

Final Report

Universität Karlsruhe
Vrije Universiteit Amsterdam
Meta4 Spain S.A.
empolis UK ltd.
empolis Polska Sp. z o. o.
Fundación IBIT
Dresdner Bank AG

Executive Summary.

SWAP EU IST-2001-34103 Final Report

This report presents the overall achievements and learned lessons of the SWAP project. Further, it summarizes dissemination and exploitation activities. An overview of generated deliverables, publications, and two system descriptions are listed in the annex of this document.

Document Id. SWAP/2004/final/v1.0
Project SWAP EU IST-2001-34103
Date November 30, 2004
Distribution Restricted

SWAP 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-2001-34103. The partners in this project are: Institute AIFB / University of Karlsruhe (coordinator, Germany), Vrije Universiteit Amsterdam VUA (Netherlands), Meta4 (Spain), empolis UK ltd. (UK), empolis Polska (Poland), and IBIT (Spain).

University of Karlsruhe

Institute AIFB
Englerstr. 28
D-76128 Karlsruhe
Germany
Tel: +49 721 608 7362, Fax: +49 721 693717
Contactperson: Marc Ehrig
E-mail: ehrig@aifb.uni-karlsruhe.de

Meta4 Spain S.A.

Rozabella, 8 Centro Europa Empresarial
28230 Las Rozas (Madrid)
Spain
Tel: +34 91 634 85 00, Fax: +34 91 634 86 86
Contactperson: Emilio Martin Cros
E-mail: emiliom@meta4.com

empolis Polska Sp. z o. o.

Plocka 5a
01-231 Warsaw
Poland
Tel: +48 22 535 88 06, Fax: +48 22 535 88 14
Contactperson: Mariusz Olko
E-mail: mariusz.olko@empolis.pl

Vrije Universiteit Amsterdam (VUA)

Division of Mathematics and Informatics W&I
De Boelelaan 1081a
1081 HV Amsterdam
The Netherlands
Tel: +31 20 4447786, Fax: +31 20 4447653
Contactperson: Hans Akkermans
E-mail: hansakkermans@cs.vu.nl

empolis UK ltd.

Unit B, The Dorcan Complex, Faraday Road
SN3 5HQ Swindon
United Kingdom
Tel: +44 1793 485465, Fax: +44 1793 485451
Contactperson: Justine Taylor
E-mail: jt@empolis.co.uk

Fundación IBIT

Reverend Francesc Sitjar 1
07010 Palma de Mallorca
Spain
Tel: +34 971 177 271, Fax: +34 971 177 279
Contactperson: Esteve Lladó Martí
E-mail: esteve@ibit.org

Changes

Version	Date	Author	Changes
0.1	22.11.04	Marc	document setup
0.2	26.11.04	Marc	integrated version
1.0	30.11.04	Marc	final version

Contents

1	Introduction	3
2	Overall Achievements	4
2.1	Conceptual Analysis and Methodology	4
2.2	Conceptual Design	8
2.3	Methods	10
2.4	Tool Development	13
2.5	Integration	16
3	Lessons Learnt from the Use Cases	19
3.1	IBIT Case Study	19
3.2	Bibliographic Case Study	26
4	Dissemination	37
5	Exploitation	39
6	Summary	42
A		45
A.1	Deliverables	46
A.2	Publications	48
A.3	XAROP — A Report on Introducing a Decentralized Semantics-based Knowledge Sharing Application	55
A.4	Bibster — A Semantics-Based Bibliographic Peer-to-Peer System	67

Acknowledgement

We would like to thank our project officers Brian Macklin and Hans-Georg Storck. They always helped us on smoothly administrating an IST project. Also they were very interested and supportive in regard to the topic of the project. We also would like to mention our reviewers Javier Garcia Sedano, Paul Ormerod, Dave Robertson, and Carles Sierra. They had to join the review meetings and had to read thousands of pages. In return, they provided as with many valuable ideas and criticisms that help us in improving the performance of the project.

Chapter 1

Introduction

This document marks the final report of the project Semantic Web and Peer-to-Peer (SWAP) funded by the European Union. With this report we reflect on the whole project once again and summarize the key results of the project.

We will stress the main achievements of each work-package, ranging from analysis, architectural considerations, developed methods, and implemented tools, to integration aspects. Further, we reflect on the lessons learned from the two case studies and the corresponding systems. Dissemination and exploitation activities, which took place during and will continue after the project, are presented in the subsequent section. We conclude with an overall summary of the project referring to the originally defined goals. The annex finally lists created deliverables, which are available via the project website, as well as the SWAP-related publications. Two comprehensive system descriptions of the main applications Xarop and Bibster conclude the annex.

Chapter 2

Overall Achievements

Each work-package has achieved many individual results, of which the most important ones are summarized in this chapter.

2.1 Conceptual Analysis and Methodology

2.1.1 State of the Art

As a first step in the project, the state of the art with respect to relevant technologies was analyzed and requirements for the use of technologies as foundation for the developments in SWAP were formulated.

P2P systems

We analyzed existing P2P solutions and identified aspects of these systems that are key features with respect to the goals of the project:

Interoperability Interoperability means the easiness of consistency across platforms, applications and programming languages. The deliverable shows that most systems have standardized communication protocols and languages. Many implementations use the JXTA protocols for connecting different systems due to its ease of use. Therefore JXTA has been chosen as communication protocol for the SWAP platform.

Scalability Due to the fact that we only use a 'real' P2P system, which means that there is no hierarchy in capabilities, scalability was an issue according to network usage (nr of messages needed to fulfill a query). From all the systems which are investigated in the deliverable only two use local semantic matching algorithms for routing messages. Note that this deliverable has been written one and half years ago, and due

to this popular research area, other systems that do semantic routing (e.g. PSearch) have come up in the meantime.

Security The Dresdner Bank project partner had security requirements, so therefore also this characteristic was investigated in the analysis of the different systems. The deliverable concludes that only the two big approaches, JADE and JXTA, address the security issue.

Privacy Further, the IBIT partners needed some privacy functionality in their application. Therefore this characteristic is important. To provide privacy, JXTA and JADE allow the use of router peers that maintain a list of pseudonyms accompanied with the real address. Empolis has implemented its own privacy method.

Both security and privacy were key issues dealt with in the implementation of Xarop.

Ontology-based KM solutions

One of the most significant application areas for Ontologies is knowledge management. Nevertheless, the ontology and the KM communities are developing in parallel, leaving potential synergies unexploited, despite a match between needs and potential solutions. Without bringing these communities closer there is a feeling that the ontology field would be in clear danger of producing tools and techniques that prove to be a mismatch to real needs, while the KM community would miss out on the opportunities presented by the appropriate use of ontologies. Ongoing and future research projects which build on existing technology and improve on it in the name of knowledge management should become aware of the issues surrounding the state-of-the-art. A more thorough and sober approach to ontology-based KM will hopefully lead to the acceptance of ontologies as tools that are no cure for all, but may improve knowledge processes to a considerable degree when applied appropriately. As a response to this observation, the developments in SWAP were driven by the needs of the knowledge management applications rather than by ontological principles. We see this as a strength rather than a weakness of the approach.

2.1.2 Combination of Technologies

A main question in the project dealt with the possible fruitful combination of principles and technologies in the areas of ontologies, P2P systems and knowledge management. This question was studied in the deliverables that are summarized below.

