HARMOSEARCH
the future of information services

FP7-SME-1
Project no. 262289

HARMOSEARCH

Harmonised Semantic Meta-Search in
Distributed Heterogeneous Databases

D5.1

Registry Requirements Analysis Report

Due date of deliverable: 2011-08-31
Actual submission date: 2011-08-31

Start date of project: 2010-12-01 Duration: 24 month

Project funded by the European Comission within the Seventh Framework Programme

Dissemination Level		
PU	Public	**X**
PP	Restricted to other participants (including the Commission Services)	
RE	Restricted to a group specified by the Consortium (including the Commission Services)	
CO	Confidential, only for members of the Consortium (including the Commission Services)	

PROJECT ACRONYM: **HARMOSEARCH**

Project Title: Harmonised Semantic Meta-Search in Distributed Heterogeneous Databases

Grant Agreement: 262289

Starting date: December 2010 **Ending date:** November 2012

Deliverable Number: D5.1

Title of the Deliverable: Registry Requirements Analysis Report

Lead Beneficiary: CPR

Task/WP related to the Deliverable: WP 5, Task 5.1

Type (Internal or Restricted or Public): Public

Author(s): Christoph Herzog, Claudio Prandoni

Partner(s) Contributing: TU-Wien, CPR

Contractual Date of Delivery to the CEC: August 31, 2011

Actual Date of Delivery to the CEC: August 31, 2011

PROJECT CO-ORDINATOR

Company name:	[X+O]
Name of representative:	Manfred Hackl
Address:	HAMBURGERSTRASSE, 10/7, A-1050 Vienna, Austria
Phone number:	+43-676-842755-500
Fax number:	+43-676-842755-599
E-mail:	manfred.hackl@xpluso.com
Project WEB site address:	www.harmosearch.org

TABLE OF CONTENTS

1 INTRODUCTION

1.1 PURPOSE OF THE DOCUMENT

In order to provide fast and reliable results even when thousands of partners can be searched, a semantic index is needed, which contains knowledge about the participants and the data they can provide. In this document we start from the requirements identified in deliverable D2.1 and conduct a detailed requirements analysis for the semantic registry, considering functional, non-functional and technical requirements. Based in these requirements we perform an evaluation of currently existing frameworks with respect to the identified requirements. Based on this evaluation, we chose the most suitable framework for building the HarmoSearch semantic registry on top of the selected base technology. Finally, we present an architectural overview of how the semantic registry will be built.

1.2 DEFINITIONS OF TERMS AND ABBREVIATIONS

Harmonise: name of the existing technological solution. The current version is Harmonise 2.0, which includes the Harmonise Ontology, the Harmonise Service Centre and the Harmonise Portal.

Metasearch: HarmoSearch component which provides distributed search capabilities to the integrated data sources.

Query Processor: HarmoSearch component which translates a query from one query language to another.

Harmonise Participant: A Harmonise participant is an actor in the Harmonise network who can provide or consume data and services.

Data Provider: A source providing data in one the subdomains of the Harmonise ontology (Events, Accommodation, etc.). Each Harmonise participant can operate several data providers.

Data Registry: The part of the semantic registry concerned with describing data providers and the content they offer and matching it against search queries.

HarmoSearch Workflow: A workflow describes the composition and configuration of functionalities into a flexibly managed process. In the HarmoSearch context this means the flexible configuration and extension of search and other data transfer processes.

Service Provider: A service provider offers an external service that can be integrated into a Harmosearch workflow. Such a service normally operates on the results of a HarmoSearch search query. Each Harmonise participant can operate several service providers.

Service Registry: The part of the semantic registry concerned with registering data about data providers and offering a search interface for accessing these data.

1.3 RELATIONSHIP WITH OTHER DOCUMENTS

Inputs to this document come from the deliverable D2.1, Use Case Specification, which defines the use cases and scenarios relevant for the system. User requirements are derived from this specification.

Furthermore, the document is based on the specification of the HarmoSearch query language (D4.1) and the ontology for the query model (D3.1). The technical requirements are derived from these deliverables.

Finally, the HarmoSearch architectural design (D2.2) builds the basis for a first overview of the architectural design of the Semantic Registry.

1.4 STRUCTURE OF THE DOCUMENT

This document is structured in the following main sections:

- The expected functionality and requirements for the semantic registry are identified in sections 2 and 3. These requirements include requirements explicated by the users in the first month of the project and more technical requirements arising from the expected interactions with other components.
- Section 4 provides an overview of relevant existing solutions which are candidates for being used as base technology for implementing the semantic registry. These are compared against each other with respect to functional and non-functional requirements derived from the detailed requirements analysis.
- Finally, section 5 presents an architecture for implementing the HarmoSearch semantic registry based on the selected base technology. It also outlines the interactions of the semantic registry with other Harmonise components with respect to the main use cases of the semantic registry.

2 FUNCTION OF THE SEMANTIC REGISTRY

2.1 PRIMARY FUNCTION: DATA REGISTRY

In order to provide fast and reliable results even when thousands of partners can be searched, a service is needed, which contains knowledge about the participants and the data they can provide. Here, the semantic registry is introduced, which has the purpose of narrowing the metasearch down from a resource-intensive broadcast-search to a relatively small number of relevant data providers to query.

The primary purpose of the semantic registry is to facilitate search by storing information about HarmoSearch data providers. This information has to contain all data required to contract the data provider automatically (i.e., public search webservice URL and credentials). Furthermore, it has to contain a description of what kind of data the provider offers.

Once a search query is issued to the HarmoSearch system, the semantic registry matches the data provider descriptions against the search criteria specified in the query. It returns all candidate data providers which may have relevant data with respect to this query.

Keeping a complete index of the data providers' data items is highly problematic to keep up to date and does not scale appropriately. Also aggregating and maintaining such an index contradicts the fundamental ideas of the HarmoSearch project. Therefore, the description has to contain only information that does not change very often, outlining the available data from the data providers.

For the purpose of the semantic registry in the HarmoSearch system, retrieving more candidates than required (wrong positives) is not a big issue, while leaving out relevant data providers (wrong negatives) is to be avoided.

