SEKT Portal - Mediation (WP4) EU-IST Project IST-2003-506826 SEKT
SEKT: Semantically Enabled Knowledge Technologies

D4.6.1.1 Report on ontology mediation for case studies V1

Jos de Bruijn[1] and Cristina Feier[1]

with contributions from Nuria Casellas[2], Pompeu Casanovas Romeu[2], Mercedes Blázquez Cívico[3], Jesús Contreras[3], Ian Thurlow[4], Roland Zeilbeck[5]

Abstract
The aim of this deliverable is to identify the requirements for mediation for the SEKT case studies. The data sources from each case study are investigated together with the relationships between them and with the scenarios in which two or more of these data sources are used in conjunction, i.e. where data integration is needed. The requirements for mediation are identified based on these scenarios. We should note that as a result of our analysis we identified the opportunity of some architectural changes for two of the case studies. The new data source landscapes proposed together with guidelines about different mediation approaches should serve as a pillar for the further development of the case studies. Also the identified requirements show that the main mediation functionality on which the tools developed by the WP4 should focus on is ontology alignment.

Keyword list: SEKT case studies, data mediation, ontology mapping, ontology merging, ontology alignment

WP4 Ontology Mediation	
Report	PU
Contractual date of delivery	Actual date of delivery
31.06.2005	05.08.2005

1 University of Innsbruck
2 Universitat Autonoma de Barcelona
3 Intelligent Software Components S.A.
4 British Telecommunications plc.
5 Siemens Business Services

SEKT Consortium

This document is part of a research project partially funded by the IST Programme of the Commission of the European Communities as project number IST-2003-506826.

British Telecommunications plc.
Orion 5/12, Adastral Park
Ipswich IP5 3RE
UK
Tel: +44 1473 609583, Fax: +44 1473 609832
Contact person: John Davies
E-mail: john.nj.davies@bt.com

Empolis GmbH
Europaallee 10
67657 Kaiserslautern
Germany
Tel: +49 631 303 5540
Fax: +49 631 303 5507
Contact person: Ralph Traphöner
E-mail: ralph.traphoener@empolis.com

Jozef Stefan Institute
Jamova 39
1000 Ljubljana
Slovenia
Tel: +386 1 4773 778, Fax: +386 1 4251 038
Contact person: Marko Grobelnik
E-mail: marko.grobelnik@ijs.si

University of Karlsruhe, Institute AIFB
Englerstr. 28
D-76128 Karlsruhe
Germany
Tel: +49 721 608 6592
Fax: +49 721 608 6580
Contact person: York Sure
E-mail: sure@aifb.uni-karlsruhe.de

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello St.
Sheffield S1 4DP
UK
Tel: +44 114 222 1891
Fax: +44 114 222 1810
Contact person: Hamish Cunningham
E-mail: hamish@dcs.shef.ac.uk

University of Innsbruck
Institute of Computer Science
Techikerstraße 13
6020 Innsbruck
Austria
Tel: +43 512 507 6475
Fax: +43 512 507 9872
Contact person: Jos de Bruijn
E-mail: jos.de-bruijn@deri.ie

Intelligent Software Components S.A.
Pedro de Valdivia, 10
28006
Madrid
Spain
Tel: +34 913 349 797
Fax: +49 34 913 349 799
Contact person: Richard Benjamins
E-mail: rbenjamins@isoco.com

Kea-pro GmbH
Tal
6464 Springen
Switzerland
Tel: +41 41 879 00
Fax: 41 41 879 00 13
Contact person: Tom Bösser
E-mail: tb@keapro.net

Ontoprise GmbH
Amalienbadstr. 36
76227 Karlsruhe
Germany
Tel: +49 721 50980912
Fax: +49 721 50980911
Contact person: Hans-Peter Schnurr
E-mail: schnurr@ontoprise.de

Sirma AI EAD, Ontotext Lab
135 Tsarigradsko Shose
Sofia 1784
Bulgaria
Tel: +359 2 9768 303, Fax: +359 2 9768 311
Contact person: Atanas Kiryakov
E-mail: naso@sirma.bg

Vrije Universiteit Amsterdam (VUA)
Department of Computer Sciences
De Boelelaan 1081a
1081 HV Amsterdam
The Netherlands
Tel: +31 20 444 7731, Fax: +31 84 221 4294
Contact person: Frank van Harmelen
E-mail: frank.van.harmelen@cs.vu.nl

Universitat Autonoma de Barcelona
Edifici B, Campus de la UAB
08193 Bellaterra (Cerdanyola del Vall` es)
Barcelona
Spain
Tel: +34 93 581 22 35, Fax: +34 93 581 29 88
Contact person: Pompeu Casanovas Romeu
E-mail: pompeu.casanovas@uab.es

Siemens Business Services
Otto-Hahn-Ring 6
81739 München
Germany
Contact person: Dirk Ramhorst
E-mail: dirk.ramhorst@siemens.com

Executive Summary

This document is a report on ontology mediation for the case studies. It provides guidelines regarding different tasks related to mediation, by identifying all the possible approaches for each such task. Then, based on the case study descriptions it gathers all the mediation requirements for the specific use case.

The methodology used for gathering the requirements for mediation was the following: all the partners responsible for one of the case studies were asked to complete a questionnaire according to which they had to describe all the data sources related with the specific case study and to answer to a set of questions about the relationships between those data sources. Then, having the data source landscapes in mind, the use cases for each case study were analysed for identifying the situations where mediation is needed.

The results were as follows. For the Siemens Case Study, some requirements were identified, but they are not very precise due to the fact that the data sources and the relationships between those in the new ontology-model of Siemens Business Services are not clearly defined yet. For the Legal Case Study precise guidelines for mediation were suggested that would lead also to a simplification of the architecture of this case study. For the Digital Library Case Study, while there are some open issues regarding the architecture of the case study, some clear requirements for mediation were derived and some suggestions were made to the partners for enhancing the architecture.

The need of ontology alignment was prevalent in all the case studies, thus the tools developed by WP4 should focus on providing this functionality.

Contents

1 Introduction

The intent of this document is to gather requirements on ontology mediation from the SEKT case study partners.

The SEKT data manual [Ehrig et al., 2004] captures the nature of the data sources used in the SEKT case studies and captures differences and overlaps between the data sources. However, the data manual does not capture the description of the data sources used in the Siemens case study and does not capture the actual requirements on ontology mediation in the case studies.

For gathering information regarding the case studies, besides the information contained in the SEKT deliverables that describe them, we asked the partners that are concerned with these case studies to provide a description of the data sources that appear in the case study and to answer to a set of questions meant to elicit/clarify the mediation requirements.

Section 2 of this deliverable is concerned with general guidelines regarding how different types of mediation can be performed, that together with the specific requirements identified for each case study, should offer a comprehensive picture of the role of mediation and the way this can be implemented for each case study.

In Section 3 of this deliverable each case study is analysed for gathering specific requirements for mediation. The description of each case study is structured as follows. We summarize the data sources used in the case study. We then describe the use cases identified for this case study, from which we derive scenarios in which mediation can be used. The requirements for mediation are derived from this scenarios.

Some conclusions are drawn in Section 4.

Some of the questions that were posed to our partners together with the answers received from them are listed in the Appendix.

2. Mediation Guidelines

The SEKT Work Package 4 provides a number of tools to the case study partners to enable ontology mediation. Roughly, these tools are:

- **Mapping editor.** The editor allows the user to specify the relationships between ontologies in a graphical way and also to edit existing mappings. The editor is the main entry point for the ontology editing activity. The mappings are retrieved from and saved to the ontology mapping store. The mapping discovery component is invoked from the editor.
- **Mapping discovery component.** The mapping discovery component can be used for two different tasks: (1) to discover (parts of) the mapping between two ontologies and (2) to discover correspondences between ontologies; these correspondences can be used to identify which concepts in two ontologies need to be merged in an ontology merging scenario.
- **Mapping store**. The ontology mappings are stored in a central location, from which they can be retrieved.
- **Querying**. Given a target ontology and one or more source ontologies and their mappings, the querying component can be used to query all ontologies and their underlying databases in terms of the target ontology.

