	<p>Even light-weight inference can have negative impact on performance. For instance, backward-chaining repositories can suffer poor query evaluation performance. On the other hand forward-chaining based machines, need to pay specific attention to loading performance and transaction handling.</p>
	<b>Opportunities</b>	<b>Threats</b>
	<p>Same as the opportunities for the RDF stores.</p> <p>Semantic repositories are well suited for application in OLAP and BI systems.</p>	<p>Same as the threats for the RDF stores.</p> <p>Semantic repositories should allow manageable and predictable reasoning behaviour. Otherwise they could fail to cover the</p>



	There are further opportunities for applications where automated validation (consistency checking) of large volumes of data is required.	expectations for DBMS-level scale, efficiency and reliability.
<b>RDF Indexers</b>	<b>Strengths</b>	<b>Weaknesses</b>
	<p>Allow for pre-selection of relevant RDF data from large number of sources.</p> <p>RDF indexers usually provide hybrid querying functionality, combining relevance ranking, based on similarity (as in IR), and some structured constraints (as in RDBMS).</p> <p>RDF indexers can add value by using semantic repositories to interlink data from different RDF graphs (as Sindice does).</p>	There are still no proven relevance ranking approaches. This is a serious weakness, because (i) relevance is crucial for the usability and the acceptance of such engines and (ii) developing the acceptable ranking schemata can take considerable time.
	<b>Opportunities</b>	<b>Threats</b>
	The major opportunity for RDF indexers is to become the search engines of Web 3.0.	Considering the scenario of a public RDF, search engine, the major risk for the development of such system is that there is still not enough RDF data and enough users interested in it. The lack of critical mass of both users and data can make maturing such engines an impossible task.

**HPC and Distributed Computing products SWOT analysis**

The objective of this section is to analyse the possibilities to exploit existing trends and developments on HPC infrastructure level (as described above) and distributed computing solutions for the LarKC platform, rather than analysing the concrete HPC and distributed computing products and services described in previous sections. This is motivated by the fact that a SWOT analysis of concrete individual products in this field would not be meaningful. The analysis can be found in the following table:



**Table 12 HPC and Distributed Computing products SWOT analysis**

	Supporting the realization of the objective	Potential risks for the realization of the objective
<b>Strength and weaknesses of proposed approach</b>	<p><b>Strengths</b></p> <p>Within LarKC experts from semantic reasoning, Grid and HPC are closely working together</p> <p>A wide range of different platforms including resources ranked in the top500 list are available</p> <p>Modular concept of the platform enabling distributed and coupled parallelism</p> <p>Distributed solution addresses legal constraints that only derived data might be shared externally</p>	<p><b>Weaknesses</b></p> <p>Existing programming models and environments are targeted for Fortran, C &amp; C++ based codes</p> <p>Distribution solutions such as BOINC have not been designed for low carbon footprint</p> <p>Security models are quite different for the different distributed computing toolkits</p>
<b>External developments supporting or preventing success</b>	<p><b>Opportunities</b></p> <p>Parallel Programming is developing towards a mainstream technology leading to better programming environments</p> <p>The dominating role of cluster system for the mid performance tier make a availability of the technological baseline for LarKC likely</p> <p>The traditional players using HPC technology face increasingly the problem of very large amounts of data that needs to be analysed</p>	<p><b>Threats</b></p> <p>An optimized solution for a specific high end computing system is likely to be obsolete in short time frame due to highly dynamic development of technology at the moment</p> <p>Legal issues with analysis of data</p> <p>The costs associated with using HPC computing facilities is quite high and while the market for semantically analysed data is existing there are cost and legal constraints</p>

## 6. Target standardisation bodies

This chapter gives an overview of the target standardization bodies and groups of interest for the LarKC project. As the project progresses, the groups to be followed more closely and potentially influenced by the project will be identified. The list of groups of interest may be also enlarged if it is considered necessary for the project's interests.

Before describing these standardisation groups in detail, we summarise their relation to the LarKC project. The table below shows the relation between the most relevant standardisation bodies and the main technology areas within the LarKC project.