Ontologies for P2P and P2P for Knowledge Management

The purpose of D1.3 was to analyze how ontology technology can help to employ the full potential of the P2P paradigm for knowledge management. It is an analysis document

that centers on the “what” and not the “how”; in other words, it is not an attempt to find specific solutions, rather it presents the current shortcomings of the P2P systems for knowledge management and attempts to find the contributions that ontology technology can make to support and strengthen them. Based on the analysis we conclude that:

- Ontologies represent the basis for shared understanding in a group of individuals.
- P2P infrastructures are highly recommended for implementing decentralized applications.

Their flexibility also make it possible to cover many of the requirements of distributed KM. For the use case of a search for knowledge, ontologies offer the ability to provide an interpretation context for any query or results, as well as to solve the problem related to the concepts of synonyms and homonyms. They also allow us to define different fuzzy levels for knowledge other than simple documents or files. The deliverable states that the technology must be adapted to be fully useful in these distributed environments. The primary aspects that could be considered as areas of study are:

- automatic maintenance of the ontologies in each knowledge space,
- automatic or semiautomatic support tools for mapping among ontologies to facilitate the management and integration of multiple ontologies with no cost for the end user,
- automatic detection of similar concepts among ontologies from different areas to build knowledge communities, publication of the knowledge from one group to other groups as well as the detection of new knowledge.

Additionally, the inference techniques must be adapted to achieve more reliable results from multiple ontologies with overlapping, incomplete, or insufficient concepts, and to deduce new knowledge. Clearly these latter techniques are very difficult to develop and therefor it would be too optimistic to hope that these could be developed in the scope of the SWAP project.

P2P for Ontologies

In deliverable D1.4 we examined what peer-to-peer can offer to ontologies. The objective of the deliverable was not to provide solutions to particular problems but to provide an overview in which directions current and future research can go. We use the Strengths - Weaknesses - Opportunities - Threads (SWOT) analysis method to identify directions which are most promising. The SWOT method requires to examine the strength and weaknesses of a given technology and to define the opportunities and threads which arise from the combination with another technology. Consequently we identify the strength and weaknesses of peer-to-peer systems. We contrast each challenge with the strength of

peer-to-peer systems and develop the opportunities arising from the combination of the both technologies. In particular we look into the research areas concerned with ontology engineering methodologies, ontology reconciliation, ontology learning and instance generation. We found as strengths of peer-to-peer systems Decentralization, Distribution of Content, Redundancy, Information Sharing, Zero Administration, Collective Assembly of a Database, Seamless Communication and Connectivity, Scalability and Distributed Computation while weaknesses are Unreliability, Security, Free Riders, Resource Allocation and Quality. Ontology reconciliation can profit from the peer-to-peer paradigm. Feedback from user queries and the sharing of mappings, viz. the redundancy of mappings between ontologies, can help to automate the reconciliation task. peer-to-peer systems will also gain from pure ontology reconciliation research. More vague is the impact on Ontology Learning. A major source for existing and new algorithms could be the feedback available in a peer-to-peer network through its ability to collectively build a database. However, tests must be performed to check whether users are really willing to provide the feedback.

2.1.3 SWAP Methodology

Based on the Analysis of the relation between the relevant technologies, the project developed a methodology for applying peer-to-peer and ontology-based methods in developing distributed knowledge management. The methodology answers the meta-level questions of when and how to apply the technology developed in the project. This methodology is therefore a crucial component of this technology-driven system: with appropriate scoping and support, we can establish the correct positioning and appropriate use of the technology for applications in knowledge management. The methodologies were applied to the case studies in SWAP and updated based on the experiences resulting from its use. The updated version of the methodology (Deliverable 1.6) was published at the end of the project, and will hopefully continue to feed from a growing expertise in delivering DKM systems based on ontologies and peer-to-peer. The activities of the methodology called HOPE are naturally grouped in five phases of development, namely Analysis, Design, Development, Deployment and Evaluation. Phases are ordered in the sense that each phase results in well-defined deliverables used by subsequent phases of development. In order to avoid the pitfalls of a flat, linear, waterfall' style process, HOPE advocates an evolutionary approach, where development is an iteration of the development cycle rather than a one-shot execution of a sequential process.

In a second step the methodology was updated to incorporate the DILIGENT methodology for ontology engineering in distributed, loosely controlled environments. The refinement takes into account that ontology engineering is a continuously ongoing activity in HOPE that deserves special attention: the performance of the system requires some degree of control over the evolution and drift of the ontologies of peers without breaking with the peer-to-peer nature of the system. Ontology development also merits a separate treatment as it proceeds along a lifecycle of its own, which is likely to complete faster

(more frequently) than the main system development cycle.

2.2 Conceptual Design

This section provides the description of the global architecture for SWAP System. The architecture meets two important tasks. First it gives a high level view of the system so many people can understand it. Second it sets the basis for developers to implement the final solution. To fully describe a complex system as SWAP its architecture must explain the system from different points of view. That is why Conceptual Design is divided in four sections describing different aspects of the system. These are: Conceptual Architecture, Modular Architecture, Technical Architecture and Integration Architecture.

2.2.1 Conceptual Architecture

The Conceptual Architecture describes the SWAP System from a technical point of view.

First, it describes what we understand the SWAP system to be and how it is organized in terms of its parts. Here, three main parts have been detected: the SWAP System itself, the SWAP Nodes and the External Knowledge Sources.

The Conceptual Architecture also establishes a logical separation of the different layers that constitute a SWAP Node. This will provide the different topologies that can appear for a SWAP Node. These layers are: the Business Logic layer (provides the basic functionality), the Middleware layer (provides the business logic layer with access to external resources), the Local Node Repository (provides a placeholder for the Node to keep persistent information), and the Presentation layer (provides the human user with access to the SWAP system).

Finally the Conceptual Architecture provides the list of software components that conform to the SWAP System. A component is defined as a piece of software that is able to perform a specific technical task and it is combined with other components to form software modules. It describes in detail each of the components in the list, providing their definition, the function they perform inside the overall architecture, and their API (Application Programming Interface). These components are the guiding line for software developers and are: Ontology extractor, Ontology merger, Ontology Editor, Visualiser, Query formulator, Query replier, Query rewriter, Inference engine, Peer selector, Advertisement, Discoverer, Ontology integrator, Metadata integrator, Conflict resolver, Resource retrieval, Local node repository adapter, Communication adapter, and External knowledge source adapter.

2.2.2 Modular Architecture

The Modular Architecture describes the SWAP System from a functional point of view.

It describes the different roles of users that may interact with the SWAP system. These are: Ontology engineer (responsible for creating and updating ontology definition), Knowledge provider (responsible for populating the system with information), Knowledge seeker (the information consumer), and System administrator (maintains the system infrastructure).

The Modular Architecture presents the three functional requirements met by the SWAP System. These issues are: Ontology management (creation, visualization, update, and mapping of ontologies), Information management (creation, retrieval, merging, inferencing, and visualization of knowledge), and System management (technical and functional aspects).