These main components can be used in many different ways, based on the needs of the actual ontology mediation scenario. In this section we describe a number of ontology mediation scenarios and indicate how the tools provided by the ontology mediation Work Package should be used.
Ontology mediation, and in fact all Semantic Web technology, is mainly used to achieve information integration. With information integration we mean the use of information originating from different (possibly heterogeneous) sources for a specific purpose. In a stricter sense, information integration is the interlinking of a number of different information sources in order to achieve a single view of the information in the different sources.
We will first describe the generic information integration scenarios. We will then describe how ontology mapping can be used in these scenarios. Please note that this chapter should not be viewed as a user manual for the ontology mediation tools developed in the Work package. For the user manual and user guide for the mediation tools we refer the reader to deliverable D4.5.3.

2.1 Information Integration Scenarios

We identify two major paradigms in information integration: (1) merging data models into a central model and (2) aligning and mapping models. In the ontology engineering community these approaches are known as Ontology Merging and Ontology Aligning**.**
[Noy and Musen, 1999] clarify the difference between ontology merging and ontology aligning. When merging two ontologies, a single coherent ontology is created that is a merged version of the two original ontologies. When aligning two ontologies, the two original ontologies persist, with a number of links established between them, allowing the aligned ontologies to reuse information from one another. Therefore, the alignment of ontologies is usually part of the ontology merging process.

Solutions can be further classified along two dimensions: a run-time and a design-time dimension. The run-time dimension concerns with the way the user views the data in the system during operation. The design-time dimension concerns with the way the models of the disparate data sources are integrated.

In the run-time, or user-centered dimension we distinguish two approaches: (1) the local model and (2) the global model approach. The difference between these two approaches is whether, in interactions with the system, the user can use his/her own local data model, or whether the user needs to conform to a global model when interacting with the system:

- In the local model, or local ontology approach, the user is represented by an agent in the system and this agent represents the user with its own local data model. The agent performs the translation between the user's local model and either the global model or other local models in order to allow interaction with multiple data sources in the system. An example of the local model approach is the KRAFT project [Preece et al., 2001].
- In the global model, or global ontology approach, the user will view the system through the global data model using a mediator, which is "a system that supports an integrated view over multiple information sources" [Hull, 1997]. Note that in the local model approach, a user agent will in most cases also contact a mediator in order to allow inter-operation with the system, which contains multiple information sources. An example is the approach taken in the COG project [de Bruijn, 2004].

In the design-time dimension we distinguish (1) one-to-one mapping and (2) using a single-shared ontology:

- One-to-one mapping of ontologies. Mappings are created between pairs of ontologies. Problems with this approach arise when many such mappings need to be created, which is often the case in organizations where many different applications are in use. The complexity of the ontology mapping for the one-to-one approach is $O(n^2)$ where n is the number of ontologies. An example of the one-to-one approach is OBSERVER [Mena et al., 2000]. Figure 1 illustrates one-to-one mapping of ontologies. There exists a mapping between every pair of ontologies. In the worst case, these mappings are only one-way. This means that a single mapping can only translate from one model to another, not the other way around.

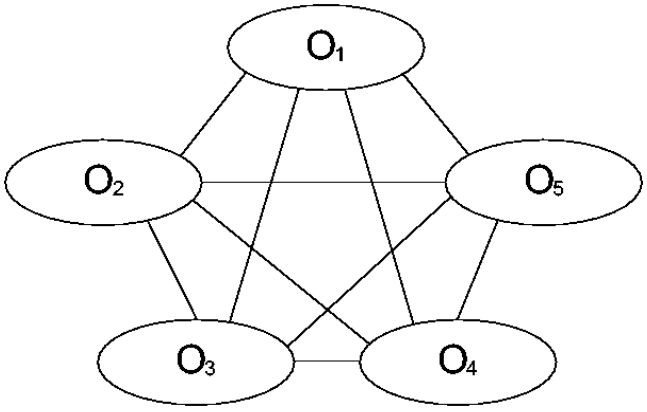

Figure 1. One-to-one mapping of ontologies

- Using a single-shared ontology (Figure 2). Drawbacks of using a single-shared ontology are similar to those of using any standard [Visser and Cui, 1998]. For example, it is hard to reach a consensus on a standard shared by many people (it is always a lengthy process), who use different terminologies for the same domain and a standard impedes changes in an organization (because evolution of standards suffers from the same problems as the development of standards). Examples of the single-shared ontology approach are MOMIS [Bergamaschi et al., 2001] and the Semantic Information Management [Schreiber, 2003].

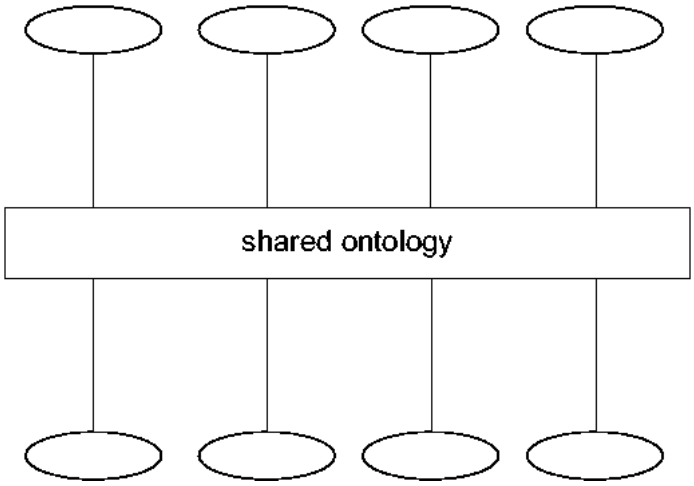

Figure 2. Single shared ontology

Within the paradigm of single-shared ontology mapping, we distinguish two forms:

- Removing the old local data models. All applications use the new global data model. A drawback of this approach is that applications depending on the local data models will break and have to be adapted to the now global model. Another drawback is the fact that groups in the organization can no longer maintain their own terminology; everybody will have to submit to the new global model [Uschold, 2000].
- Keeping the local data models and creating a mapping to the new global data model. Local models can remain in place; applications will not break because of the new global model. An advantage of this approach compared with the one-to-one mapping approach is that there is a smaller number of mappings that need to be defined and maintained for integrating a given number of data models. A drawback of this approach is that still old (possibly not so good) data models remain and mappings need to be maintained. They need to be updated with every update of the local model and with every update of the global model.

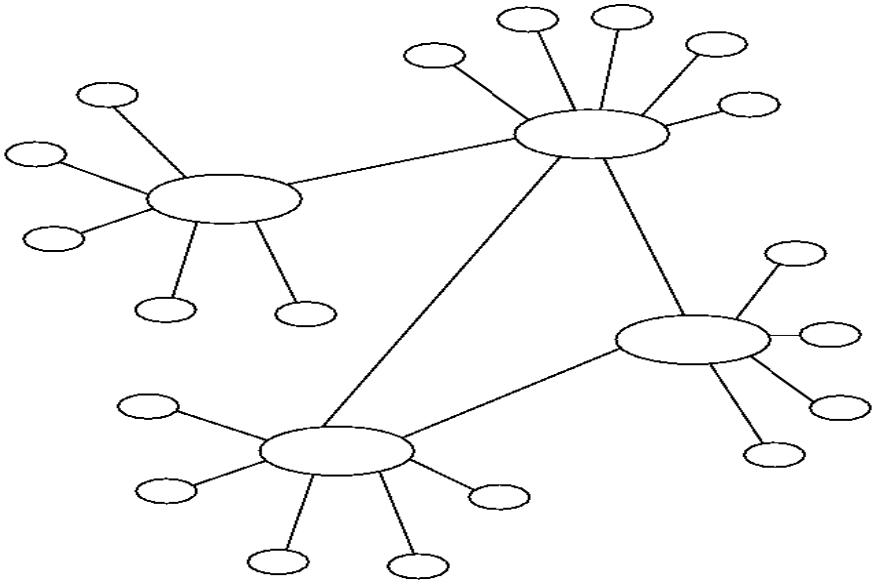

Figure 3. Ontology "islands": large ellipses depict locally global ontologies; small ellipses depict locally local ontologies

We can learn from the data integration systems, which provide services for query answering over distributed heterogeneous data sources. However, the current setting of these integration systems is inside the enterprise, which is still a more-or-less controlled area. On the Web, not much control over the use of ontologies can be expected and the global integration scenario is not expected to scale, because eventually different organizations will use different ontologies and will not want to commit to a new ontology. However, the one-to-one integration approach is also not expected to scale, because it would require the maintenance of too many mappings between ontologies. Therefore, we expect a hybrid approach will appear, where we have several "islands" around influential domain ontologies, where within the island there is a form of global integration; one ontology would be the global ontology of the islands and a number of local ontologies are mapped to this global ontology. Then, there would be mappings between the islands, as illustrated in Figure 3.