2.2 SECONDARY FUNCTION: SERVICE REGISTRY

The second major functionality of the registry is the role of the service registry. This task has been identified through the use cases (see deliverable D2.1) and further detailed in the activity diagrams of the architectural design (see deliverable D2.2).

The goal of this aspect of the semantic registry is to store all information required to access and configure additional services in HarmoSearch workflows. These services are distinct from the data providers in that they work with or refine results of HarmoSearch queries. In this definition of "service" the service providers do not deliver "first hand" data like the data providers but offer functionalities which can be integrated into a HarmoSearch workflow and work with data from a preceding HarmoSearch query.

As an additional task the registry has to provide all required functionalities for these additional services to be published, looked up and used.

3 REQUIREMENTS

There are two kinds of user requirements relevant for the registry. On the one hand, some of the user requirements for search are also of importance for the registry, since the registry has to supply the metasearch process with information about the relevant data providers. On the other hand, there are specific user requirements regarding the semantic registry.

3.1 USER REQUIREMENTS

These requirements, gathered during the first months of the project, are presented in the following section. It starts with requirements for the search process per se. These are not directly requirements for the semantic registry, but the semantic registry has to offer capabilities to support these search requirements.

This is followed by a list of requirements which are directly targeted at the semantic registry.

For each user requirement the following information is provided:

- ID: unique identifier of the requirement
- Author: partner who primarily provided the requirement
- Group: category which the requirement belongs to
- Action: system functionality which the requirement refers to
- Requirement: brief description of the requirement
- Description: detailed description of the requirement
- Comment: additional notes which are useful for the implementation of the requirement
- Priority: importance of the requirement
    - High (i.e. mandatory)
    - Medium (i.e. desiderata)
    - Low (i.e. for the future)
- Examples: examples to aid the understanding of the requirement

ID	SEARCH01
Author	Afidium
Group	Search
Action	Quick Search
Requirement	It should be possible to search for specific items by specifying the name or a unique code

Description	Sometimes items are identified by a unique code or by a key; it should be possible for the user to search by specifying one of these fields in order to get a short list of results (or even just one result) without inserting many search parameters.
Comment	--
Priority	High
Examples	Find the exhibition which has a particular code. Find a museum by its complete name

ID	SEARCH02
Author	Afidium, eCTRL, SPK, EC3, [x+o]
Group	Search
Action	Basic Search
Requirement	It should be possible to search for items by specifying a single criterion or a combination of criteria
Description	It should be possible to query different data providers by specifying one or more search criteria. The search results have to match either all the different conditions (AND) or at least one of them (OR) or a combination of them. The data types of the search fields include numbers, texts, dates, etc.
Comment	In the user interface, the search criteria should be organised and grouped by category in order to improve the usability
Priority	High
Examples	Find all exhibitions in Rome or Napoli on a given date Find all accommodations of category 3, 4 and 5 stars in a given city

ID	SEARCH03
Author	SPK, eCTRL, [x+o]

Group	Search
Action	Basic search with enumeration values
Requirement	It should be possible to fill in some of the search criteria by choosing their values from enumerated value domains
Description	Some of the search parameters are not free text or numeric fields, but drop-down lists whose values must be selected by choosing among a set of predefined values. In order to translate a query from one query language to another, these reference lists have to be translated too.
Comment	It would be useful to have a component to manage the reference lists, i.e. to dynamically retrieve, add, edit or remove values of an enumerated value domain.
Priority	High
Examples	Find all the exhibitions pertaining to "modern art" Find all the "3 stars" accommodations

ID	SEARC004
Author	Afidium, EC3, SPK, [x+o]
Group	Search
Action	Basic search with geographical data
Requirement	It should be possible to search by specifying geographical data and/or the indication of a specific area of interest
Description	In some cases the user has the need to find items which are located in a particular area or close to a specific point of interest
Comment	The search engine should be able to handle geographical data and to compute distances between them. It requires a geographic hierarchy holding several levels of geographic entities and their relations. It requires geocodes of each geographic entity, with distance calculation functionality.
Priority	Medium

Examples	Find all the accommodations close to the centre of Berlin Find all the exhibitions within 1 km from a certain place

ID	SEARCH05
Author	Afidium
Group	Search
Action	Basic search with flexible dates
Requirement	It should be possible to get back not only the results which match exactly the specified dates, but also the ones which are available one or two days before and after
Description	The user searches for a specific item on a specific date and gets back among the results also the items matching the search criteria which are available one or two days before or after the specified date
Comment	
Priority	Medium
Examples	Find all the exhibitions which will take place on May 1st or close to this date

ID	SEARCH06
Author	SPK, [x+o]
Group	Search
Action	Basic search with priority criteria
Requirement	It should be possible to distinguish between criteria which are mandatory and criteria which are optional
Description	From the user's point of view some of the search criteria could be mandatory, while others could be just preferred, i.e. it is not necessary that the search results match these latter criteria but it is a desiderata
Comment	The search engine should handle criteria with different priorities and show to the user the search results ranked according to these priorities

Priority	Low
Examples	Find all the accommodations in Berlin that possibly have a swimming pool Find all the exhibitions pertaining to "modern art" which take place possibly in Berlin

ID	REGISTRY01
Author	Afidium, SPK, [x+o]
Group	Registration
Action	Register data source
Requirement	It should be possible to add or update a data source to the semantic registry and associate it with a Harmonise participant
Description	The description of the data provider should contain all technical aspects required to access the data provider and it should be possible to add additional information about the Harmonise participant besides the Harmonise ID.
Comment	A Harmonise participant can operate several data providers, possibly giving distinct views of an underlying data source
Priority	High
Examples	Register the Euromuse search webservice in the registry, describe its access parameters and associate it to the Euromuse Harmonise user.

ID	REGISTRY02
Author	Afidium
Group	Registration
Action	Register Mappings
Requirement	Mappings stored in the mapping store should be registered with data providers.

Description	Each participant can operate several data providers (e.g., offering event and accommodatin data) and for each data it should be possible to assign a distinct Harmonise mapping.
Comment	
Priority	High
Examples	A harmonise participant offering accommodation and attraction data registers two data providers, one for accommodation and one for attractions. Each provider is assigned a different mapping.