**Table 13 Standardisation matrix**

<b>Stand. body</b>	<b>OASIS</b>	<b>W3C</b>	<b>MPI</b>	<b>OGF</b>
<b>Tech. topic</b>				
data formats and language profiles		X		
service description and invocation	X	X		
distributed service architectures	X		X	X
resource description and invocation	X			X
parallel programming models			X	

In the following sections, we describe the relevant standardisation bodies in more detail. For each group, information is provided on the following aspects:

- Description: Brief description of the group goal
- Relevance for LarKC: Areas of work of interest for LarKC
- Current status: Current status of the group, current trends under discussion,...
- Monitoring: How can the group be monitored by LarKC partners (e.g. possibilities for attendance of periodic f2f meetings, following mailing list,...)
- Participation and contributions: Procedure within the group for active participation and contributions (e.g. submission of draft by email, formal contribution in f2f meeting, procedures for discussions and approval,...)
- LarKC partners involvement: names of the LarKC partners involved in the group, degree of involvement,...

It is important to note that the standardisation bodies are listed in alphabetical order, which is not necessarily the same as the order of relevance to the project.

### **6.1. MPI Forum**

#### **Description**

The MPI (Message Passing Interface) [33] forum is an open group with representatives from various organisations, that together define the MPI standard for inter-process communication, focusing primarily on the message-passing parallel programming model.



## Relevance for LarKC

MPI is relevant to LarKC as one of the parallel programming models to be consider inside the plug-ins and maybe for communication between the plug-ins in certain cases of parallel execution (to be analysed). Available implementations of MPI in Java programming language will be considered for LarKC.

## Current status

The latest version of MPI, MPI-2.1, was approved by the MPI Forum on September 4, 2008 with the second and final official vote.

Currently discussions are already taking place on:

- MPI-2.2: Small changes to the MPI-2.1 standard. A small change is defined as one that does not break existing user code, either by interface changes or by semantic changes, and does not require large implementation changes.
- MPI-3.0: Additions to the MPI 2.2 standard that are needed for better platform and application support. These are to be consistent with MPI being a library that provides parallel process management and data exchange capabilities. This includes, but is not limited to, issues associated with scalability (performance and robustness), multi-core support, cluster support, and applications support.

## Monitoring

There is a public MPI Forum Working Groups Wiki [37] where discussions on standard modifications, future versions, etc. take place. The wiki is publicly readable but one need to be subscribed in order to write new entries. There are also mailing lists organized by standard versions and groups, publicly readable and also with participation subject to subscription.

## Participation and contributions

The MPI Forum organizes face to face meetings approximately every two months, where every interested person can attend, normally after the payment of a fee to cover the meeting costs.

Contributions can be made both through wiki/ mailing lists (for registered users) and in the f2f meetings.

But in order to have the right to vote, a participant must attend at least every second f2f meeting. If somebody is interested in making a relevant contribution, it is highly recommended that he/she defend it in a f2f meeting.

## LarKC partners involvement

LarKC partner HLRS is involved in the MPI standardization committee.

## 6.2. OASIS

### Description

The OASIS group (*organisation for the advancement of structured information standards*) [31] is a non-profit consortium developing open information standards. OASIS was initially founded to promote



interoperability between SGML (predecessor of XML) vendors and users, and expanded its scope towards general interoperability. OASIS is active in technology areas such as Web services, e-commerce, XML processing, supply-chain management, and trust and security management. OASIS has some 5,000 participants and some 600 members (organisations and individuals).

### **Relevance for LarKC**

Several OASIS working groups are relevant to LarKC:

- *Unstructured Information Management Architecture (UIMA)*: focusing on content analysis, automatic semantic annotation, and semantic exploration of unstructured information. The work in this group is directly relevant to LarKC, since several LarKC use-cases require unstructured information to be analysed and extracted into structures with semantic annotation.
- *Semantic Execution Environments*: its mission is to standardise the reference ontology and reference architecture for Semantic Execution Environments of Semantic Web Services. The committee also develops guidelines and implementation directions for deploying semantic Web Services into service-oriented architectures. As LarKC can be seen as such an "execution environment", and as LarKC plug-ins could be seen as "semantic Web services", the work in this group is clearly relevant.