2.2 Application of Ontology Mapping

After having decided on the general architecture for your ontology integration system (one-to-one, single-shared, etc.), you need to decide how the ontology mediation components fit in the architecture.
The general mapping language which is developed in the Work Package 4 allows to specify both uni-directional and bidirectional mappings between ontologies. A uni-directional mapping can serve to transform data from the source ontology to the target

ontology and thus also to query the source ontology in terms of the target ontology. A bidirectional mapping can be used also to transform data from the target ontology to the source. Furthermore, it allows to perform arbitrary reasoning (e.g., subsumption) over both ontologies at the same time.
The mapping tools developed in the Work Package 4 have a bias towards uni-directional mappings, since these are more suited for efficient data transformation and query answering. Furthermore, the reasoner of choice, OntoBroker[6] [Fensel et al., 1998], supports only rule-type logical formulas which are inherently uni-directional (the body of a rule is used to derive the head).

2.3 Ontology merging

There exist several misconceptions on what the term "ontology merging" actually means. Intuitively the terms means "putting a number of ontologies together in some way". In this Section we aim to clarify what this actually means. In order to clarify the meaning we distinguish three different ontology merging scenarios.
As an illustration we take the simple case where two source ontologies *O1* and *O2* are merged into one target ontology, named *O3*. This is illustrated in Figure 4.

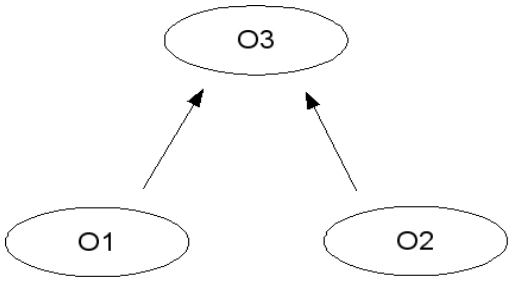

Figure 4. Basic ontology merging

There are two important questions with respect to the figure, namely "what happens to O1 and O2 after the merge?" and "what do the arrows mean?". There are essentially two possibilities as to what happens with O1 and O2 after the merge. Namely, either they remain or they disappear:

- In case O1 and O2 *remain* after the merge, all three ontologies can be used after the merge. This would be the case if (1) one does not have control over the availability of O1 and O2, (2) making O1 and O2 unavailable would break existing applications or (3) the merge is only temporary.
- In case O1 and O2 do not remain after the merge, the ontologies have been replaced and, depending on the way the merge is done, applications which use O1 and O2 might have to be updated to use the new ontology.

The question what the arrows in the Figure mean has a lot to do with the form of O3. In fact, the meaning of the arrows, together with O1 and O2, completely determines

6 http://www.ontoprise.de/content/e3/e27/index_eng.html

what O3 looks like. This brings us to the three basic ontology merging scenarios. We distinguish three basic scenarios:

1. O3 is simply the *union* of O1 and O2 (written as O3 = O1 ∪ O2) without taking any special care about mismatches or overlap between the source ontologies. This form of ontology merging has a number of characteristics:
   - If there is some conceptual overlap between O1 and O2, but there is still some difference in how it is written down, there will be redundancy in O3. For example, if O1 talks about *Cars* and O2 talks about *Automobiles*, O3 will talk about both.
   - If O1 and O2 use the same terminology, but with a slightly different meaning, inconsistencies might be introduced in O3. For example, according to some English ontology, a marriage may only be between a man and a woman, whereas according to some Spanish ontology, a marriage may be between any two people, but both ontologies have the restriction that a person may only be married to one other person. Now, according to some Arab ontology, this restriction does not hold.
   - If O1 and O2 cover different domains and use different terminology, the**re will be no redundancy and no** inconsistency in O3.
   - This kind of merging corresponds with the usual notion of ontology import, as it is defined in, for example, OWL and WSML. In a sense, no mediation between ontologies is performed, because the ontology merging in this scenario is just a simple concatenation of ontology elements.
2. O3 is obtained from the *union* of O1, O2, *and* a number of mapping rules M (O3 = O1 ∪ O2 ∪ M). These mapping rules would resolve conceptual overlap between O1 and O2. For example, in the case of *Cars* and *Automobiles*, the mapping rules would state that the concepts are equivalent. Note that such mapping rules could introduce inconsistencies, as mentioned in the second bullet in the previous scenario. Typically, ontologies are specified using different namespaces and thus the names in the ontologies would only rarely overlap. However, for example, a mapping rule stating the equivalence between *http://spain.com#Marriage* and *http://england.com#Marriage* could introduce an inconsistency.
   - This procedure is usually used when one wants to reason over multiple ontologies, but one does not want to replace the original ontologies. The merged ontology is created on the fly from existing ontologies and pre-specified mapping rules. This is the main scenario supported by Ontology Mediation in SEKT.
3. O3 is obtained from O1 and O2 using the following procedure: if an element in O1 overlaps with an element in O2, a new element in O3 is created which replaces the source elements from O1 and O2. Any element in O1 (or O2) which does not overlap with any element in O2 (or O1) is simply added to O3.
   - This merging scenario is typically geared towards replacing the original ontologies O1 and O2. However, it is not always possible to remove such ontologies and removal of such ontologies would break existing applications.

- Creating a merged ontology in this scenario requires all stakeholders for the ontologies O1 and O2 to agree on the merged ontology. Thus, this form of ontology merging can be seen as a special case of distributed ontology engineering, namely one with typically a large group of stakeholders (because this group is the merge of the stakeholders of O1 and O2) and with already some input ontologies which should be reused to a large extent.
- Some tools have been developed for this type of ontology merging, the most prominent being PROMPT [Noy & Musen, 2000]. However, it is arguable whether such a centralized tool specifically for this type of ontology merging is beneficial. We believe that it is more useful to see this kind of ontology merging, where the merged ontology replaces the source ontologies, as a special kind of ontology engineering.

3 Eliciting Requirements for Mediation from the Use Cases

In this section we derive requirements for mediation for each of the SEKT Case Studies. The section contains three subsections, one for each case study. Each subsection starts with a short description of the corresponding case study followed by two main parts: the description of the data sources for the corresponding case study and a part in which the requirements for ontology mediation are explicityly derived. The description of the data sources was provided by the partners responsible for each case study. The process of eliciting requirements for mediation is based on the analysis of the use cases and on the data source landscape described in the first part of each subsection.

3.1 SBS Case Study (Heterogeneous Groups in Media)

The objective of the Siemens / Siemens Business Services case study is to investigate and verify how semantically enabled technologies can improve the productivity of IT and business consultants.

Data Source Descriptions

The following are the data sources that were indicated by our partners to be used in this case study:

- **Intranet**

  Nature: web pages

  Structure: unstructured

  Size: 5000 WebSites

  Purpose: to provide information for IT Projects

  Content: Knowledge Base, business process support, requirements capture, searching and browsing, alerts, knowledge capture and reuse, knowledge sharing, expertise location.