The Modular Architecture also provides the list of software modules that conform the SWAP System. A module is defined as a piece of software that is able to perform a specific functional task, working alone, independently of other modules. It describes in detail its definition, the function it performs inside the overall architecture, the components needed to implement the module, the relation between those components, etc. These modules are: Ontology extraction module, Ontology edition module, Query formulation module, Query reply module, Visualizer module, and Expertise module.

2.2.3 Technical Architecture

The Technical Architecture presents the technical environment that must be used to develop the SWAP project.

First, it discusses the main goals for development, such as open standards, free software, TCO, multi-platform, technology prospects or SWAP partners skills.

Then, taking into account those goals, it chooses the development paradigms to be used for: Operating system, Database, Database connectivity, Communication software, User interface, Programming language, Ontology language, and Query language.

2.2.4 Integration Architecture

The Integration Architecture finally describes three main issues.

First it describes how internal system components are integrated, this is, how they interact with each other, what are the “patterns of integration”, etc. To do so, the Integration Architecture describes the Component Model (set of contracts between components), the Component Container (environment where components meet each other) and the Data Model (the basis for components to talk to each other).

It also describes how User Interface elements are integrated and what to do if one wants to add a new User Interface element. For the user interface the eclipse platform has been selected as it provides: Platform Runtime and Plug-In Architecture, SWT (a widget set and graphics library), JFace (an UI toolkit), Workbench (an UI integration face), Help plug-in, and Update Manager (for software distribution).

Finally, it describes how the SWAP System can interact with external applications. In this respect integration with MS Office, Adobe Acrobat Reader, OntoScrape, OntoEdit, and Full-Text Search Engine is explained.

2.3 Methods

The fulfillment of the projects objectives required not only the recombination of existing methods but in many cases the development of new ones. First of all it was necessary to design a metadata model 2.3.1 to allow for efficient exchange of common information. As known from other distributed systems heterogeneity is a major problem in peer-to-peer systems which led us to the development of novel mapping methods 2.3.2. We also had to investigate possibilities of efficient query routing mechanisms to guarantee the scalability of our system 2.3.3. In particular for the Bibster system we developed a recommendation system for bibliographic entries 2.3.4. Also in the context of Bibster the repository was confronted with highly complex queries which made particular optimization of our RDF story necessary 2.3.5. Finally we were able to extend the OWL ontology language with context information 2.3.6.

2.3.1 Metadata model

Metadata plays a central role in the effort of providing search techniques that go beyond string matching. Ontology-based metadata facilitates the access to domain knowledge. Furthermore, it enables the construction of semantic queries. Existing approaches of ontology-based information access assume a setting where information providers share an ontology that is used to access the information. In a peer-to-peer setting, this assumption does no longer hold. We rather face the situation, where individual peers maintain their own view of the domain in terms of the organization of the local file system and other information sources. Enforcing the use of a global ontology in such a setting would mean to give up the benefits of the peer-to-peer approach mentioned above. Therefore, one has to find a way to deal with the existence of multiple, distributed and frequently changing views on the domain. We have defined an RDF(S) based metadata model that combines ontological structures with information needed to align, evolve and use these structures for query processing and peer selection.

2.3.2 Mapping

Semantic mapping¹ between ontologies is a necessary precondition to establish interoperation between peers using different ontologies. Existing mapping methods were not suitable for the ontology integration task confronted in SWAP, as existing methods lay focus exclusively on improving the *effectiveness* and neglected *efficiency*. To illustrate our requirements: We have been working in realms where light-weight ontologies are applied such as the ACM Topic hierarchy with its 10^4 concepts or folder structures of individual computers, which corresponded to 10^4 to 10^5 concepts.

We have developed a mapping method which takes into account quality of mapping results as well as the run-time complexity. Our hypothesis was that mapping algorithms may be streamlined such that the loss of quality (compared to a standard baseline) is marginal, but the improvement of efficiency is so tremendous that it allows for the ad-hoc mapping of large-size, light-weight ontologies. We have shown in experiments and theoretically that our new efficient mapping approach QOM delivers mapping results which are comparable to the most sophisticated mapping methods consuming only a fraction of the time.

2.3.3 Peer selection

peer-to-peer systems are distributed systems without any centralized control or hierarchical organization, in which each node runs software with equivalent functionality. A review of the features of recent peer-to-peer applications yields a long list: redundant storage, permanence, selection of nearby servers, anonymity, search, authentication, and hierarchical naming. Despite this rich set of features, scalability is a significant challenge: peer-to-peer networks that broadcast all queries to all peers don't scale - intelligent query routing and network topologies are required to be able to route queries to a relevant subset of peers. Little effort has been made with respect to rich semantic representations of metadata and query functionalities beyond simple keyword searches. For such networks we have proposed two innovative peer selection methods. In the first one peers use a shared ontology to advertise their expertise in the peer-to-peer network. The knowledge about the expertise of other peers forms a semantic topology, independent of the underlying network topology. If the peer receives a query, it can decide to forward it to peers about which it knows that their expertise is *similar* to the subject of the query. The advantage of this approach is that queries will not be forwarded to all or a random set of known peers, but only to those that have a good chance of answering it. The second peer selection method REMINDIN' is based on social metaphors to select appropriate peers. It learns from queries routed through a peer and from answers of remote peers. Additionally it uses query relaxation to select appropriate peers for a query. REMINDIN's main advantage is that it does not require peers to distribute any information besides queries and answers.

¹Frequently also called alignment.

Both models were successfully evaluated. The first mechanism is deployed in the BIBSTER system.

2.3.4 Recommender

These recommender functions build upon two main features of the peer-to-peer system:

- *Semantic representation of metadata*: When bibliographic entries are made available for use, they are structured and classified according to two bibliographic ontologies, the SWRC² ontology and the ACM³ topic hierarchy. This ontological structure is then exploited to help the user formulate semantic queries. Query results again are represented according to the ontology. These semantic representations of the knowledge available on the peers, the user queries and relevant results allow us to directly create a semantic user profile and rich semantic similarity functions as a basis for recommending information that may potentially be interesting to the user.
- *Peer-to-Peer infrastructure with a semantic topology*: The peer-to-peer infrastructure reflects the distributed, decentralized and dynamic nature of creation of bibliographic metadata in a research community. In fact, a centralized solution does not exist and cannot exist, because of the multitude of informal workshops that researchers refer to, but that do not show up in centralized resources such as DBLP⁴. The decentralized peer-to-peer architecture can immediately be exploited for recommending newly created data as soon as it becomes available in the network.

Using semantic descriptions of the knowledge available on the peers, we are able to create semantic topologies that reflect the social networks of research communities: Peers with similar interests and expertise are clustered, such that the semantic neighborhood of a peer automatically covers a set of peers that contain relevant information for the specific community of interest. Furthermore, to address the cold start problem that recommender systems typically have to face, we can make use of the of the peer's semantic neighbors to create an initial user profile.

2.3.5 Processing Complex Queries

The methods for peer selection and query processing considered in SWAP so far make the silent assumption that queries are rather simple and that they can be answered as a whole by a single peer. As the power of semantic web technologies, however, lies in the ability to describe and query complex conceptual models, these assumptions come with severe limitations of the approach. In order to address this problem, we have designed methods

²<http://www.semanticweb.org/ontologies/swrc-onto-2001-12-11.daml>

³<http://www.acm.org/class/1988/>

⁴<http://dblp.uni-trier.de/>

for indexing information in a distributed system based on more complex data structures, i.e. on specific RDF schema paths. It is represented through the query language SeRQL.