The relevance of other OASIS groups for the LarKC project will be further analysed as the project progresses.

### **Current status**

The technical work of OASIS is driven by its members; technical committees (TCs) are formed based on the proposals of the OASIS members, and the TCs set their own agendas and schedules. OASIS provides the guidance, process, and infrastructure necessary for its members to do the work.

Currently there are TCs in a number of areas including the following:

- Horizontal and e-business framework
- Web Services
- Security
- Public Sector
- Vertical industry applications

### **Monitoring**

Governance of OASIS is transparent and open. OASIS offers a possibility of membership. Becoming a member provides more leverage to influence the standardisation process. The membership is open, any individual or an organisation company is eligible that is somehow involved or can benefit from the standardisation process. The direction of OASIS Consortium is determined by the Board of Directors and Technical Advisory Board. Members of these strategic boards are elected by an open ballot and they serve two year terms.

The work of the OASIS is open to a global community. Information on the status of standards is regularly published in six languages. The archives of discussion groups and technical committee documents are open to public. This enables to monitor the activities and provides an opportunity for wide collaboration.



## Participation and contributions

There are various ways to contribute to the OASIS effort and membership is not necessarily required. After obtaining a login to a Focus Area website the communication can take various forms, for example:

- editing wiki pages discussion topics related to the use and understanding of standards
- joining the discussion group
- making comments to standard proposals

The Focus Areas contain information on Products, Services, Forums, Blogs and News. This information can be directly edited by users.

The work on individual standards takes place within Technical Committees. There are various rules stipulating the participation, licensing, disclosure, transparency and accountability aspects of this work. Before getting involved in this work the participants are encouraged to apply for membership. There are various levels of membership that reflect the level of involvement in the standardisation process.

### LarKC partners involvement

LarKC partners participate in two OASIS working groups:

- The University of Sheffield participates (and is a founding member) in the technical committee on *Unstructured Information Management Architecture (UIMA)*.
- STI Innsbruck participates (and is a founding member) of the technical committee on *Semantic Execution Environments*.

## 6.3. Open Grid Forum

### Description

The Open Grid Forum (OGF) [34] is a community to drive the evolution and adoption of distributed computing. OGF shares best practices and consolidates these into standards to easy deployment of distributed computing techniques.

The Open Grid Forum (OGF) was formed from the merger of the Global Grid Forum (GGF) and the Enterprise Grid Alliance (EGA). GGF had a rich history and established international presence within the academic and research communities along with a growing participation from industry. EGA was a consortium focused on developing and promoting enterprise grid solutions. OGF provides an open forum that brings together key individuals and organizations from the grid community to align requirements; identify and remove barriers; workshop best practices that will expedite grid adoption. As an open standards organization, OGF collaborates extensively with other standards development organizations to align with existing industry standards and develop new specifications to enable grid software interoperability.

### Relevance for LarKC

The initial identification of relevant working groups for LarKC within the OGF refers to the Grid Resource Allocation Agreement Protocol Working Group (GRAAP-WG) [40], the Semantic Grid Research Group (SEM-RG) [39], the Open Grid Services Architecture Working Group (OGSA-WG) [44] and the Reference Model Working Group (RM-WG) [45].

The goal of the GRAAP-WG is to produce a set of specifications and supporting documents which describe methods and means to establish Service Level Agreements between different entities in a distributed





environment. This group is relevant for LarKC for the SLAs to be negotiated between the LarKC platform and the different plug-ins, in order to guarantee the required level of QoS.

The goal of the SEM-RG is to realise the added value of emerging Web technologies and approaches, in particular Semantic Web and Web 2.0, for Grid users and developers. This group is relevant to LarKC for the semantic description of distributed resources, QoS parameters, etc.