Data format: html, MS Office, text, pdf, zip

API/data access: Web, CMS (SIX-CMS) System & Web Indices

- **Livelink / Knowledgemotion** (3 different instances)
  Nature: Document Management System, metadata in Oracle DB, generic files in EFS (extended file system)

  Structure: structured via folders and attributes (system attributes, custom attributes)

  Size: 1 TerraByte

  Purpose: to provide information for IT Projects

  Content: all working documents for the employees

  Data format: MS Office, text, PDF, html, xml, mpg, jpg, tif

  API/data access: Livelink API (API for Java, C, VB with complete Livelink functionality), XML Export/Import, Web-Service

Knowledgemotion is a document management system that has four underlying pillars, depicted in the figure below**:**

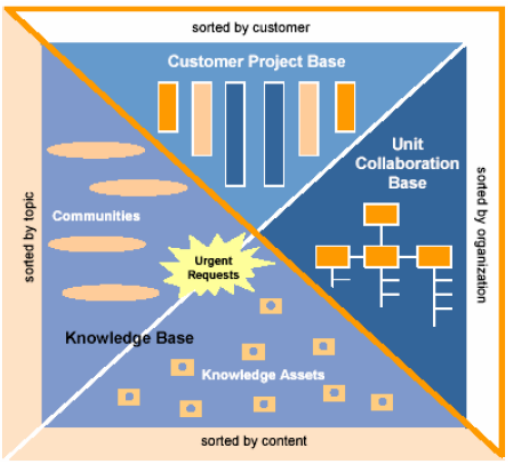

Figure 5. The four pillars of Knowledgemotion

- **Siemens Business Services Specification Information Ontology (Communities)**
  Size: approx. 50 MB

Purpose: searching, identifying and receiving answers and similar solutions and develop solution ideas using multi-perspective reasoning. Similarity identification, analogies, comparisons, identify idea attractors, reasoning.

Content: MS Office Files, Proposals, News, Offerings, Events, Project Reusables, References/Success Stories

- Find Reusables
- Upload Reusables

Requirements on Ontology Mediation

Some use cases for this case study are presented in [Zeilbeck et al., 2005]. These are:

1) *Use Case 1 – Solution Design for a Proposal – Phase S40*: a Proposal Manager / Project Manager Candidate has to propose a solution for the customers solution in sufficient detail, so that all requirements are covered and the time and the effort to be spent on implementation can be reliably estimated. This implies locating the relevant knowledge, which can be either in external or internal sources and knowledge sharing.
2) *Use Case 2 – Contribution of a document/Knowledge Asset* : users should be able to make knowledge–assets candidates available to their colleagues;
3) *Use Case 3 – Reuse Initiative* : replicating project results and outstanding successes to as many customers and with as many minimal effort as possible (reference selling)

Unfortunately, due to the summary description of the data sources provided by the partners from this case study and to the high-level description of the use cases (it was not clearly specified which data sources should be accessed in specific situations) the requirements for mediation derived for this case study are very vague. One requirement for mediation that was indicated by our partners and that is depicted by Figure 6, is the need for ontology mediation between the ontologies created on the basis of the individual's data sources (e.g. filesystem) and the shared ontologies.

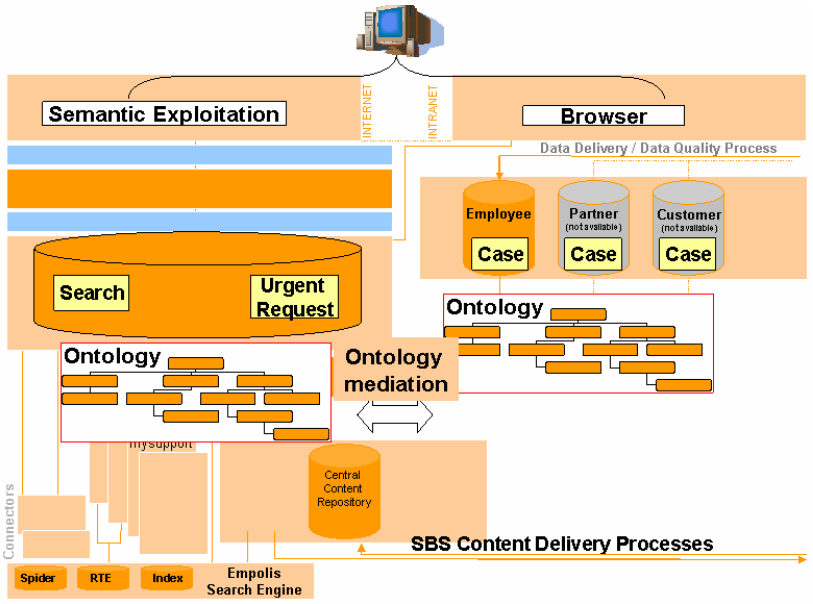

Figure 6. Siemens Business Services Infomodel (ontology approach)

Also, as the use cases clearly suggest, there is a need for searching in multiple repositories at the same time; thus, the ontologies corresponding to the different data sources will have to be aligned. It is not clear yet, if this will be done using one shared ontology, a one-to-one mapping or a hybrid approach. It was suggested by our partners that the interaction should take place on the top level ontology (SBS Ontology), so we assume that the global approach is considered in this phase.

3.2 Legal Case Study

The objective of this case study is to provide the young judges from Spain with a Semantic Web-enabled search system that helps them making decisions in their first destination.

Data Source Descriptions

Figure 7 presents the architecture for the case study introduced in [**Rodrigo et al., 2004**], which includes all the data sources that are used by this use case and the relationships between those. Two categories of knowledge are captured in these data sources:

i) *Expert Knowledge* – the judges expertise
ii) *Jurisprudence* - the existing body of law (rulings).

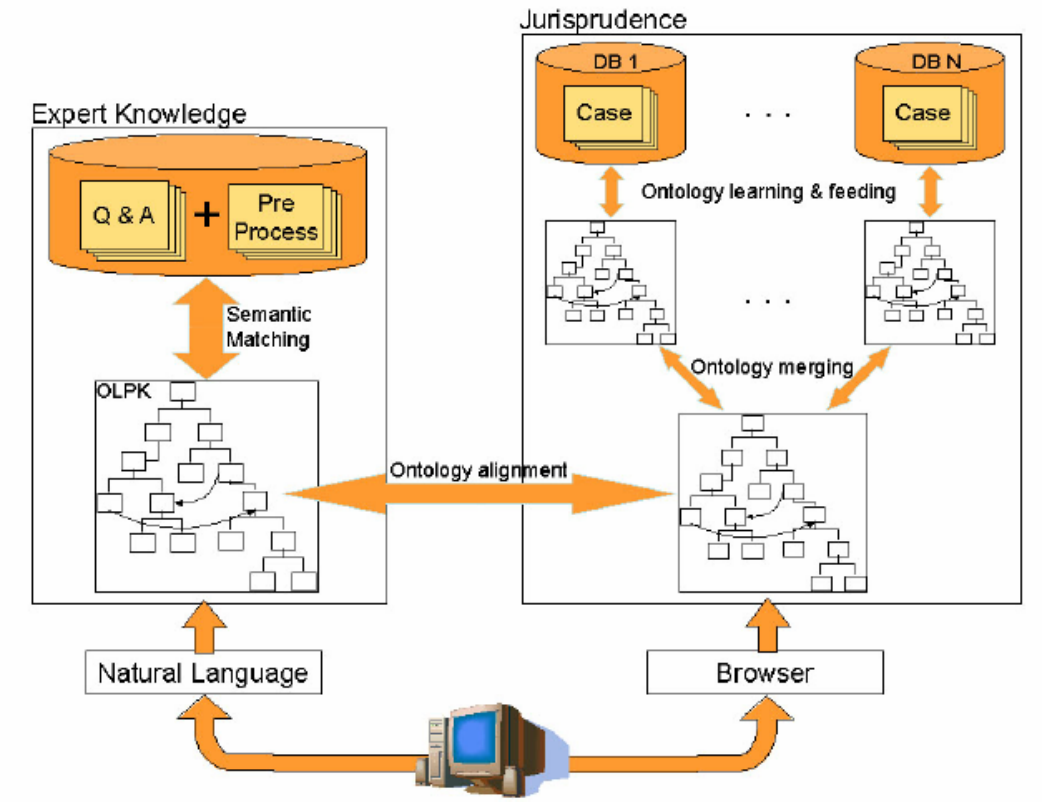

Figure 7. Legal Case Study Initial Architecture

Below is given a description of the data sources that capture the Expert Knowledge:

Q & A database:

Nature: question-answer pairs

Structure: unstructured

Size: approximately 800 question-answer pairs.