2.3.6 An OWL Extension for P2P systems

Ontologies are *shared* models of a domain that encode a view which is common to a set of different parties. Contexts are *local* models that encode a party's subjective view of a domain. We show how ontologies can be contextualized, thus acquiring certain useful properties that a pure shared approach cannot provide. We say that an ontology is contextualized or, also, that it is a *contextual ontology*, when its contents are kept local, and therefore not shared with other ontologies, and mapped with the contents of other ontologies via explicit (context) mappings. The result is *Context OWL (C-OWL)*, a language whose syntax and semantics have been obtained by extending the OWL syntax and semantics to allow for the representation of contextual ontologies. The nature of this mixed approach makes C-OWL better suited to the needs of P2P systems than pure ontology languages.

2.4 Tool Development

The SWAP project was explicitly aimed at not only investigating the combination of Semantic Web and P2P technologies for knowledge management but also at providing supporting technologies. In order to live up to this claim, a number of tools were developed that support the structured exchange of information in P2P networks. In particular, these tools aim at supporting the following functions:

- Publishing local information in the network
- Access distributed information using RDF
- Exploring information available in the network

We briefly discuss the methods that have been developed to support these three functions. We focus on tools that have actually been implemented in the main SWAP system and are ready to be used in applications.

2.4.1 Content Publishing

The project has developed tools for migrating user information into an RDF schema representation. These tools can roughly be distinguished by their two main functionalities:

- Extracting knowledge from user data

- Semi-automatic refinement and integration of extracted knowledge

The corresponding tools are described in this deliverable.

The **OntoScape** tool extracts knowledge from selected parts of a users file system, email folders, browser bookmarks and address books as well as a special tool for generating RDF representations of a users' bibtex files. The latter will be used in the bibtex case study. Further, we use an extension of the **OntoEdit** Ontology editor that is specialized in browsing and editing the knowledge structures extracted by OntoScape. The system also supports the manual merging of concepts in order to integrate the different knowledge structures extracted from a user's computer.

P-Up is an extension of the OntoScape tool which allows the user to extract the shared folders faster to the local node repository. P-Up detects new/changed folders on the local machines. The output of P-Up can thus be used as input for our semi-automatic ontology learning process from folder structures. The process defines a way to construct ontologies from folder structures. For this process it uses different input sources. Input sources can be dictionaries, existing ontologies and named entity lists. Given an existing ontology on the local machine and a dictionary the process can be used to suggest changes to the local ontology based on the new folders.

In order to support the creation of *common* ontological knowledge that links the data structures extracted from different users data we adapted practices for ontology creation to fit the special requirements of collaborative and distributed modeling. In order to support this process in the SWAP project a process model has been defined and the editing functionality of the OntoEdit tool has been extended to support this process.

2.4.2 Distributed RDF Processing

Sesame is an open source RDF repository for storing and retrieving metadata with support for RDF Schema inferencing and querying. It is developed and maintained by Aduna (formerly known as Administrator). The main goal of Sesame is to provide a middleware for effective and scalable storage and retrieval of metadata represented in RDF and RDF Schema. We have chosen Sesame as the basis for implementation of a distributed RDF repository because of its flexible and extensible architecture.

In the project we extended the functionality of Sesame to fit the needs of distributed information access. We embedded Sesame in a JXTA based P2P platform this enabling different Sesame systems to communicate in a P2P fashion by sending RDF queries to selected peers in the network and collecting the answers. This platform was successfully used to implement applications in the two case studies carried out in the project.

The nature of peer-to-peer information sharing leads to a situation where redundant information is present in the network. This redundancy concerns documents as well as other objects such as representations of authors or organizations. Using the RDF infrastructure described above, duplicate results are retrieved if they are represented by RDF resources

with different URIs and have to be identified and eliminated by the user. The project has designed and implemented heuristic methods for detecting resources that correspond to different real world objects in order to free the user from this task (**P-Map**). Besides general considerations about a strategy for duplicate detection some concrete detection methods have been implemented.

One of the main shortcomings of the naïve use of the P2P metaphor is the lack of a strategy for peer selection. Queries had to be sent to all peers in the network (broadcast), to a random set of peers or a number of peers had to be selected manually by the user. System tests have shown that the scalability of the system requires intelligent selection strategies that prevent a flooding of the network with requests. In the project an **automatic peer selection strategy** was implemented that uses descriptions of the expertise of a peer in terms of topics they offer information about.

2.4.3 Visualization

We developed a result-centric visualization of information available in the peer network. The visualization is based on the **Spectacle Clustermap Technology** that has proven to be one of the most useful in an evaluation of different existing techniques for presenting RDF data. We extended the existing technology with the possibility to multiple search criteria including different classification hierarchies and full text search. This is a major step towards an efficient exploration of distributed information sources. The visualization described has been designed in cooperation with the case study partners and has been integrated into the Xarop system used in the IBIT case study. First user tests have been carried out in the context of general system testing. A more detailed evaluation of the visualization has to be carried out in the future in order to identify remaining problems. Such user studies have been carried out for similar visualizations in the area of providing access to distributed document repositories. The results from these related studies were quite positive and indicated that the cluster map technology is suited for supporting the exploration of distributed information sources. An open question for such an evaluation is how well the dynamics of a peer-to-peer system can be captured.

2.4.4 Conclusions

The project has provided an infrastructure and supporting tools for publishing and exchanging user data in a P2P setting using RDF technology. The infrastructure combines the benefits of P2P information sharing (scalability, robustness) with the advantages of RDF (limited reasoning, visualization). This combination makes the SWAP tools valuable for many information sharing applications.

We also have to conclude that some goals of the project were set too ambitious as many methods did not make it beyond an experimental stage. Examples for such methods are the automatic alignment of heterogeneous schemas and the processing of complex

queries that involves partial answers from different peers. While important insights about these methods were gained in the project and published at leading conferences they need further investigation before they can be integrated in stable user friendly tools.

2.5 Integration

Goals The goal of this work-package was integrating tools developed within the project into a common, reusable software platform and instantiating this platform in the two case studies.

2.5.1 SWAP Integrated Application Platform

During the project, the consortium developed the SWAP Integrated Application Platform - a framework for building semantic-driven peer-to-peer applications.

The platform is highly modular in its design to allow for easy instantiation for specific case studies. Each kind of functionality or semantic processing of data is encapsulated in a reusable components, which allows to easily switch implementations of e.g. the semantic repository, network communications layer or query routing mechanisms. The platform setup can be easily changed, e.g. a component implementation can be switched by a simple change in a single XML file.

The very basic concepts that have been formulated for the platform have been the following:

- there are many peers,
- each peer has some local knowledge and resources to share,
- there is nothing like a global ontology shared by all peers,
- therefore each peer is itself responsible for being able to communicate with other peers, e.g. for translating queries it formulates into terms understandable for remote peers.

The platform has the following architecture:

- delivers a single component container, which holds and manages all component instances, separates component implementations from each other allowing component reusability and easy setup change mentioned above,
- defines collaboration between components, which generalizes functionality of semantic-driven peer-to-peer application.