ID	REGISTRY03
Author	[x+o]
Group	Registration
Action	Register external service
Requirement	It should be possible to add or update an external service to the semantic registry.
Description	External services can be used in custom workflows. It should be possible to register all information required to access such an external service as well as information describing the functionality of the service. Services should be associated with a Harmonise participant.
Comment	Each participant can operate an arbitrary number of services.
Priority	Medium
Examples	An Harmonise participant registers a service for adding user ratings to accommodation data in the Harmonise format. This service can be used to enrich accommodation search results with user ratings for the accommodation.

ID	REGISTRY04
Author	Afidium, [x+o]
Group	Registration

Action	Configure access control for registered data providers and services
Requirement	It should be possible to configure who may use a service or access a data provider based on a flexible access control mechanism.
Description	It should be possible to allow or deny service or search access based on users and user groups. Further possible criteria for granting or denying access can be based on the description of the Harmonise participant asking for access.
Comment	
Priority	High
Examples	A data provider registers his service and configures access to be granted only to a group of Harmonise participants paying for access.

ID	REGISTRY05
Author	[x+o]
Group	Registration
Action	Describe data offered by a data provider aided by the mapping.
Requirement	The description of the data offered by a data provider should be aided as far as possible by the mapping registered with the data provider.
Description	If a mapping is registered for a data provider, then it already contains useful information like which domain of the ontology is mapped, etc. Such information should be extracted and automatically assigned to the data description where useful.
Comment	The user should be aided by the information extracted from the mapping in limiting the offered choices for describing the data.
Priority	Low
Examples	Based on the mapping, a data provider offering events is presented with options for describing events only.

ID	REGISTRY06
Author	Afidium, eCTRL
Group	Access
Action	Discover services to be used in a workflow
Requirement	Services must be searchable and browsable. All information required to use a service must be accessible.
Description	External services for adding functionality to HarmoSearch workflows must be discoverable in order to be used. They should be browsable and searchable. Service descriptions must contain technical access information and possibly also information on how to gain access rights to the service.
Comment	
Priority	Medium
Examples	Harmonise participants looking for a service to add user ratings to accommodations should be able to discover the service added in REGISTRY03.

ID	REGISTRY07
Author	Afidium, eCTRL
Group	Access
Action	Discover data providers
Requirement	Harmonise participants should be able to look up other participants in order to agree on data exchange.
Description	Data providers should be searchable and browsable based on the data they provide as well as on the participant's description. Besides technical information, also information to contact the Harmonise participant should be available.
Comment	In order to access non-free data providers, agreements are required. There has to be enough information on the platform to be able to start negotiating this.

Priority	Medium
Examples	A harmonise participant wants to access data on specific events which are not freely accessible. He must look up the specific provider and contact him in order to negotiate access.

ID	REGISTRY08
Author	Afidium
Group	Registration
Action	Specify filter for notification when matching data providers become available
Requirement	Harmonise participants should be able to specify filter criteria, indicating an interest in specific data. When such data becomes available a notification should be sent.
Description	A harmonise participant should be able to specify criteria for data he is interested in like when discovering data providers (REGISTRY07). These criteria should be stored in the registry and an alert (e.g., email notification) be sent when a new or updated data provider matched the specification.
Comment	
Priority	Medium
Examples	A harmonise participants wants to be notified whenever new accommodation providers in Italy become available on the Harmonise network.

3.2 TECHNICAL REQUIREMENTS

This section contains requirements for the semantic registry derived from the overall architecture and the foreseen interaction of the different components (see deliverable D2.2). Furthermore, additional technical requirements are listed here.

For each user requirement the following information is provided:

- ID: unique identifier of the requirement
- Group: category which the requirement belongs to
- Requirement: brief description of the requirement
- Description: detailed description of the requirement

- Comment: additional notes which are useful for the implementation of the requirement
- Priority: importance of the requirement
    - High (i.e. mandatory)
    - Medium (i.e. desiderata)
    - Low (i.e. for the future)
- Examples: examples to aid the understanding of the requirement

ID	TECH01
Group	Data Description
Requirement	Data description should outline the set of all provided data items rather than index them.
Description	Indexing all data items of a large number of data providers does not scale and soon becomes infeasible. Therefore, the description of data in the registry should rather outline the data in a stable way.
Comment	The description should be done in such a way that it doesn't change very often.
Priority	High
Examples	A data provider offering information about Spa-Hotels in Austria describes this fact, but does not add any volatile details.

ID	TECH02
Group	Data Description
Requirement	Data should be described in terms of the Harmonise Ontology
Description	The domain dependent part of the registry data schema should use the Harmonise ontology directly.
Comment	In this way no separate domain dependent registry ontology is introduced.
Priority	High
Examples	

ID	TECH03
Group	Data Description
Requirement	The Harmonise ontology should be stored only in one place.
Description	There should be no duplicate storage of the Harmonise ontology. In this way it is ensured that there are no different versions of the Harmonise ontology in use.
Comment	Whether Harmonise ontology should be stored in the registry or in some other component is subject to discussion.
Priority	Medium
Examples	

ID	TECH04
Group	Data Description
Requirement	It should be possible to load the description of data providers from an external source, i.e., from the data provider itself.
Description	There should be the possibility for data providers to maintain their description not on the semantic registry itself but on their own server, providing the description in a simple file. This file should then be loaded into the semantic registry on demand.
Comment	When to load and how to cache is subject to discussion.
Priority	Low
Examples	A data provider offers the description of his data in a file on his own webserver.

ID	TECH05
Group	Data Description

Requirement	Data providers must be able to describe what kinds of data – i.e. what fields of the Harmonise ontology – they offer.
Description	This requirement is based on the concept of subdomains in the HarmoSearch query language and specified a kind of compliance level with a given set of data fields. A data provider must be able to either select a predefined compliance level ("subdomain") or create an individual description.
Comment	There should not be the need for the data provider to know all existing compliance levels in order to create an individual one.
Priority	High
Examples	Euromuse wants to specify which data fields a data supplier has to provide in order to be acceptable as a Euromuse data source.