Purpose: to provide frequently asked questions and answers for young judges in their first destination in Spain.

Content: description of frequently recurring problems faced by judges in Spain, along with ways to deal with the problems.

Data format: text

API/data access: by ODBC

Legal Ontology: - Ontology of Judiciary Professional Knowledge (OLPK in the figure, OJPK further in the document)

Nature: ontology

Structure: structured

Size: nearly 50 concepts, 100 relations and more than 300 instances

Purpose: holds and structures the expert knowledge; allows the system to reply through the same set of basically related concepts that users (young judges) will have in mind in their consultations; it enables discovering hidden relationships between terms in user queries and terms in the QA database by providing a reference terminology for describing both: similarities between the questions posed by the user (in natural language) and the questions stored in the database are derived based on the semantic distance between the ontology concepts identified at those questions.

Content: background knowledge about the legal system in Spain; it models the different parts of legal processes, their agents and the roles that each agent plays in the different parts of those processes, the documents associated to different processes like complaints, certificates, etc. This knowledge is related with the experience of the judge's daily work.

Data format: OWL (motivated by the fact that the ontology had to be integrated in Proton Ontology, which at its turn is an OWL ontology).

The data sources that capture the Jurisprudence Knowledge proposed by the above architecture are:

- **Jurisprudence databases :**

  Nature: a diverse collection of existing jurisprudence documents stored in databases

  Structure: semi-structured

  Size: million of documents.

  Purpose: to provide explanations of the answers provided by the system based on the available jurisprudence.

  Content: facts, law and jurisprudence regarding past legal cases.

  Data format: There are several providers of databases, but the format of the databases is not known. It has been only achieved web access to them.

  API/data access: Web access, based on username and password.

- **Database ontologies :** *they are not created yet.*
  Nature: ontology

Structure: structured

Size: *not yet known*

Purpose: to have available a set of ontologies (one per database) reflecting the knowledge included in the databases; it is envisioned to use the technology of the WP1 to generate ontologies automatically from the content of the databases of cases. These ontologies are domain ontologies, for example, about gender violence, economic offence, etc., and they allow accessing the content of the databases using natural language techniques.

Content: formalized representations of the cases from the databases

Data format: OWL

- **Jurisprudence ontology:**
  Nature: ontology

  Structure: structured

  Size: *not yet known*

  Purpose: to have a general ontology covering all the jurisprudence knowledge included in the system, which will constitute a central point of the system reasoning.

  Content: the result of merging the ontologies learned from the databases.

  Data format: OWL

Requirements on Ontology Mediation

Six use cases were identified in **[Casanovas et al., 2005]** for the Legal Case Study:

1) *Question answering* - this is done using the information from the FAQ repository. OJPK is used for providing a common terminology in which user queries and question - answer pairs are translated in order to be matched.

2) *Answer explanation* - the user wants to get additional information to the answer to his question, in the form of an explanation. This is done by accessing the cases from the Jurisprudence Ontology that are related to the user question. This imposes the need of alignment between the OJPK and the Jurisprudence Ontology.

3) *FAQ Updating* – including new question-answer pairs in the FAQ repository in order to cover knowledge gaps. This may involve updating professional knowledge, in which case also the alignment with the jurisprudence has to be revised.

4) *Cases Updating* - including new cases in the databases. This implies revising the ontology databases and the Jurisprudence Ontology and further on, revising the mappings (the alignment) between the Jurisprudence Ontology and OJPK.

5) *Database Ontology Learning* – learning the domain ontologies – we will further discuss this issue below.

6) *Jurisprudence Ontology Learning* – that's the official name of this use case in **[Casanovas et al., 2005]**, but actually in the same document it is said that it is intended to create this ontology by merging the domain ontologies. – as with the previous use case we will elaborate on this below.

Thus, an explicit need for ontology mediation appeared in the use case 2) (alignment) and in the use case 6) (merging). We further analyse these use cases with respect to the mediation requirements.

Use case 2) requires *aligning* the Jurisprudence Ontology with the Ontology of Judiciary Professional Knowledge. In this way, queries on the Legal Ontology can be used directly for querying also the jurisprudence databases. There is some overlap between these two ontologies, which is an effect of the overlap between the knowledge from the FAQ repository and the Judgment knowledge captured by the Jurisprudence databases. Figure 8 describes the overlap between these two sources of knowledge: the *question* is related with the *case history* and the *answer* with the *decision grounds* and the *ruling*.

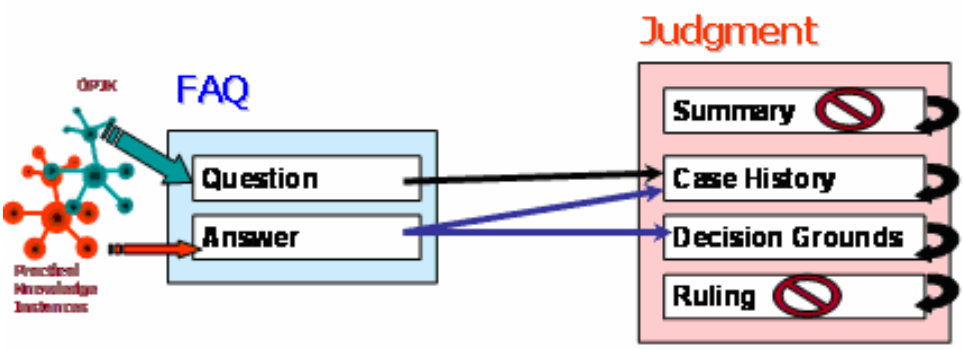

Figure 8. The relation between the knowledge from the FAQ and the Jurisprudence Knowledge

As Figure 8 suggests, and also, as it is explicitly specified in the answer to one of the questions from the appendix related with the mediation requirements for this case study (Question 3), the knowledge extracted from (contained in) the databases seems to follow a certain pattern, that of a judgement. The main parts of a judgment (and the structure that it is common to all providers) are: *the case history*, *the decision grounds* and *the ruling*. This leads us to the use cases 5) and 6), related with learning ontologies from the databases and merging them in the Jurisprudence Ontology. We assume that the ontologies learned from databases will differ only with respect to instances and will have the same schema corresponding to the structure of a judgment. But, in this case, the utility of these ontologies is questionable. A single schema

should be learned (or should be assumed from the beginning) and different instance stores could be used for storing the data from the knowledge bases according to that schema. The Jurisprudence Ontology will consist of the learned schema and links to these instance stores. This would rule out the need for ontology merging from use case 6).

We can see that use cases 3) and 4) also impose the need for alignment revision in case one or the other of the ontologies is updated.