The components defined by a platform and their role are the following:

- Local Node Repository - storing local peer knowledge in an RDF repository, with the ability to query for this data and an inference engine built-in,
- Communication Adapter - peer-to-peer network communication layer, allowing peers to query each other, send the query results, advertise its local knowledge into the network and share local resources (e.g. files or mail messages) with other peers,
- Query Translator - component responsible for translating queries formulated in terms of local ontologies into language of remote peer's ontologies before sending the query to the remote peer,
- Query Replier - responsible for performing the query and sending back the results,
- Advertiser - responsible for initiation of advertisements of local peer knowledge into the network,
- Ontology Integrator - ontology integration mechanism,
- Meta Integrator - component responsible for integration of meta-layer of knowledge,
- Merge - duplicate detection and similarity mechanism,
- Extractor - knowledge extracting mechanism, allowing to extract knowledge from local peer resources - like files or browser bookmarks,
- Peer Selector - semantically driven query routing mechanism,
- Controller - the component that controls the collaboration among all other components

2.5.2 Platform Enhancements

Additionally, two major platform enhancements have been developed:

- security framework based on peer-to-peer principles, allowing for access control of local resources with trust delegation mechanisms,
- integration with search engine, allowing for both full-text and semantic search in local peer resources (like files).

2.5.3 Two Platform Instantiations

The integration effort was also put into instantiating the platform in the case studies, resulting in two applications - Xarop (for Virtual Enterprise Case Study) and Bibster (for Scalability Case Study).

2.5.4 Strong Collaboration Among Partners

This one is not a “hard” deliverable, but also a very important achievement of the project. Delivering all things mentioned above was possible only because of really close collaboration between all involved partners. So we managed to:

- set up the common code repository, all the development was done against this repository, so the integration efforts have not been left for the last project phase, but spread over the whole project (which minimized the risk of not finishing them in time),
- perform many tests of the platform, which was a key success factor for ensuring it really works in real-world settings.

Chapter 3

Lessons Learnt from the Use Cases

3.1 IBIT Case Study

The main goal of this section is to present the evaluation results and lessons learned from the use of XAROP, the prototype developed for the IBIT Case Study in SWAP under the WP7. The section presents the analysis of the data gathered in the evaluation, and draws the main conclusions of this evaluation.

3.1.1 Evaluated User Interfaces

We have evaluated the two user interfaces of XAROP: The Knowledge Searching Interface, that is the main user interface to structure, share and retrieve knowledge; and the OntoEdit plug-in, the interface embedded in Xarop to change/update the local ontology, and to assist the steps of the DILIGENT methodology defined to maintain a common ontology.

3.1.2 Evaluated Users

Initially, we planned to test the system on 28 peers from 7 organizations, but due to some users leaving their organizations, and in other situations some users not being able to use the system for technical reasons (not appropriate hardware), we finally evaluated the system on 20 peers from 6 organizations. The evaluation plan of XAROP was executed between July and August 2004.

3.1.3 Evaluation methodology

The evaluation methodology was based on User Feedback methods. We obtained the user feedback from two sources: an on-line questionnaire and through interviews with the users. The on-line questionnaire, with 55 questions, was divided by evaluation goals: 30 questions on the user interface, 13 questions on the functionalities for knowledge structuring and knowledge searching, 8 questions on knowledge sharing, and 4 questions on organizational and economical impact. User Interviews were particularly useful in eliciting information about user attitudes, impressions, and preferences.

We also used Observational Methods that involved real people using the system. The users were observed directly doing special or ordinary tasks in their work, while the observer made notes about their behavior and performance that were considered important to record. Unfortunately, we didn't have access to information about the number of queries sent around by peers to perform a System Evaluation.

3.1.4 Analyzing General Results on the User Interfaces

During the installation phase, no installation problems occurred, and the XAROP tools didn't have blocking bugs that impede normal use of the tools. In general terms, the performance of the tool was good. The users could run without problems the platform simultaneously with other applications that they normally use in the everyday work.

Taking into account the indicators obtained in the general questionnaire on the user interfaces (NOTE: all the gathered data from the questionnaires can be consulted in the Deliverable-7.5), it is reflected that the users had a positive impression on the user interface. Some graphical elements, as the cluster map, contributed to give this impression. The users encountered the organization of the information and the arrangement of menus and menu items clear and usable, respectively. Regarding the screen features, the screen layout is well organized for the users, separating clearly the side for visualising/presenting query results and the side for browsing folders and sending queries.

The terminology used by the system was considered consistent, and the multi-language interface contributed to improve this opinion, but the system still requires that the users have to understand specific terms on P2P networks and the Semantic Web domains. This is specially dramatic when using the OntoEdit plug-in, where words as 'concept', 'instance', 'relation', 'attribute' are not at all familiar to the users without a background in knowledge engineering, in the context of the application.

The users penalized the lack of information about the state of progress of some tasks performed by the system. This is the case when sharing folders with OntoScrape and when downloading files.

The users penalized the 'learnability', this is the case with which the users can begin effective interaction and achieve maximum performance with the system. It was not easy

for the users without a previous experience in the use of knowledge management tools to know what is the 'process' to follow to share/retrieve knowledge: (1) sharing folders and bookmarks, (2) full indexing the documents, (3) classifying documents using the local ontology, (4) establishing permissions of access to files. Further, it seemed, that it was not easy for the users to discover new functionalities of the system. Also, the difficulty for learning by trial and error could be attributed to specific functionalities for the system, for example, the 'drag and drop' functionality to select documents from the results list panel to classify them under concepts of the local ontology. This difficulty for learning is extreme when we assume that the users have to be able to understand/learn how to manage the local node repository: After classifying manually documents under concepts of the local ontology, the users have to launch the OntoEdit plug-in if they want to remove the entries of classified documents. It is not intuitive for the users that the files can be saved to a local shared folder by dragging them from the results list panel to a local folder in the peer folders tree.

The lack of some supplemental information (FAQs, Glossary, Help) was also penalized for the users.

Regarding the obtained results about the system capabilities, the users consider that the system does not perform the tasks 'fast enough'. This appreciation could be due to the fact that the users normally don't understand/assume the asynchronous behavior of the system when sending queries. It seems that not experts users, familiar with Google, don't assume that the system can take some time getting the results from other users' PCs in the P2P network.

Regarding the specific questions on the user interfaces, the users considered that the icons should be more concrete and familiar, specially for the cluster map, where it is a bit confusing to differentiate what is a concept query and what is a folder query. This situation was also produced when the user adds a folder as a new concept of the local ontology using OntoEdit. In this case, there is no way to distinguish queries on the cluster map, because the folder and the created new concept have the same name.

It is relevant to emphasize that an important number of users considered that the menu structure of the system does not match the task structure. This can be related to the previously mentioned difficulty for a considerable number of users to understand the process of sharing, structuring and retrieving knowledge.

3.1.5 Analyzing Results on the Knowledge Searching and Providing functionalities

Regarding the questions about knowledge providing functionalities (sharing folders and bookmarks), the stated results show that the general opinion is positive, but some users pointed a missing feature: to give the possibility of sharing individual files, not only entire folders.

Surprisingly, the users have a negative opinion about the helpfulness of the automatic classification feature, and the process by which it is performed (during the full indexing of documents). In interviews with the users, some of them stated that they prefer basically to classify manually documents under concepts of the local ontology. This would mean that automatic classification is still an advanced feature for normal users. They normally conceive the full indexing process only as a previous step to perform keyword-based searches. So, some users were confused when after the indexing process some documents were automatically classified.