ID	TECH06
Group	Data Description
Requirement	Providers of additional services must be able to describe the kind of input and output data their services expect and deliver.
Description	Harmonise participants offering additional services to be used in HarmoSearch workflows must be able to define what kind of data items they expect as input and what kind of data items they deliver as output. The description should be possible either as a predefined or as a specifically created compliance level ("subdomain").
Comment	Subdomains should be described like in TECH05. Extensions of this mechanism due to implementation detail may become necessary thought.
Priority	High
Examples	A service for enriching Euromuse-Compliant exhibition data with expert reviews defines its input data as subdomain "Euromuse" and it's output data as an individual subdomain, enriched by the reviews.

ID	TECH07
Group	Reasoning
Requirement	The registry must have semantic reasoning capabilities
Description	There are several places where semantic reasoning capabilities are required. For example when checking a HarmoSearch query against the data provider's data description or when applying geo-reasoning processes on the registered data. An important application of semantic reasoning is the matching of subdomains of the Harmonise ontology against each other.
Comment	Given the availability of the Harmonise ontology as an RDF description and the mature tools for RDF and OWL reasoning, an OWL based reasoning mechanism is desirable.
Priority	High
Examples	Euromuse wants to query all Harmonise participants supplying exhibition data conformant to their data requirements (e.g., having a title, description, etc.). The registry should automatically check the data conformance of all registered data sources and return the appropriate ones.

ID	TECH08
Group	Reasoning
Requirement	All data required for reasoning should be loaded from external sources where possible.
Description	There is need for semantic reasoning in the registry (see TECH07), and some of the reasoning processes require additional data (e.g., geo-reasoning). This data should be loaded from external sources or external reasoning services should be employed wherever possible.
Comment	This reduces the need for manually updating externally provided data in the system.
Priority	Low
Examples	

ID	TECH09
Group	Integration
Requirement	User interfaces for using and accessing the registry must be integrated into the overall HarmoSearch solution.
Description	The registry must provide several user interfaces for defining and managing the descriptions and configurations stored in the semantic registry. These have to be integrated in the appropriate places in the overall solution.
Comment	Integration should be seamless and might require that a part of the user interface is programmatically in another component.
Priority	Medium
Examples	The user interfaces for creating a data provider and associating it with a Harmonise participant should be located in the participants' management section.

ID	TECH10
Group	Integration
Requirement	The semantic registry should offer webservice interfaces for all relevant functionalities.
Description	In order to allow integration in distributed systems, the semantic registry should offer webservice interfaces for all relevant functionalities. Special consideration must be given to checking access rights in these webservice interfaces.
Comment	Integration with other HarmoSearch components may be done via these webservice interfaces or more directly through the registry's API.
Priority	Medium
Examples	

4 STATE OF THE ART

The semantic registry has the task of storing metadata about the content different data providers have available. This includes reasoning on data items and on the available mappings as well as reasoning on user configurations and created workflows in order to find suitable data providers for a given query by a given user.

Furthermore, not only the definition of available data but also the definition of interests for specific content items with respect to configured workflows and access rights must be supported. Finally, also the description and discovery of third party services to be included in the workflows must be possible.

These specific requirements cannot be directly fulfilled by existing registry implementations. However, an existing system could be used as a basis on top of which the extensions for the HarmoSearch project are implemented.

The following sections provide an overview of the identified candidates for base technologies, which were analysed to estimate their suitability for HarmoSearch. This is followed by a list of functional and non-functional requirements for building the HarmoSearch semantic registry.

Based on the assessment of the candidates against these requirements, a base technology is chosen. The architecture for building the HarmoSearch semantic registry based on this choice is then outlined in section 5.

4.1 THE OMAR EBXML REGISTRY

The Object, Metadata and Artifacts Registry (OMAR)[1] is an implementation of the ebXML registry specification, supporting XML[2] based business interactions. It provides a set of services which enables the sharing of content and metadata between different participants. It allows managing any content type and the standardised metadata that describe it.

OMAR offers several features that make it an interesting candidate of a base technology for the HarmoSearch Semantic Registry. Among these is a role based access control, facilities for the cataloguing XML content as well as content based event-notation. OMAR offers Java user interfaces as well as API access for all relevant user actions.

OMAR is built in Java as a web application running on an application server (Apache Tomcat is recommended). It needs a relational database system to operate. Besides Derby and HSQLDB which are shipped with the application, also PostgreSQL and Oracle databases can be used. Compatibility with other database management systems needs to be checked.

[1] OMAR is the OASIS ebXML reference registry, *http://ebxmlrr.sourceforge.net/index.html*

[2] Extensible markup language, *http://www.w3.org/TR/2006/REC-xml11-20060816/*

OMAR is distributed as open-source software under the very liberal "freebxml License", which makes no restrictions on deriving and selling software based on the OMAR registry.

The OMAR registry allows to define own data schemata and offers indexing and querying capabilities. But it is not, out of the box, suitable to describe data content on a meta-level or query data in terms of a specified pre-existing ontology (i.e., the Harmonise ontology). Also it does not feature any reasoning support. Therefore, in order to be used as basis for the Harmonise registry, some heavy modifications of OMAR's data layer will be necessary.

The last version of the OMAR registry was released in July 2007. Therefore, the base technologies used in the development of OMAR are somewhat outdated by now. With its acceptable quality of the documentation, an adaptation of the OMAR registry for the purposes of the HarmoSearch semantic registry appears feasible.

However, first experiments showed that the OMAR project has a high complexity and is not trivial to operate, let alone modify. For use as HarmoSearch semantic registry, a relatively large number of modifications would be required, especially with respect to semantic metadata description and reasoning. Therefore, we estimate the effort to adapt the OMAR registry to the HarmoSearch requirements as very high - similar to the creation of a new registry.

4.2 THE FUSION SEMANTIC REGISTRY

The *FUSION Semantic Registry*[3] is a semantically-enhanced service registry. It is based on the UDDI[4] specification but adds machine understandable semantics for specifying and discovering services. Therefore, unlike its UDDI base, the FUSION Semantic Registry supports fully automated service discovery.

It was developed in the context of the IST Research Project "FUSION", funded by the European Commission in the 6th Framework Programme. , led by SAP AG and a consortium consisting of 14 partners from five European countries (Germany, Poland, Greece, Hungary, Bulgaria).