In conclusion to this analysis, we acknowledge the need for alignment between OJPK and the Jurisprudence Ontology and we propose the following simplification of the initial architecture in which the ontologies learned from each jurisprudence databases are eliminated, the Jurisprudence Ontology being directly created (populated) from these databases:

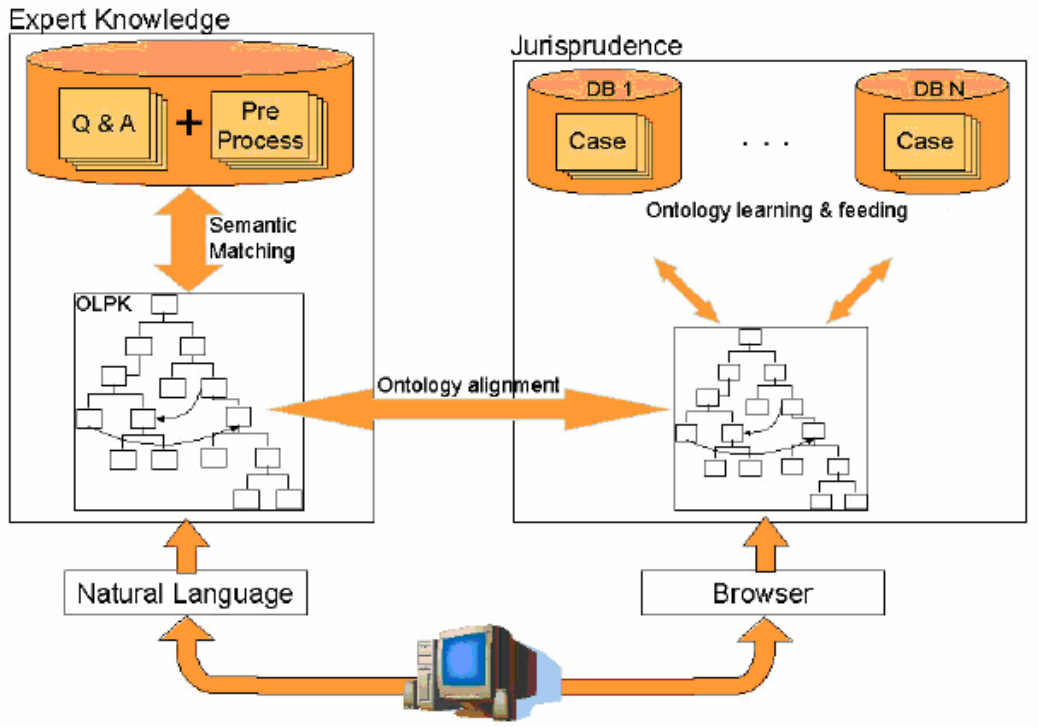

Figure 9. Proposed architecture for the Legal Case Study

3.3 Digital Library

The BT digital library offers users the capability to search and browse an extensive on-line collection of technical and business journals, conference proceedings, and electronic books. The Digital Library provides access to over 5 million records from the Inspec and ABI/INFORM databases. BT subscribes to approximately 1000 on-line publications, giving users access to the full-text of over 900,000 scientific and business articles and papers **[**Alsmeyer et al., 2005**]**.

Currently, the user queries an index via a web interface. The majority of user queries are simple one or two keyword type queries. Little use is made of the '+' prefix (term

must occur) or keyword truncation with the wild-card (*). By default the search is performed against the title, abstract and subject index fields, but searches can also be specified in Inspec and ABI Inform for authors, journal titles, publication year and document title, using the au=, so=, py=, and ti= operators (not very user friendly). Results are shown as a simple list. The user has the options to refine/filter their search based on the descriptors (controlled indexing terms), date of publication, company (if available) and type of article (Inspec only).

The aim of this case study is to have a semantically-enabled digital library based on the same data sources that will allow the user to perform semantic search, to access semantically enabled public and private information spaces, to annotate and share digital library documents, etc. Users may also have personal search agents, and may choose to share web pages in a knoweledge sharing application (which will be integrated with the Digital Library).

Data Source Descriptions

Figure 10 depicts the data sources and the relationships between those that were initially envisioned to be used by this case study :

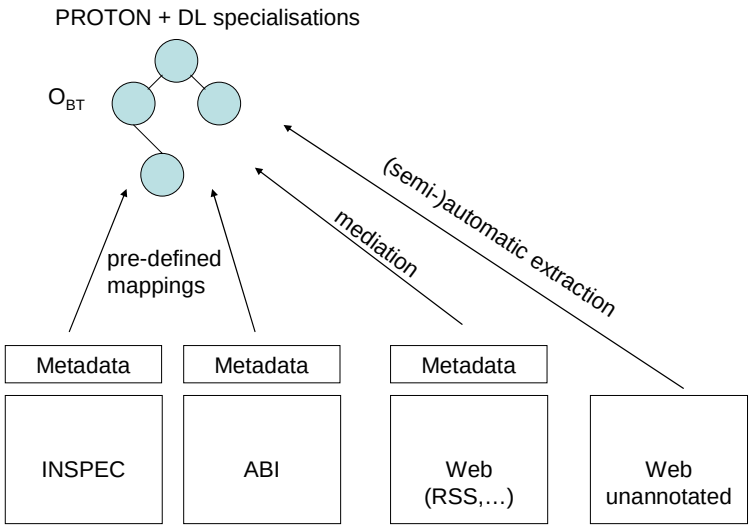

Figure 10. Data sources for the Digital Library Case Study

A more detailed description of data sources is given below:

- **ABI**

  Nature: bibliographic database

  Structure: structured

Size: > 2.000.000 records

Purpose: to provide bibliographic information.

Content: bibliographic information *from business and management journals*

Data format: MARC

API/data access: currently, a proprietary index built from information contained in the ABI database. The index is accessed via a CGI interface. The overall architecture for the BT Digital Library will be an integrated KAON2/SIP platform. The architecture is still under consideration, being discussed between the various technical workpackages.

- **Inspec**

  Nature: bibliographic database

  Structure: structured

  Size: > 2.000.000 records

  Purpose: To provide bibliographic information.

  Content: bibliographic information *from technical journals and conferences*

  Data format: MARC

  API/data access: currently, a proprietary index built from information contained in the Inspec database. The index is accessed via a CGI interface. The overall architecture for the BT Digital Library will be an integrated KAON2/SIP platform. The architecture is still under consideration, being discussed between the various technical workpackages.

- **Web database**

  Nature: a collection of references to web pages. Built Web index from focused crawlers (and from information, e.g. Web pages, shared using the WP5 knowledge sharing tools).

  Structure: semi-structured (attempt to extract metadata where possible, e.g. create a summary, named entities, author, and classify against the BT digital library ontology).

  Size: still to be determined. A focused crawler could be provided for each 'Information space' (there are approximately 200 information spaces). Each crawler could provide data for 10s of 1000s of pages, let's say 1000 pages per information space, giving a total of 200000 records).

  Purpose: to collect references to web pages of interest to the user, user group/project, domain expert.

Content: references to web pages along (Web unannotated in the figure) with user annotations (and summary extracted from Web page, and other extracted data such as named entities).

Data format: HTML, PDF, etc.

API/data access: currently, a proprietary index built from information contained in the Web database. The index is accessed via a CGI interface. The overall architecture for the BT Digital Library will be an integrated KAON2/SIP platform. The architecture is still under consideration, being discussed between the various technical workpackages.

- **Web RSS –** *not created yet*

  Structure: structured

  Size: unknown

  Purpose: keeping users up-to-date with information that is relevant to their work

  Content: collection of newsfeeds that are relevant for users

  Data format: RSS (different variants), Atom

  API/data access: news agregators

- **Library Ontology**
  Nature: ontology

  Structure: structured

  Size: to be estimated (approx 5KB per record – based on preliminary RDF ontology of 4000+ Inspec records).

  Purpose: aim to improve user's ability to find relevant information in the library, either individually, or as a group of users.

  Content: concepts related to bibliographic information

Requirements on Ontology Mediation

The following use cases were identified in [Alsmeyer et al., 2005] for the Digital Library Case Study :

1) End user use case 1: *Search and browse* - semantically-enabled searching using a search context based on the interaction of the users with the digital library and other information sources and on their profile. Users will be able to

browse a more limited view of the digital library ontology based on their particular interests.

2) End user use case 2: *Information spaces* - serve specific categories of users by having attached queries based on topics from the digital library ontology. Users are given functions to subscribe to selected sub-topics within the domain of a public information space. Users will have the possibility to set private information spaces.

3) End user use case 3: *Knowledge sharing* - a user can annotate web pages and library content for their own benefit. Knowledge sharing emerges as a side effect of this annotation process, the annotations of the users being done using the same ontology.