The Open Grid Forum (OGF) has embraced the OGSA as the blueprint for standards-based grid computing. 'Open' refers to the process used to develop standards that achieve interoperability. 'Grid' is concerned with the integration, virtualization, and management of services and resources in a distributed, heterogeneous environment. It is 'service-oriented' because it delivers functionality as loosely coupled, interacting services aligned with industry-accepted Web service standards. The 'architecture' defines the components, their organizations and interactions, and the design philosophy used. OGSA-WG is developing the architecture and its constituent specifications and profiles in collaboration with a number of fellow working groups.

The RM-WG aims to:

- Capture all of the common and abstract components (services and resources) that comprise a Grid.
- Describe, both formally and informally, these components, together with their life-cycles and their relationships with one another.
- Pragmatically reconcile these with other extant standards, including but not necessarily limited to the other OGF standards (planned for Reference Model v2.1), the DMTF's CIM (Common Information Model), the SNIA's SMI-S etc., ITIL, DCML and so forth (planned for Reference Model v2.2).
- Define ways of representing and serializing the model, including, but not necessarily limited to RDF/RDFS/OWL Vocabulary, SML, XML Schemas, Java etc (planned for Reference Model v3.0).

### **Current status**

The focus of GRAAP is currently on the interoperability testing of different WS-Agreement implementations (to evaluate the specification and to evolve from an OGF Proposed Recommendation to an OGF Recommendation) and on use cases. In addition, the group discusses the definition of a re-negotiation protocol to be used in the context of WS-Agreement.

The SEM-RG currently engages in a diverse range of activities which includes (1) the application of Web technologies within the lifecycle of information, including scientific data and other digital artifacts of the e-Science process such as workflows and provenance; (2) application of Web technologies within the infrastructure, for example resource and service descriptions, in order to facilitate automation in discovering and combining resources; and (3) application to the social dimension of e-Science, from social networks to collective intelligence.

The OGSA-WG published in July 2008 the *Information and Data Modelling in OGSA Grids*, which provides information to the Grid community on the direction for information and data modelling of OGSA resources.

The RM-WG is currently working in the *OGF Reference Model V2.0*, which will provide information to the Grid community on a reference model for grid components and how they can be organized and managed.

### **Monitoring**

OGF organizes international events three times a year to align requirements, identify and remove barriers to grid adoption, explore related technologies, and record best practices of grid usage. During these 3-4 days sessions, OGF working groups discuss and advance draft specifications toward publication and Research Groups explore future trends and applications of grids in a variety of commercial and research communities.



New groups and communities are proposed through Birds of a Feather (BoF) sessions and tutorials offer hands-on training and experience with grid implementations and technologies.

Beside the f2f meetings, it is also possible to monitor the activities of the different groups subscribing to the corresponding mailing lists and following the updates on the wikis, for which it is not necessary to be an OGF member. The OGF membership allows participants to have a more influent participation and advantages in the meetings fees.

### **Participation and contributions**

The primary product of the OGF is the specifications and best practices that result from working group and research group activities. When a group has reached consensus on a draft document (called a Grid Working Draft or GWD), it is submitted to the OGF Editor, who manages a virtual "pipeline" of documents that move through a prescribed process toward publication. It is acceptable for an individual or set of authors not in a recognized OGF group to submit a draft to the OGF Editor. In this case, the Editor will work with the submitting parties to ensure that the document is in the proper format.

While a draft is in the pipeline, it is reviewed by the Area Directors, the OGF Editor, and eventually by the community as a whole (during the public comment process). When a draft has progressed through the prescribed process and has gained final approval from the OGF Editor and the Steering Group, it is promoted to a Grid Final Document (GFD), is given a unique number, and becomes part of the OGF Document Series.

### **LarKC partners involvement**

LarKC partner HLRS is actively involved in the GRAAP-WG working group on grid resource allocation, attending to the periodic f2f meetings and following the different monitoring channels. Furthermore, HLRS monitors the SEM-RG, the OGSA-WG and the RM-WG.