The FUSION registry uses of SAWSDL[5] annotating the service interface descriptions. Furthermore, it makes use of OWL-DL[6] for describing service capabilities and reasoning.

FUSION is implemented in Java and runs as a standalone web application on a standard web application server (e.g., Apache Tomcat), using a UDDI compliant

[3] FUSION Semantic Registry, *http://www.seerc.org/fusion/semanticregistry/*

[4] Universal Description, Discovery and Integration; A standard for registering and locating web services. *http://uddi.org/pubs/ProgrammersAPI-V2.04-Published-20020719.htm*

[5] Semantic Annotations for WSDL and XML Schema. *http://www.w3.org/TR/sawsdl/*

[6] OWL Web Ontology Language. *http://www.w3.org/TR/owl-features/*

server (e.g., the open-source JUDDI[7] server implementation). The FUSION registry itself is released as open source software under the GPL v3.0[8] software license.

UDDI, the base technology on and for which the FUSION semantic registry is built, has been outdated by now. UDDI was written in August 2000, at a time when the authors had a vision of a world in which consumers of Web Services would be linked up with providers through a public or private dynamic brokerage system. In this vision, anyone needing a service such as credit card authentication would go to their service broker and select one supporting the desired service interface and meeting other criteria.

However, UDDI has not been widely adopted in the way its designers had hoped. IBM, Microsoft, and SAP announced they were closing their public UDDI nodes in January 2006. The group defining UDDI, the OASIS Universal Description, Discovery, and Integration (UDDI) Specification Technical Committee has also been closed.

This development certainly had an impact on the Fusion semantic registry, which had its last version update in October 2008. However, it does not necessarily mean that it is unsuitable for the requirements of the HarmoSearch semantic registry. A modification of the service description and reasoning structures to fit the needs of HarmoSearch appears possible. There effort required, especially with the low documentation quality, will be significant though – rivalling that of a new implementation from scratch.

4.3 SEMANTIC DATA STORES

From semantic web research, some predominant formats and technologies have established themselves for meaningful description and semantic processing of data. These technologies include the Resource Description Framework (RDF)[9], a variety of data interchange formats and notations such as RDF Schema (RDFS)[10] and the Web Ontology Language (OWL), all of which are intended to provide a formal description of concepts, terms, and relationships within a given knowledge domain. These technologies have already reached a very mature and stable status, so that they are well suited to be employed where challenging data and metadata centric problems need to be solved.

In the context of the HarmoSearch semantic registry, building a new and custom-made semantic registry based on existing semantic web technologies appears feasible. The strengths of these technologies lies in describing data on a meta-level and reasoning about these descriptions. On an abstract level these are the main tasks the semantic registry has to fulfil.

[7] An open source Java implementation of the UDDI specification. *http://juddi.apache.org/*

[8] GNU General Public License, a copyleft (“viral”) open source license, *http://www.gnu.org/licenses/gpl-3.0.html*

[9] *http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/*

[10] *http://www.w3.org/TR/2004/REC-rdf-schema-20040210/*

The benefit over a custom-made solution over the adoption of existing registries is that the work can be focused on the tasks of the semantic registry instead of caring about the requirements the base registry imposes. Furthermore, while the existing registries offer superior support regarding access control, their ability to integrate well into the HarmoSearch solution and handle data and reasoning as desired for the HarmoSearch semantic registry is insufficient to be used out of the box. In essence we estimate that the effort for adopting the existing solutions and modifying them to suite the requirements of the HarmoSearch project is comparable to building a new solution based on the more mature and well-maintained semantic web technologies.

For the task of creating a semantic registry, first of all the choice of base technology is of relevance, followed by the selection of a framework for implementing it and a semantic storage component.

Since the Harmonise ontology is distributed in RDF format, adopting RDF as base technology is a natural choice. However, in order to integrate more advanced semantic reasoning, the extension to OWL, is desirable for its more powerful description and reasoning capabilities.

The Web Ontology Language (OWL) is knowledge representation languages. It is characterised by formal semantics and RDF/XML-based serializations for the Semantic Web. OWL is endorsed by the World Wide Web Consortium (W3C) and has attracted wide interest and support. OWL Full is intended to be compatible with RDF Schema (RDFS), and to be capable of augmenting the meanings of existing Resource Description Framework (RDF) vocabulary. For performance reasons a subset of OWL, OWL-DL (description logic) is used in real applications. It allows for efficient reasoning while retaining almost all of the capabilities of OWL-Full.

Data repositories for OWL and RDF data and metadata are called triple stores, since data is stored in triples of subject, predicate and object (e.g., "Fritz", "is-a", "Cat"). There exist a large number of open source triple stores, the most popular of which are Allegrograph[11], Jena[12], Mulgara[13], Sesame[14] and Virtuoso[15].

All of these are built for high performance data storage and reasoning, however a comparison of supported reasoning mechanisms and current state of support for the

[11] A closed source graph database, capable of storing RDF data, *http://www.franz.com/agraph/allegrograph/*

[12] The Jena Semantic Web Framework is an open source a semantic data store and Java API, *http://jena.sourceforge.net/*

[13] Mulgara is a scalable open source RDF database, *http://www.mulgara.org/*

[14] OpenRDF Sesame is an open source a semantic data store and Java API, *http://www.openrdf.org/*

[15] OpenLink Virtuoso is a „universal server", combining relational database, RDF, XML, free-text, web application server and file server functionality, *http://virtuoso.openlinksw.com/*

new SPARQL 1.1[16] query specification found Jena to be the most promising candidate as a semantic data store and framework.

Jena also supplies a powerful Java framework for working with and reasoning on OWL and RDF data, which is used in many of the other systems as well and, especially in combination with the use of Jena's triple store, is a good choice for an implementation framework. Furthermore, Jena has a comprehensive documentation and an active and supportive community.

The Jena framework is distributed under an own, very liberal licence which has the retaining of existing copyright notices as its only requirement. Apart from that it may be redistributed in source or binary form with or without modifications without restrictions. The latest version of the Jena framework was released in May 2011.