4) End user use case 4: *Expertise location* – identification of experts by monitoring the level of difficulty of what people are reading in specific topic areas - implies classifying the publications in the digital library according to a level of reader difficulty

5) End user use case 5: *Personal search agent* - this use case describes a semantic search agent that can be configured to query the digital library, WWW or Intranet based search engines for information on behalf of the user.

6) End user use case 6: *Personal information-based content delivery*- keeping users up-to-date with information that is relevant to their work (events within their own organization as well as breaking news on a client they are about to visit). The available information should be analyzed in context with a user's diary events.

7) End user use case 7: *Profile management* – the applications will make profile recommendations to the users, as they interact with the system, suggest membership of relevant information spaces based on the accessed content, etc.

8) End user use case 8: *Notification* – digital library tools will send notifications to users (under user control).

9) Domain expert use case 1: *Focused crawling* - extending the content of the library with relevant information from the WWW by selectively retrieving and annotating information from WWW sources.

10) Domain expert use case 2: *Information space configuration tool* – similar to the process that a user follows when creating a private information space.

11) Administrator use case 1: *Instance update tool* - the instance data will be updated with content from: a) successive weekly updates of the databases, b) from documents discovered by automated search agents and crawlers, and c) from electronic documents added by people using the knowledge sharing tools;

12) Domain expert and system administrator use case 1: *Ontology extension and merging tool* - the digital library will be capable of (semi-)automatically

learning an ontology (or extensions to an ontology) from an additional corpora of documents; managing updates to the digital library topic ontology: a need to merge the evolving digital library topic ontology with new topic/terms in the latest thesaurus.

In most of the previous use cases, knowledge from different data sources (ABI, Inspec, WWW, RSS) must be accessed in a uniform way for answering user queries, helping users to subscribe to information spaces, to annotate/share their knowledge, etc. The BT Ontology (O_{BT} in the figure) is created manually (it models people, articles, roles, organisations, etc), and agreed in collaboration between SEKT work packages. For this, PROTON will be extended with BT domain-specific classes/properties. Instances of objects (e.g. articles) will be extracted from the current ABI and Inspec databases (and updated when weekly Inspec/ABI updates are received). For making this possible, the content of the databases must be aligned with the content of the BT Ontology. A separate 'news' ontology will be created and populated using the newsfeeds. Items from the news ontology are mapped to O_{BT}, thereby allowing people who are searching the digital library to also search the 'news' index.

Based on the use cases, the following scenarios were derived where mediation can be used:

Scenarios where mediation could be used:

1. The initial Digital Library Ontology will have to be aligned with the information contained in the purchased databases (ABI and Inspec) in order to be able to populate its instance store using the content of those databases. For example, the 'topics' from the ontology could be aligned with the Inspec and ABI Thesaurus 'preferred terms'. When the alignment between the two schema is in place, the set of instances of the Digital Library Ontology can be populated by instance transformation techniques using the content of the databases.
The Digital Library Ontology will have to be updated when: a) regular ABI/Inspec updates are received, b) as focussed crawler software retrieves relevant content (and builds an index). As a result of b) new topic areas may be identified, and there will be a need to extend the topics of the Digital Library Ontology, e.g. extending the set of instances of the "Topic" concept from the Digital Library Ontology. On subsequent (yearly) updates of the ABI/Inspec Thesaurus, some of the topics created as a result of b) may now align with new 'preferred terms'. Ontology mediation will therefore be required between the new topics derived as a result of b) and the equivalent new topics defined by preferred terms in the ABI/Inspec thesaurus.

2. Newsfeeds are intended to be collected (see use case 6). An ontology will be created for annotating the news (called News Ontology in Figure 11). This ontology will have to be populated with the content of the newsfeeds. We envision that some mappings will exist between this ontology and the syndication formats (RSS, Atom, etc.), but some natural language techniques might also have to be used for wrapping the content of the newsfeeds to this ontology, due to the limited structure of the syndication formats. At its turn, this ontology will have to be aligned with the BT DL in order to make possible the search using the BT DL search interface.

The next picture depicts the new relationships identified between the data sources:

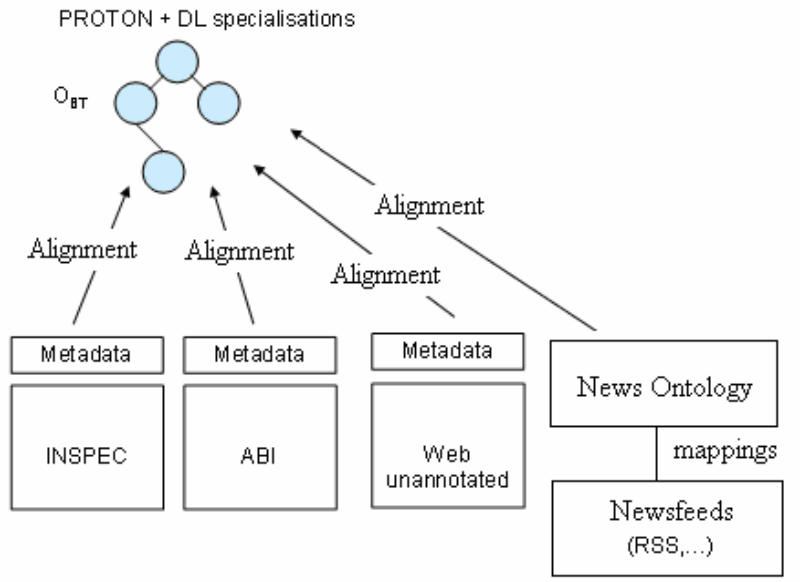

Figure 11. Proposed architecture for the Digital Library Case Study

4. Conclusions

This deliverable perused the descriptions of the SEKT case studies with the purpose of identifying the mediation requirements for each of them. Each partner was asked to provide a description of the data sources used by the case study and to answer some clarifying questions regarding the relationships between these data sources. These contributions together with the descriptions of the use cases from the corresponding deliverables [Alsmeyer et al., 2005, Casanovas et al., 2005, Zeilbeck et al., 2005] enabled us to derive such requirements. While the Siemens case study started later than the other two case studies and the relationships between the data sources are not precisely defined yet, for the other two case studies we were able to derive precise requirements and to suggest a new data source landscape The need of ontology alignment was prevalent in all the case studies, thus the tools developed by WP4 should focus on providing this functionality. Both the case studies and WP4, the SEKT work package responsible for mediation, will benefit from the analysis performed in this deliverable.

Appendix A: Questions and Answers Regarding Mediation Requirements for the SBS Case Study

This appendix contains a set of questions regarding the relationships between the data sources and the mediation requirements for the Siemens Business Services Case Study together with the answers we received from our partners.

(1) **Q**: How will the mappings between the personal and the shared ontologies be exploited? Can individuals use the mappings to query data which is annotated using shared or other people's ontologies?
A: It is intended to match different ontologies (personal, public): Search the full text of your email, files, viewed web pages, and chats. Since you can easily search information on your computer, you don't need to worry about organizing your files, email, or bookmarks. You can just do a quick search for what you remember seeing, instead of having to remember exactly what file, email, or web page had that information, and where that item is now located on your computer.- so, one individual will search for (use) information located only in the personal ontology and not in the shared ontology or in other individual ontology.

(2) **Q**: Is there any overlap (redundancy) in data contained in different data sources (e.g. two data sources, which both contain data about persons)?
A: normally, there are specific data sources for singular purpose; e.g. central corporate directory or partner database; there is a temporary redundancy of information in intranet and knowledgemotion

(3) **Q**: Are there any current or envisioned applications that use more than one data source? And if so, which data sources do they use in combination?
A: knowledgemotion as multicluster-architecture (3 large databases getting accessed by diverse webserver(frontends), having search environment for each database at the moment SIX-CMS for Intranet Content

(4) **Q**: What are the current strategies for dealing with redundancy?
A: To avoid redundancy different workgroups are set up to define the scope of location for information (where to store some data, either on intranet pages or in document management systems like knowledgemotion); a rule is defined to put all officially announceable information on webpage.