## **6.4. W3C**

### **Description**

The W3C (*World-Wide Web consortium*) [32] is an international consortium whose mission is to "lead the Web to its full potential by developing protocols that ensure long-term growth for the Web". The W3C, headed by Sir Tim Berners-Lee, the inventor of the Web, has developed many fundamental Web standards such as HTML, XML, and related languages such as XSL, and XQuery. The W3C has standardised some Web service protocols, such as SOAP and WSDL. The W3C also guides the development of Semantic Web standards, such as RDF, OWL, and SPARQL.

### **Relevance for LarKC**

There is a wide range of W3C standards applicable to LarKC and, therefore, different groups are of interest for the project, such as the *Health-Care and Life Sciences* interest group (HCLS), the *RIF working group* or the *Data Access working group*.

### **Current status**

W3C Activities are generally organized into groups: Working Groups (for technical developments), Interest Groups (for more general work), and Coordination Groups (for communication among related groups). These groups, made up of participants from Member organizations, the Team, and Invited Experts, produce the bulk of W3C's results: technical reports, including Web standards, open source software, and services



(e.g. validation services). These groups also ensure coordination with other standards bodies and technical communities.

### Monitoring

Working groups can be partially tracked through W3C mailing lists and forums that are publicly visible, which provides a channel for interacting with working groups prior to the release of a candidate recommendation.

### Participation and contributions

W3C standards are authored by W3C Working Groups. Working groups are formed from W3C members. At the time of writing, the membership fee for a non-profit organisation is €6,500 per annum. Once a working group has completed a draft that becomes available for public review in the form of a candidate, recommendation and comments can then be submitted to the group for consideration. Comments submitted in this manner and their responses are made public subject to the group director's approval.

### LarKC partners involvement

Many LarKC partners are active within groups of the W3C:

- AstraZeneca participates in the *Health-Care and Life Sciences* interest group (HCLS), which develops and advocates for the usage of Semantic Web technologies in health-care, translational medicine, and bio-sciences. AstraZeneca is particularly interested in the *Linking Open Drug Data initiative* (LODD), which aims at linking together publicly available drug information, such as drug impact on gene expressions and clinical trial results. The HCLS group and LODD initiative share a common interest with LarKC's use cases 7a and 7b on linking and reusing (publicly available) drug information.
- Both STI Innsbruck and CycEurope participate in the *RIF working group*, which aims at a common standard for interchanging (logical) rules on the Web. Rules are relevant for LarKC as a crucial form of declarative knowledge representation. Currently no W3C standard exists for rules on the Semantic Web. Some proposals have been made, but combining rules with existing Semantic Web standards (ie: OWL) while staying within reasonable computational characteristics is not trivial.
- CEFRIEL monitors the *Data Access working group* of the W3C, which develops standards to query RDF data over the Web. The working group produced the SPARQL recommendation, but may in the future propose extensions to this standard to address requirements that are not addressed by the current standard. CEFRIEL is currently evaluating proposing a SPARQL extension for querying streaming data (named C-SPARQL) to this working group in 2009. A first contribution should be ready in the spring 2009 and will be mainly focused on aggregation operators for C-SPARQL, while a second contribution concerning RDF streams as well as continuous SPARQL queries should be ready in the fall of 2009.

### 6.5. Others

#### Knowledge discovery standards

The large variety of data and model formats that researchers and practitioners have to deal with and the lack of procedural support in Knowledge discovery have prompted a number of standardization efforts in recent years, led by industry and supported by the knowledge discovery community at large [41]. *Sarabjot Singh Anand, Marko Grobelnik, Frank Herrmann, Mark Hornick, Christoph Lingenfelder, Niall Rooney, Dietrich Wettschereck, Knowledge discovery standard*, published in 3 September 2008, provides an overview of the



most prominent of these standards and highlights how they relate to each other using some example applications of these standards.