4.4 REQUIREMENTS EVALUATION OF BASE TECHNOLOGIES

In this section the base technology choices presented in the previous paragraphs are evaluated against the requirements identified in section 3 and additional technical requirements like community support and maturity of the solution. We then evaluate the fitness of the different base technologies with respect to these requirements. Finally, based on this evaluation, we select the most appropriate base technology to build the HarmoSearch semantic registry.

ID	Requirement	OMAR	FUSION	TRIPPLE STORE
FR1	Describe data of data providers	+	+	+++
FR2	Analyse mapping to aide data description	+	+	+
FR3	Describe external services	++	+++	++
FR4	Associate data and services with harmonise participants	+++	+++	+++
FR5	Provide flexible access control	+++	+++	+
FR6	Match HarmoSearch query with data providers' data descriptions	+	+	++
FR7	Match query and data providers using Harmonise value lists	+++	+++	+

[16] SPARQL Protocol And RDF Query Language, a query language for RDF, *http://www.w3.org/TR/rdf-sparql-query/*

FR8	Provide possibility to select data providers directly as per query language specification	++	++	++
FR9	Possibility to use or provide geo-reasoning services	+	+	++
FR10	Possibility to specify filter criteria for notification	+++	+	++
FR11	Describe data – don't index items	+	+	+++
FR12	Describe data in Harmonise terms	++	+	+++
FR13	Import data from external sources on demand	+	+	++
FR14	Provide reasoning capability on Harmonise ontology	+	++	+++
FR15	Provide reasoning capability on external data (e.g., geo-reasoning)	+	++	+++
FR16	Integrate user interfaces into Harmonise platform	+	+	+
FR17	Offer webservice interfaces for relevant functionality	++	+++	+
NFR1	Open Source solution	+++	+++	+++
NFR2	Documentation quality	++	+	+++
NFR3	Community support	+	+	+++
NFR4	Continuous development activity	+	+	+++
NFR5	Conformance with widely accepted standards (non-proprietary solution)	+++	++	+++
NFR6	Effort of adoption for envisaged Harmonise solution	+	+	+
Overall Evaluation		**40**	**39**	**51**

4.5 SELECTION OF BASE TECHNOLOGY

As can be seen in the evaluation table, all candidate base technologies have their strengths and weaknesses. However, the implementation based on a semantic data store is evaluated best.

This is mainly due to the fact that the other base solutions are lacking some distinctive features which make for a relatively high effort in adapting these solutions. Since the effort is estimated to rival that of a new implementation, a semantic data store is a viable option as a base technology.

Regarding most of the desired features for the HarmoSearch semantic registry, a new implementation is favoured above adaptations of an existing tool. This is due to the highly specific requirements of the HarmoSearch project, which are hard to be fulfilled by general purpose registries without extensive modifications.

Finally, the state of documentation and ongoing development and support clearly favours the semantic data store approach over the other solutions.

Therefore, a new implementation of the semantic registry based on the Jena Semantic Web framework is chosen for the HarmoSearch semantic registry. The next section describes the architecture for the semantic registry following from this decision.

5 SEMANTIC REGISTRY ARCHITECTURE

This section gives an overview of how the semantic registry can be built based on a semantic data store. The basic components and their interactions with each other and with other components of the HarmoSearch system are outlined.

Furthermore, the interaction of the semantic registry with the rest of the HarmoSearch system with respect to the main use cases metasearch and service configuration are described in some detail.

5.1 SEMANTIC REGISTRY ARCHITECTURE OVERVIEW

The internal components of the semantic registry are depicted in *Figure 1* below. For each component a short description follows.

5.1.1 Query Manager

The query manager's task is to provide an interface for the metasearch and/or workflow component to access the data registry part of the semantic registry. It exposes all relevant interfaces regarding data providers. Its main task is therefore, to accept a HarmoSearch query and return a number of data providers relevant for the query. Furthermore it offers facilities to search for data providers and return detailed information.

5.1.2 Service Manager

The service manager exposes interfaces for accessing data about the services registered in the service registry part of the semantic registry. The main task is to allow to search for registered services and to return all data required in order to use the services in a HarmoSearch workflow.

5.1.3 Query Transformer

The query transformer is the registry component which analyses a provided HarmoSearch query (see deliverable D4.1) and translates it into a SPARQL query, which in turn is processed by the SPARQL processor of the RDF/OWL store. The result, containing information about which data provider to query and how to access them, is sent back to the query manager.

5.1.4 RDF/OWL Data Store

The semantic data store contains all components required to operate the semantic database for the HarmoSearch semantic registry. As the most suitable technology to provide this service, the Jena framework has been identified. Together with its standalone implementation of Fuseki, a SPARQL powered RDF/OWL triple store and the Pellet OWL reasoned, the following subcomponents are covered:

- RDF/OWL Store
- OWL Reasoner
- SPARQL Processor
- Interface to access external RDF sources

5.1.5 Data Description Interface

This component exposes the interfaces for registering data providers in the semantic registry and describing the data they offer. This description process is aided by the mapping analyser and its result stored in the semantic data store. Furthermore, this component interacts with the access control manager in order to check and define access control policies. This interface may be exposed as a webservice and is used by the semantic registries graphical user interface.

5.1.6 Mapping Analyser

This component has the sole purpose of aiding the data description process by analyzing the relevant mapping and pre-processing and/or limiting the user's choices in the data description GUI. It directly interacts with the Harmonise Mapping store in order to retrieve the mapping files for analysis.

5.1.7 Notification Manager

This component's task is to allow registering criteria for defining an interest in specific data providers and to define a workflow to be triggered should a relevant data provider become available. The notification manager therefore listens to changes on the data description interface and triggers the predefined actions in the workflow engine (e.g., email notification).

5.1.8 Service Description Interface

Like the data description interface exposes all functionalities for registering new data providers, the service description interface exposes all the required functionalities for registering new services to be used in HarmoSearch workflows. This mainly encompasses the description of the service and the ability to search for them. The component directly interacts with the semantic data store in order to store and access this information.

5.1.9 Access Control Manager

The access control manager directly interacts with the Harmonise Access Control Module and offers all relevant features related to access-control for the other registry components. Possible additional required features exceeding the capabilities of the Harmonise Access Control Module are implemented in this component.