(5) **Q**: How are shared ontologies constructed? Are they intended to be a merge of the personal ontologies or are they engineered from scratch?
A: It is supposed to merge personal ontologies (from scratch).

(6) **Q**: Are there any other requirements with respect to differences between data sources and ontologies and with respect to querying across data sources in this case study?
A: The main barrier to electronic commerce lies in the need for applications to meaningfully share information, not in the reliability or security of the Internet. This is due to the variety of enterprise collaboration systems deployed by businesses and the way these systems are variously configured and used. It is the central goal of an Ontology to solve this problem.

Appendix B: Questions and Answers Regarding Mediation Requirements for the Legal Case Study

This appendix contains a set of questions regarding the relationships between the data sources and the mediation requirements for the Legal Case Study together with the answers we received from our partners.

(1) Q: How is the merged ontology for the jurisprudence databases going to be constructed? Will it be constructed solely on the basis of the jurisprudence databases or will it be constructed independently and later mapped to the jurisprudence databases?

A: We don't construct independently the merging ontology, but we have considered the definition of [Noy and Musen, 1999] about Ontology Merging: "Ontology merging proposes to generate a unique ontology from the original ontologies". We are thinking on the generation of the merge ontology from domain ontologies.

(2) Q: Why is there a separate merged ontology for the jurisprudence? Why not use the Legal Ontology directly and map the individual domain ontologies to this ontology?

A: In functional terms, it does not make a difference to have a mapping between the two "big" ontologies or several mappings between the legal ontology and each of the ontologies extracted from the jurisprudence databases, this is, the result of the system would be the same in both cases. However, in design time it seemed conceptually clearer to map two ontologies that are "similar" in size.

(3) Q: How do you plan to "ontologize" the content of databases of cases? For having a different ontology for each jurisprudence database it would be necessary to take into account each database peculiarity. The format of a database is such a specific feature. Are there any specificities of the databases that are considered during the learning phase such that different ontologies (schemas, not just instances) are created? Or, the "merged" jurisprudence ontology is constructed independently and the "ontologies" corresponding to the databases are just different instance stores of the "merged" ontology?

A: We don't plan to access directly to these databases, using ODBC or other kind of connection. We plan to use the web interface of these databases. Besides, the judgments (that are our aim) have a similar structure in all databases (independently of the database provider). The main parts of a judgment (and the structure that it is common to all providers) are: *the case history*, *the decision grounds* and *the ruling*. These sections are the sections that have the relevant information for the matching between question - answer and judgments. It has not been created yet any specification of the databases ontologies.

(4) How will the ontologies learned from the jurisprudence databases be related to the databases themselves?
It is not defined yet, but we are thinking on solutions based on the definition of attributes like "Documental Reference".

Appendix C: Questions and Answers Regarding Mediation Requirements for the Digital Library Case Study

This appendix contains a set of questions regarding the relationships between the data sources and the mediation requirements for the Digital Library Case Study together with the answers we received from our partners.

(1) Q: How are queries issued to the ABI and Inspec databases and how are results returned?

A: Currently, the user queries an index via a web interface. The majority of user queries are simple 1 or 2 keyword type queries. Little use is made of the '+' prefix (term must occur) or keyword truncation with the wild-card (*). By default the search is performed against the title, abstract and subject index fields, but searches can also be specified in Inspec and ABI Inform for authors, journal titles, publication year and document title, using the au=, so=, py=, and ti= operators (not very user friendly). Results are shown as a simple list. The user has the options to refine/filter their search based on the descriptors (controlled indexing terms), date of publication, company (if available) and type of article (Inspec only).

(2) Q: How big is the overlap between the four major data sources? In other words, is redundancy in query answers likely to occur?

A: There is minimal overlap between ABI and Inspec, less than 5% because they address different areas: ABI is business oriented, whilst Inspec is scientific. Both of these databases contain articles that may be found on the web, either in the web version of a journal included in the database or in a preprint on an author's site. It is conceivable that these articles will be included in the web database and introduce a degree of redundancy.

Bibliography and references

[Alsmeyer et al., 2005] David Alsmeyer, Allyson Cheung, Michael Engler, Nick Kings, Ian Thurlow, Paul Warren: *D11.2.1 BT Digital Library Scenarios* SEKT deliverable D11.2.1, 2005.

[Bergamaschi et al., 2001] Sonia Bergamaschi, Silvana Castano, Maurizio Vincini, and Domenico Beneventano. Semantic integration of heterogeneous information sources. Special Issue on Intelligent Information Integration, Data & Knowledge Engineering, 36(1):215-249, 2001.

[de Bruijn, 2004] Jos de Bruijn: Semantic Integration of Disparate Data Sources in the COG Project, Proceedings of the 6th International Conference on Enterprise Information Systems (ICEIS2004), Porto, Portugal, 2004.

[Casanovas et al., 2005] Pompeu Casanovas, Marta Poblet, Núria Casellas, Joan-Josep Vallbé, Francisco Ramos, Richard Benjamins, Mercedes Blázquez, Luis Rodrigo, Jesús Contreras, Jesús Gorroñogoitia Cruz : *D10.2.1 Legal Scenario –Case Study-Intelligent Integrated Decision Support for Legal Professionals* SEKT deliverable D10.2.1, 2005.

[Ehrig et al., 2004] M. Ehrig, T. Gabel, P. Haase, Y. Sure, C. Tempich, and J. Voelker. *Data manual – Initial Version*. SEKT informal deliverable D7.1.1.b, 2004.

[Fensel et al., 1998] D. Fensel, S. Decker, M. Erdmann, and R. Studer: "*Ontobroker in a Nutshell*" (short paper). In C. Nikolaou et al. (eds.), Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science, LNCS 1513, Springer-Verlag Berlin, 1998

[**Mena et al., 2000**] Eduardo Mena, Arantza Illarramendi, Vipul Kashyap, and Amit P. Sheth. OBSERVER: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. Distributed and Parallel Databases, 8(2):223-271, 2000.

[Noy & Musen, 1999] Natalya F. Noy and Mark A. Musen. Smart: Automated support for ontology merging and alignment. Technical Report SMI-1999-0813, Stanford Medical Informatics, 1999.

[Noy & Musen 2000] Natalya F. Noy and Mark A. Musen. Prompt: Algorithm and tool for automated ontology merging and alignment. In Proc. 17th Natl. Conf. On Artificial Intelligence (AAAI2000), Austin, Texas, USA, July/August 2000.

[Preece et al., 2001] Alun D. Preece, Kit-Ying Hui, W. A. Gray, Philippe Marti, Trevor J. M. BenchCapon, Zhan Cui, and Dean Jones. Kraft: An agent architecture for knowledge fusion. International Journal of Cooperative Information Systems, 10(1-2):171-195, September 2001.

[**Rodrigo et al., 2004**] Luis Rodrigo, Mercedes Blázquez, Pompeu Casanovas, Marta Poblet: *D10.1.1 Legal Case Study Before Analysis* SEKT deliverable D10.1.1, 2004.

[Visser & Cui, 1998] Pepijn R. S. Visser and Zhan Cui. On accepting heterogeneous ontologies in distributed architectures. In Proceedings of the ECAI98 workshop on applications of ontologies and problem-solving methods, Brighton, UK, 1998.

[Schreiber, 2003] Z. Schreiber. Semantic information management: Solving the enterprise data problem. To be found on the http://www.unicorn.com/ website, 2003.

[**Uschold, 2000**] Mike Uschold. Creating, integrating, and maintaining local and global ontologies. In Proceedings of the First Workshop on Ontology Learning (OL-2000) in conjunction with the 14th European Conference on Artificial Intelligence (ECAI-2000), Berlin, Germany, August 2000.

[Zeilbeck et al., 2005] Roland Zeilbeck, Mark Siebert, Markus Schwemmle, and Dirk Ramhorst: *D9.1.1 Siemens Case Study - Initial Analysis and Use Case Description.* SEKT deliverable D9.1.1, 2005.