Manual classification of documents was, in general, appreciated as an interesting functionality to structure knowledge.

Regarding the kinds of searches on the local peer supported by the system, the users encountered more helpful and valuable browsing folders and the keyword-based search. The reason is obvious, taking into account that users are very familiar with the files manager of Windows and keyword searches on search engines like Google. The conceptual search was not considered, in general, useful for the users, and surprisingly this result is not coherent with the good acceptance and predisposition of the users, indicated previously, to classify documents manually. Anyway, this incoherent result could demonstrate that, after all, the conceptual search functionality was not well understood by the users. In interviews with the users, some of them didn't appreciate the benefits of getting semantic results in the searches. They didn't understand that the results of a query (in this case, of a concept query) could retrieve the documents classified under a broader term of the ontology, not only of the selected concept. Even, some of them conceive the ontology as a 'common folder structure', instead of an ontology to enrich the searches.

Some users pointed that, having the possibility of sharing a folder structure, they don't see the need of using an ontology to classify documents.

Lessons learned:

1. Keyword search - Sometimes semantic keyword search was not appreciated by users because they did not get the results they were expecting. They asked for exact matches.
2. Automatic classification - The users did not appreciate this feature at all. They would have preferred the system not to classify their documents. They wanted to classify their documents manually.

Regarding visualization of results on the cluster map, and the possibilities it offers to discover new knowledge and duplicate documents, users considered this functionality very useful and attractive, even though, some users were confused when the cluster map simultaneously presented the results of several queries. Normally, the users are not familiar with the possibility of mixing results of several queries, to detect duplicates and discover new knowledge.

Results reflect that it was very difficult for the users to use OntoEdit to change the local ontology according to their needs, because it requires knowledge engineering background that the users didn't have (only 5 peers evaluated the OntoEdit plug-in to try to change the local ontology). This result regarding OntoEdit will be more described in the next section.

In general terms, attending to the results of the final questions of this section, the users benefit from the use of XAROP to structure their local resources according to their personal views. Even though the users seem not to have been benefited from the use of an ontology, they admitted that the system contributed to improve their information searches.

3.1.6 Analyzing Results on Knowledge Sharing

The results obtained in this section reflect the parallelism with the results exposed in the previous section, when evaluating the searches on the local peer, but in this case focused on the evaluation of searches on remote peers. Users found the browsing folders functionality and keyword-based searches on remote peers more helpful than concept searches. Manual peer selection and downloading files features were also valued positively.

Regarding security, the users trust the levels of security provided by the system, with the possibility of establishing access permissions to individual files.

The final questions reflected that the system could not contribute significantly to save time when exchanging knowledge with members within the same organization. And this was similarly the case when exchanging knowledge with members of other organizations. In both cases the users preferred to browse remote folders, instead of sending queries. One of the reasons of this low contribution of the system to knowledge sharing is that a low number of peers were active all the time during the evaluation, this situation didn't contribute to motivate the users to share knowledge. This lack of motivation could force the need of establishing mechanisms such as incentives for the users to share knowledge in the future.

3.1.7 DILIGENT Methodology Evaluation Results

In this section we present the evaluation about how the use of the ontology benefited the creation of a common, inter-organizational knowledge view. We also analyzed how the individual view of each user contributed to the creation of the common view. For that, we have evaluated two iterations or cycles of the DILIGENT methodology. DILIGENT covers the following steps:

1. Build the common ontology
2. Local adaptations
3. Analysis

4. Revision

5. Local update

In the *first iteration*, the first version of the common ontology was delivered to the users by email (step 1) in a text file. Then the users extended the common ontology according to the local needs (step 2) and sent the extended ontology back to the ontology engineer (the changes were sent by email in text files too). Then, we created the central board, formed by one representative member of each organization. The representative member of IBIT also assumed the role of ontology engineer. The hotel associations/federations designated only one representative member for these organizations. The central board analyzed (step 3) and revised (step 4) the local adaptations done by the users to produce the new version of the common ontology. The board balanced the requests, and decided on the changes to be introduced in the ontology. The new version of the common ontology was sent to the users (step 5) for local updating (the ontology was distributed through the first distribution of the XAROP platform).

In the *second iteration*, some users liked the new version of the ontology (it is interesting to comment that for some users it was difficult to distinguish between topic hierarchies and concept hierarchies). The initial idea was that the users could produce the local adaptation in this second iteration using OntoEdit. But the users encountered difficulties using OntoEdit to change the local ontology. This tool requires some background in knowledge engineering that the users don't have. Even though, 5 peers from IBIT were able to do changes on the local ontology using OntoEdit, in the end, only the ontology engineer was able to benefit from the use of this tool. So, despite the users having OntoEdit available, we decided to continue collecting the local changes using text files (step 2). Each representative member compiled the proposed extensions by the users in one text file. Representative members had to specify which concepts to add and the place in the hierarchy in which they should be added.

Problems found in the extensions requested by the users: 1. Confusion between folders and concepts: (one of) the users thought that the taxonomy of concepts should mimic their folder structure. In this case the board decided not to follow the user's suggestions and introduce the requested changes, since this would compromise the quality of the ontology. It is interesting to notice that some users didn't understand why the ontology has to change. In one organization most users were active in suggesting changes and in others they preferred to delegate the task to one person. The reason could be that in some organizations, like CITTIB, there is a fixed organizational structure with a well-defined workflow. They can not decide nor contribute to change the process in the organization.

Nonetheless the folders provided insight as to some of the concepts users were missing in the ontology. Some required action could be: some more training on the differences of what concepts in an ontology are and how they should be related to the information users have in their folders and documents.

2. Introduction of concepts that belong to other modules (parts of the ontology or subdomains of knowledge. Please, see description of the modules in D7.5, page 20): (one of) the users requested the introduction of concepts in the module which he specifically worked with that correspond to concepts that should be introduced in another module. In this case, it would be beneficial that users should have some higher level view of the main modules of the ontology. This could probably be used to remember users of other possible modules where they could find some of the concepts they are needing.

Analysis of requested changes:

1. Most of them were specializations of existing concepts and broadening with new sibling concepts.
2. Some of them were the introduction of some instantiations.
3. Some of them were renaming due to change in legislation.

The local extensions provided by the users were analyzed (step 3) and revised (step 4) in a meeting held in Palma with members of the SWAP consortium and the Instituto Superior Técnico (Lisboa). We balanced and included some suggested extensions in the common ontology, and sent the new version to the rest of members of the central board to get their approval. In two of the modules we completely followed the requested changes from some of the users' board members. In one of the modules we did not follow the requested changes. They were requested by the member of the board that used his folder structure to suggest new concepts. However, we tried to abstract from the folder structure and introduced changes representing the concepts he would require. This second iteration has not concluded yet, we are currently expecting feedback from the users.