5.1.10 Graphical User Interfaces

Those functionalities of the semantic registry which have a direct interaction with the user need to provide an adequate user interface. These interfaces should to be integrated into the overall HarmoSearch solution. Therefore, they are not necessarily implemented in the registry component itself, but could instead only make use of the exposed functionalities of the registry either through a webservice interface or more directly through the registry's API.

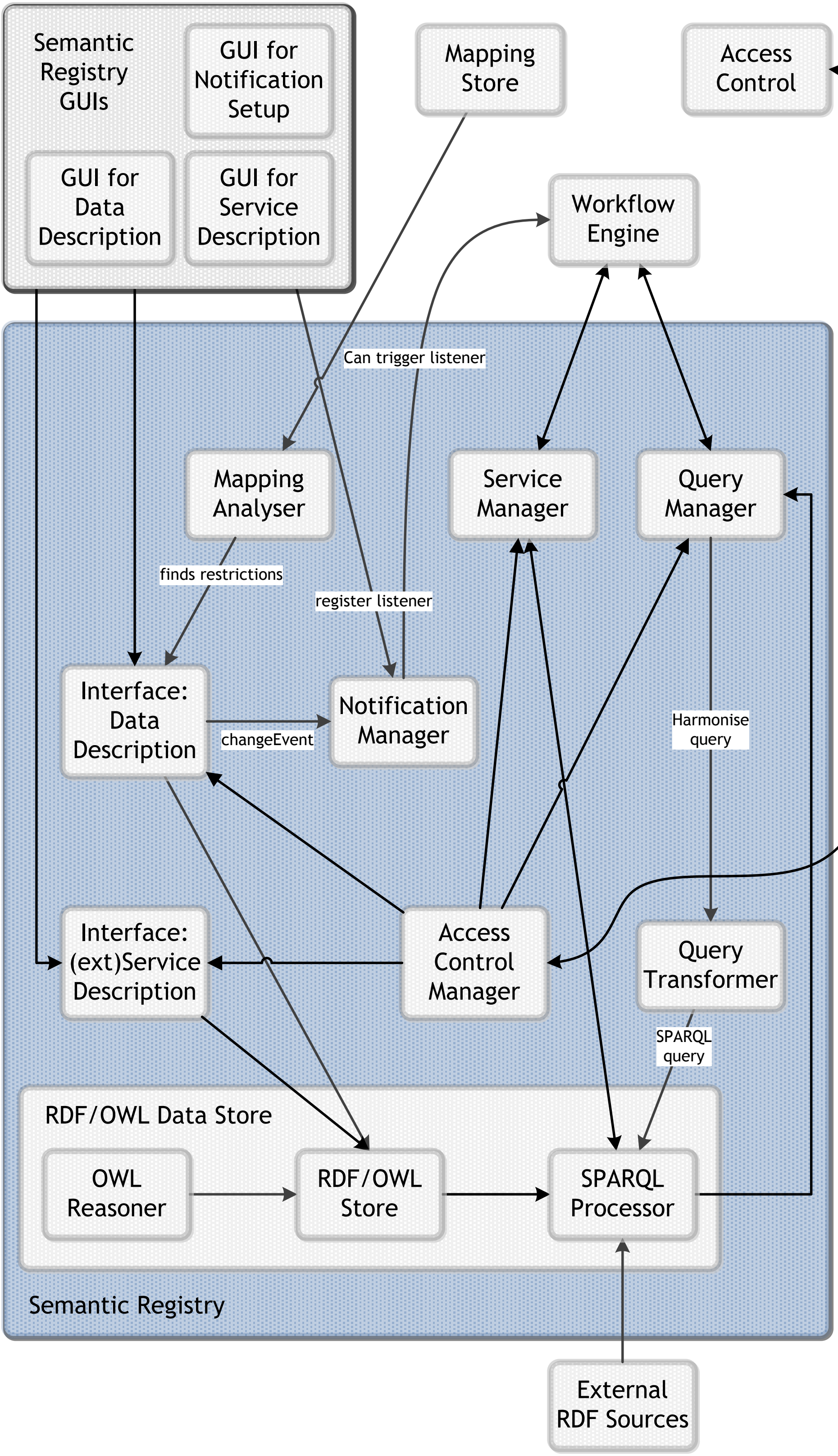

Figure 1: HarmoSearch Registry Components and Interactions

5.2 INTERACTIONS IN MAIN USE CASES

Since the semantic registry can fulfil its functions only in interaction with other Harmonise components, the required interactions give additional insight into the functionality the registry has to expose. The following sections briefly describe the interactions in the main use cases: *Metasearch* and *Service Lookup*.

5.2.1 Metasearch

The main use case for the semantic registry is to aid the metasearch in identifying appropriate data providers to connect to for a given HarmoSearch search query. In order to fulfil this goal, the semantic registry has to be integrated in the HarmoSearch search workflow. A basic overview of the interaction triggered by a HarmoSearch search query is depicted in the sequence diagram in *Figure 2*.

Note that here only a single provider is actually depicted, whereas this process is executed for all providers in parallel. Calls to methods with provider[x], and results of provider[x] stand for iterating over all providers.

Triggered by an incoming search query, first the metasearch engine is activated to codify the query using the HarmoSearch query language. Then the Workflow Engine queries the semantic registry, giving the HarmoSearch query as input. The expected output is a list of data providers relevant for the query together with all required information to access the provider's search interface. Basically, this means the information required to set up and configure the query processors and data connectors for this search.

For each of the retrieved candidate providers the workflow engine then transforms the query into the provider's format using the query processor. The transformed query is sent to the provider using the correct data connector, which retrieves the results. These results are then transformed into the Harmonise format using the reconciliation engine and stored using the metasearch engine.

An overview of the interaction between the different components is shown in *Figure 3*. The workflow engine is the central hub, coordinating the search process triggered by the incoming search request.