This area is of interest for LarKC since one task to use LarKC for is Knowledge Discovery. In fact, one could wrap knowledge discovery software as a LarKC plug-in (e.g. as TRANSFORM, previously known as "ABSTRACT"). Therefore, the standardisation activities in the area will be monitored by LarKC.

## 7. Conclusions

Having performed a preliminary analysis of the LarKC market, analysed the environment and context, and identified technology products and services, we come to the conclusion that, in order to make a thorough analysis of the LarKC related Market, we will need to identify the concrete exploitable items generated by the project. For this purpose, in future releases of this deliverable, the most significant and innovative LarKC results will be identified as potential candidates for exploitation. They will be the basis for performing a new and more exhaustive market analysis.

As the LarKC project is evolving in a rapidly changing world, it is necessary for the project researchers to be continuously aware of the status of the related technologies. Technologies used to build the project will be, as much as possible, based on mature and emerging standards. In order to reach this commitment an initial identification of bodies and groups of interest has been performed in this document. In subsequent deliverables, a detailed standardisation strategy will be established. This strategy will include plans for LarKC to appropriately influence standards activities with concrete results of the project.

## 8. References

- [1] [http://en.wikipedia.org/wiki/PEST\\_analysis](http://en.wikipedia.org/wiki/PEST_analysis)
- [2] [http://en.wikipedia.org/wiki/SWOT\\_analysis](http://en.wikipedia.org/wiki/SWOT_analysis)
- [3] [http://en.wikipedia.org/wiki/B.C.G.\\_Analysis](http://en.wikipedia.org/wiki/B.C.G._Analysis)
- [4] [http://en.wikipedia.org/wiki/G.\\_E.\\_multi\\_factoral\\_analysis](http://en.wikipedia.org/wiki/G._E._multi_factoral_analysis)
- [5] [http://en.wikipedia.org/wiki/Porter\\_5\\_forces\\_analysis](http://en.wikipedia.org/wiki/Porter_5_forces_analysis)
- [6] Georgina Gallizo, Sabine Roller, Axel Tenschert, Michael Witbrock, Barry Bishop, Uwe Keller, Frank van Harmelen, Gaston Tagni, Eyal Oren. *Summary of parallelisation and control approaches and their exemplary application for selected algorithms or applications*. LarKC project deliverable D5.1, [http://www.larkc.eu/wp-content/uploads/2008/10/larkc\\_d51\\_summary-of-parallelization-and-control-approaches-and-their-exemplary-application-for-selected-algorithms-or-applications.pdf](http://www.larkc.eu/wp-content/uploads/2008/10/larkc_d51_summary-of-parallelization-and-control-approaches-and-their-exemplary-application-for-selected-algorithms-or-applications.pdf)
- [7] Rolf Rabenseifner, High Performance Computing Center Stuttgart (HLRS), *Parallelization Tutorial at LarKC Kick-Off Meeting*, 15 April 2008
- [8] High Performance Fortran, <http://hpff.rice.edu/>
- [9] <http://www.openmp.org/>
- [10] <http://www.mpi-forum.org/>
- [11] Ian J. Taylor. *From P2P to Web Services and Grids – Peers in a Client/Server World*, Springer-Verlag London Limited, 2005
- [12] <https://jxta.dev.java.net/>
- [13] <http://www.cs.vu.nl/ibis/>
- [14] <http://www.unicore.eu/documentation/manuals/unicore6/>
- [15] <http://www.2enlightenment.com/node/7>
- [16] <http://www.cifs.dk/scripts/artikel.asp?id=1469>
- [17] [http://en.wikipedia.org/wiki/Treaty\\_of\\_Lisbon](http://en.wikipedia.org/wiki/Treaty_of_Lisbon)
- [18] <http://boinc.berkeley.edu/>
- [19] Globus Project. <http://www.globus.org>
- [20] WSRF.NET: <http://www.cs.virginia.edu/~gsw2c/wsrif.net.html>