Lessons learned:

1. To ease the task of understanding the ontology. The users should have preferably focused access to the part of the ontology they are dealing with. Otherwise it can be too confusing. However, sometimes users should browse through the whole ontology to find concepts they want, that is, identify the concepts they need in the whole ontology and not just specifically in the "module" they are mainly using.
2. Adaptation - We have to provide easy ways to the users of extending the local ontology, avoiding the need of launching and using an ontology editor to produce the local changes.
3. Reuse - Users did not reuse concepts, ontologies, or adaptations from others. This may be due to lack of appropriate tools. On the other hand, no one requested this feature to be added.

3.1.8 Conclusions

The evaluation results indicate that the XAROP platform satisfied the requirements and the major expectations of the virtual enterprise case study. The results were positive in terms of usability and about the functionalities for knowledge structuring and searching, even though during the evaluation the system could not contribute significantly to save time when exchanging knowledge within and between organizations.

A relatively low number of peers were active all the time during the evaluation, this situation didn't contribute to motivate the users to share knowledge. We have to find mechanisms to give incentives to peers to share knowledge. One possible solution in the future to increase the number of active peers could be to establish some passive and permanent active peers in some organizations, putting available documents of public interest (indicator reports emitted by public institutions).

The DILIGENT process contributed to create a common understanding of the tourist destination management domain among organizations, but we have to deliver a more friendly ontology editing tool if we want to motivate them to provide local extensions of the ontology.

Regarding the impact of the system in the organizations, the general opinion is that the system is useful to support effectively and efficiently the processes of the organization, but the organizations do not expect that the system could support those processes in the future, beyond the case study. In conclusion, this reflects that a real deployment of the system will require a major compromise of the organizations to adopt supporting tools for the collaborative decision making, and also a 'social and culture change' of the users to new ways of cooperation.

3.2 Bibliographic Case Study

In this section we describe the lessons learnt of the Bibster case study based on the evaluation report delivered in [SMM⁺]. It presents the results of carrying out the evaluation plan, which is discussed in Deliverable 10.2. In this case study, SWAP technology is employed to help researchers in sharing bibliographic data, in order to save work in trying to find this data and avoid unnecessary and error-prone duplication of the bibliographic entries.

The evaluation consisted of two parts: a user-focused evaluation and a system evaluation. Results of both types will be presented in the following.

3.2.1 User Evaluation

Methodology

The user evaluation is based on feedback obtained from an online user survey and from email and web forums.

The on-line user survey is a test for assessing the user experience with certain particular features of the Bibster platform. It is also intended to profile the users in terms of their technical environment, scientific background and existing social relationships with each other.

The second source of feedback is from email and web forums. Users can subscribe to a mailing list where they can post questions and suggestions for improvement. The Bibster system has also been discussed in a Slashdot article¹. 79 comments were posted on this forum.

Feedback from on-line user survey

Users who downloaded the Bibster application from the web site were invited to fill out an on-line user survey. The first part of this questionnaire records specific circumstances that might effect the user experience: personal information, previous experience and the technical environment in which the software is deployed. The second part contains questions about the experience of the users with the tool as well as existing social relationships of users that might affect their behavior in knowledge transfer.

Over the period of 3–30 August, 2004, the online questionnaire was filled in 72 times. Note that the actual number of answers to the survey questions varies by question as the users had the possibility to opt-out, i.e. not to answer a question if they did not want to. We note that some members of the SWAP project have also filled in the questionnaire.

User background The first and last segment of the questionnaire contained questions related to the background of the users. These questions were introduced so that we can understand who the users are and put the answers to the evaluation in context.

Of those who answered the question related to their affiliation (28), almost all of them are employed in academia (24) with the vast majority in the position of a PhD student (21). The average years in research is low accordingly: only 6 years (or 5 years when removing a single outlier). When asked to specify their research interests, the most commonly occurring keywords were (in order of importance): ontology, ontologies, Semantic Web, peer-to-peer, information/information systems, knowledge/knowledge management, internet, digital library. This would come as no surprise, since the initial roll-out targeted the academic world and in specific the Semantic Web community and those interested in

¹<http://slashdot.org/article.pl?sid=04/08/02/2153251>

peer-to-peer technology by profession. We cannot conclude, however, whether the case is that the software did not have a wider appeal or simply those who tried from outside of these communities (e.g. those who came across the software through Slashdot) did not feel obliged to fill out the questionnaire.

The distribution to the question about \LaTeX use shows that users either use \LaTeX for all their publications or none at all, with the two groups equally represented in the sample. Those who run Linux use \LaTeX more than other users (86 vs. 48 percent), which is explained by the fact that \LaTeX originates in the Unix world. Altogether, however, there are less Linux users than one would might expect from this community: the distribution of the operating systems shows a dominance of Windows.

Concerning the hardware environment, the average user of the sample has a machine equipped with 512 MB of memory (based on 42 replies) and a processor of 1600 MHz (32 replies).

Working with bibliographic items The next section of the survey inquired about the present methods that users employ to find bibliographic data. From the results it seems that people first turn to Google, followed by Citeseer, DBLP and other methods in this particular order. Among the other resources mentioned we find local libraries, colleagues, IEEE (ieeexplore.ieee.org), Web of Science, SIGGRAPH, bioone, JStor, dafodil.de, Amazon, SearchPlus@LANL, and Scirus, where many of these tools serve a specific science community. See Figure 3.1.

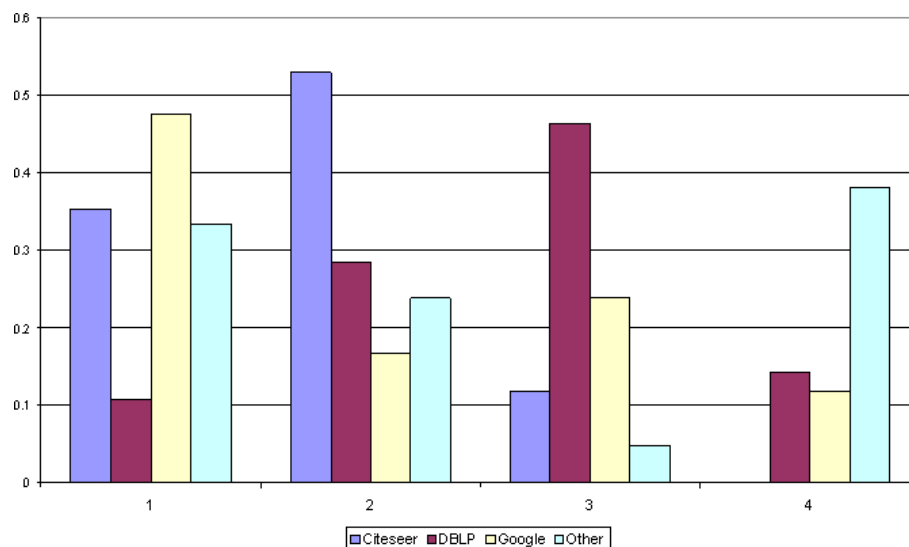


Figure 3.1: The percentage of users (y-axis) who ranked a given search tool on the first, second, third or fourth place (x-axis) when asked about their order of preference.

The high ranking of Google over the search engines specialized in bibliographic items can be explained by the unsurpassed coverage of Google in general web content including

the pages of CiteSeer itself. Also, it has been noted by many researchers in the past that web users prefer the fast answer over the precise one; it is likely that many people try Google first and then move on to a more specialized service if the answer cannot be found among the first few hits returned.