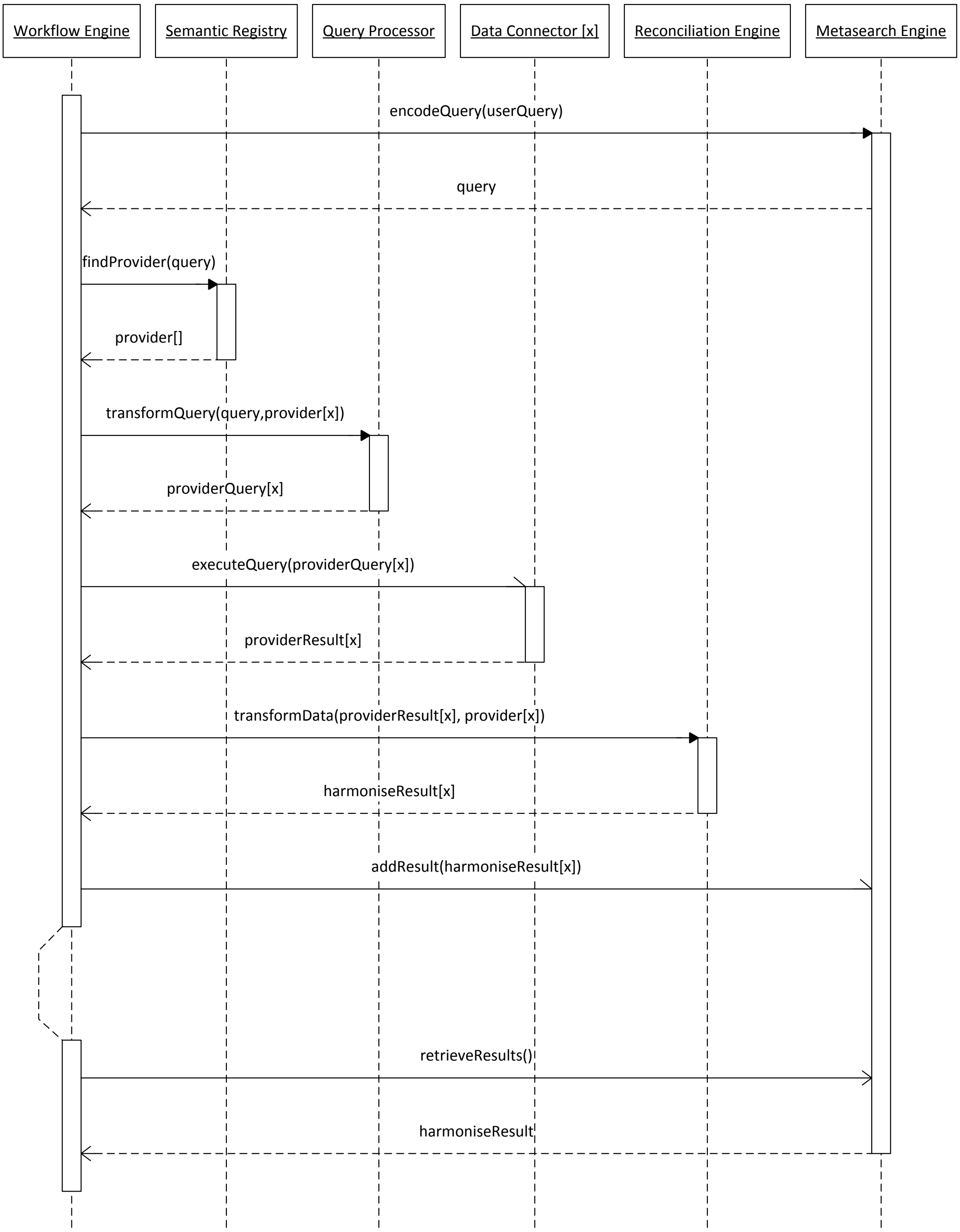

Figure 2: Sequence diagram of a Metasearch Process

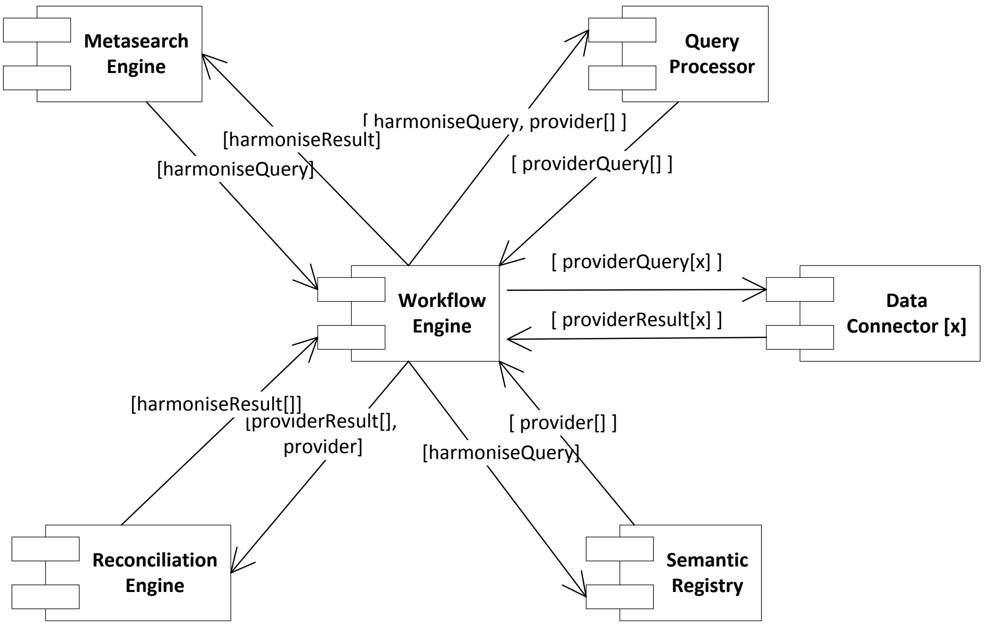

Figure 3: Interaction of Components in a Metasearch Process.

5.2.2 (External) Service Description and Lookup

In order to act as a service registry, the semantic registry exposes a specific interface for registering, describing and searching for (external) services which provide additional functionalities to be configured in HarmoSearch workflows.

The workflow engine accesses this interface for its configuration and execution. The interaction is therefore limited to, on the one hand offering a search functionality which returns all relevant data for selecting and adding a service to a workflow. On the other hand a function for retrieving relevant technical data about a service given its ID as a parameter has to be provided in order for the Workflow engine to instantiate and execute workflows containing external services.

6 LIST OF FIGURES