- [21] <http://www.unicore.eu/>
- [22] Windows Communication Foundation (WCF): <http://msdn.microsoft.com/en-gb/library/ms731190.aspx>
- [23] Carole Goble, Robert Stevens. [State of the nation in data integration for bioinformatics](http://www.sciencedirect.com/science/article/B6WHD-4RS43MK-5/2/ab521e38479b012d688a645191bfa1c6), Journal of Biomedical Informatics Volume 41, Issue 5, Semantic Mashup of Biomedical Data, October 2008, Pages 687-693. <http://www.sciencedirect.com/science/article/B6WHD-4RS43MK-5/2/ab521e38479b012d688a645191bfa1c6>
- [24] Bo Andersson, Vassil Momtchev. [Requirements summary and data repository](http://www.larkc.eu/wp-content/uploads/2008/10/larkc_d7a-11_requirements-summary-and-data-repository.pdf). LarKC project deliverable D7a.1.1, [http://www.larkc.eu/wp-content/uploads/2008/10/larkc\\_d7a-11\\_requirements-summary-and-data-repository.pdf](http://www.larkc.eu/wp-content/uploads/2008/10/larkc_d7a-11_requirements-summary-and-data-repository.pdf)
- [25] [http://en.wikipedia.org/wiki/Ishikawa\\_diagram](http://en.wikipedia.org/wiki/Ishikawa_diagram)
- [26] <http://www.prace-project.eu/>
- [27] <http://www.eternity2.fr/download>
- [28] [ftp://ftp.cordis.europa.eu/pub//docs/esfri-roadmap-report-26092006\\_en.pdf](ftp://ftp.cordis.europa.eu/pub//docs/esfri-roadmap-report-26092006_en.pdf)
- [29] [http://en.wikipedia.org/wiki/Windows\\_Communication\\_Foundation](http://en.wikipedia.org/wiki/Windows_Communication_Foundation)
- [30] Amazon Elastic Compute Cloud (EC2), <http://aws.amazon.com/ec2/>
- [31] <http://www.oasis-open.org>
- [32] <http://www.w3.org/>
- [33] <http://www.mpi-forum.org/>
- [34] <http://www.ogf.org>
- [35] [http://cordis.europa.eu/fp7/i2010\\_en.html](http://cordis.europa.eu/fp7/i2010_en.html)
- [36] <http://www.2enlightenment.com/>
- [37] <https://svn.mpi-forum.org/trac/mpi-forum-web/wiki/>
- [38] <http://lists.mpi-forum.org/>
- [39] <http://www.semanticgrid.org/OGF>
- [40] [http://www.ogf.org/gf/group\\_info/view.php?group=graap-wg](http://www.ogf.org/gf/group_info/view.php?group=graap-wg)
- [41] <http://www.springerlink.com/content/x221p87h41k18063/>
- [42] Atanas Kiryakov. *Measurable Targets for Scalable Reasoning*. LarKC project deliverable D5.5.1, [http://www.larkc.eu/wp-content/uploads/2008/07/larkc\\_d551.pdf](http://www.larkc.eu/wp-content/uploads/2008/07/larkc_d551.pdf)
- [43] Ellen J. Stokes et al., *Information and Data Modeling in OGSA Grids*, July 2008, <http://www.ogf.org/documents/GFD.137.pdf>
- [44] OGF OGSA-WG, [http://www.ogf.org/gf/group\\_info/view.php?group=ogsa-wg](http://www.ogf.org/gf/group_info/view.php?group=ogsa-wg)
- [45] OGF RM-WG, [http://www.ogf.org/gf/group\\_info/view.php?group=rm-wg](http://www.ogf.org/gf/group_info/view.php?group=rm-wg)
- [46] Fischer, F; Keller, U; Kiryakov, A; Huang, Z; Momtchev, V; Simperl, E. (2008). *Initial Knowledge Representation Formalism*. Deliverable D1.1.3 or project LarKC. <http://www.larkc.eu/deliverables/>
- [47] Atanas Kiryakov. *Measurable Targets for Scalable Reasoning*. Deliverable D5.5.1 of project LarKC <http://www.larkc.eu/deliverables/>