When it comes to local content, 56% of the users keep their bibliographic items in a single file, 41% use multiple files and only 1 person is known to use a database (of 39 responses). The vast majority of the users (75%) use a text editor to manage their file(s), 2 persons use JabRef, while 8 users use some other (unspecified) tool for this purpose. In essence, the bulk of the management of bibliographic items is still done manually without any specific tool support.

User appreciation of Bibster The main block of questions asked the opinion of the users about Bibster and its specific features.

First, we wanted to know the extent of the experience of the users with the system. It seems that about half of the users who have answered this question (24 persons) tested the system for a week or less, which suggests a high attrition rate. It is also unfortunate that a quarter of the users ran the tool for a day or less.

The overall usefulness of the system was rated as 2.65 (based on 23 answers) on a scale of 1 - Outstanding, 2 - Good, 3 - Average 4 - Needs improvement, 5 - Poor. This means that the software is considered slightly more average than good by the respondents. However, 78 percent of these respondents would like to keep the system after the evaluation period.

More targeted questions were also introduced to focus on benefits that we expected from the introduction of Bibster, i.e. the facilitation of knowledge discovery and knowledge sharing (learning) and improvement over previous search methods in terms of time and quality. Unfortunately, the results showed that for most users, the system did not increase the quality and efficiency of the search compared with previous search methods. No doubt, this had to do with the performance problems experienced with the system.

The users were also asked to rate individual features of the system. The following table shows the usefulness of certain features of Bibster as perceived by the users (based on 22 responses and using the same scale as for overall usefulness). The features are listed in decreasing order of approval by the users.

Exporting to L ^A T _E X	1.77
Searching using keywords	1.82
Searching using specific attributes (author, title etc.)	1.82
Detecting duplicate entries	2.14
Sorting by author	2.18
Adding new BibTeX entries	2.36
Looking up BibTeX items by topic	2.59
Sorting by date	2.63
Selecting peers manually	2.91
Sorting by the source of the information	2.95

Worthy to note the relatively low value attributed to looking up publications by topic as opposed to using keywords. The users in the sample also found little use to manual peer selection and sorting by peers. This seems to suggest that Bibster is considered more an anonymous peer-to-peer system (like Napster) than a community tool: the users are not concerned about the source of the information. (Note: peers were identified by user identifiers, not the real name of the users, although most users tended to choose identifiers that were reflective of their name.)

The respondents also had the opportunity to describe in free text the features that they liked the most, the ones they liked the least and any new features they would like to see. Among the positive aspects of the system, users mostly mentioned the idea itself, i.e. the combination of Semantic Web and peer-to-peer technology. Complaints mostly focused on the implementation: the download size, memory footprint and speed of the application, and the extensive use of the hard disk. Secondly, users suffered from the lack of shared data: too few publications in the system, focused on a narrow area (Computer Science).

In requests for new features, a number of users noted the need to share the actual documents (or at least share the URLs) and at least two users have mentioned the idea of a client/server version of Bibster. The idea seems to be a fix for the performance problem: a research group could set up a server (or "super-peer") that would regularly poll the client machines (the computers of individual researchers) for bibliographic data. Only the server would be visible to the outside and therefore the client machines would not have to answer to requests from outside of their group. This would naturally reduce the load on the client machines.

Feedback from email and forums

The direct feedback from the Slashdot forum and email provided more detailed information from a subset of the Bibster users on the usefulness, flaws and possible improvements. Here we analyze their comments and see which lessons we can learn from it.

There were 79 comments posted on the Slashdot forum about Bibster and +/- 40 emails from people that are known by members of the Bibster developers team. A high-level summary of the analysis of the comments can be divided in meta statements on using

a peer-to-peer approach and in comments on technical and usability improvements.

About the Peer-to-Peer approach

- **Reliability of the data:** The idea that every person can contribute bibliographic meta-data rises the question about the reliability of it. The meanings of the forum posters are divergent. One user refers to the CDDDB project which stores meta-data for CDs (Title, Artist and Track listings), where often substantial variations in the spellings, capitalization, punctuation, etc. are made between different providers. The lack of the possibility of constant peer reviewing, like in Wikipedia, could lead to inaccurate and untrustworthy info. Some possible solutions proposed by two users are a peer review mechanism or an automatic 'pagerank' inspired distributed rating mechanism.
- **Availability of the data** Some users address the positive aspects of the direct availability and redundancy of the data on a peer-to-peer network. The problem with Bibster however seems to be that at the current startup phase, not many users are online and that therefor also users will stop using it because they cannot find the information they seek. A comment via email was about solving this problem by adding some peers to the network by the developers of Bibster that have much content (like the complete DBLP database) and are constantly online.
- **Accessibility and use of standards** Some postings were about having different ways to get content from the Bibster network. One user comments that BibTeX may be widely used in the hard sciences, but which is not international-friendly, has a bad data model insufficient to the task of representing the sorts of data that scholars in the humanities deal with. One user suggests to use an open network protocol that would allow developers to make their own interfaces like a text based one. Also the ACM topic hierarchy needs to be more general to keep more users interested.

About features and possible improvements Comments and requests for new features and possible improvements concerned:

Full text and user comment sharing, support for other operating systems, improvements in the user interface, the dependency on the Eclipse framework, the installation procedure, and bug fixing.

Summary

In general, the user evaluation revealed that users like the idea of combining Semantic Web techniques with peer-to-peer for sharing Bibtex entries. However, the current Bibster implementation has such a big memory and CPU footprint and also a slow response rate

and low number of participants that many users did not like it now. The users that tested the system are working in the knowledge engineering domain, so we cannot make any statements about its usability in other research fields. The users like the possibility to export the local repository to \LaTeX and also the search possibilities on specific attributes (author, title, etc.).

3.2.2 System Evaluation

Methodology

The system evaluation is based on the analysis of system activity that was automatically logged to log files on the individual Bibster clients as described in the evaluation plan [SMMH]. Therefore, the user and system activity could be recorded and analyzed.

The log files were locally created in an XML format and periodically sent to a centralized FTP server. For analyzing purposes, the log data was written in a database and queried using Excel as statistic tool.

Environment

A total of 398 peers spread across multiple organizations mainly from Europe and North America used the Bibster system.

Figure 3.2 shows the distribution of the usage activity in terms of number of queries of the 13 calendar weeks of the evaluation period. Two effects are striking: First, because of holidays in July and August, the average activity in these months is lower than in June. However, there is a peak in calendar week 32, because of the attention drawn to the Bibster system by the positive comments at Slashdot², a worldwide well known news portal for information technology. Unfortunately, the positive effects of this interest were singular in nature and could not be retained.

Content Analysis

A total of 98872 bibliographic entries were shared by the 398 peers, with an average of 248 entries per peer. However, the distribution had a high variance (c.f. Figure 3.3): While 62% (248 peers) were free-riding³ and shared no content, 6% (24 peers) shared at least 1000 entries each, accounting for 79% of the total shared content.

²<http://slashdot.org/articles/04/08/02/2153251.shtml?tid=185&tid=1&tid=218>

³In many peer-to-peer systems (e.g. Napster, Gnutella) users are mainly interested in their own advantage and conserve their resources (i.e. bandwidth) by sharing no files. In the common literature this phenomena is called *Free-Rider* problem. Users do not have a direct incentive to share files.

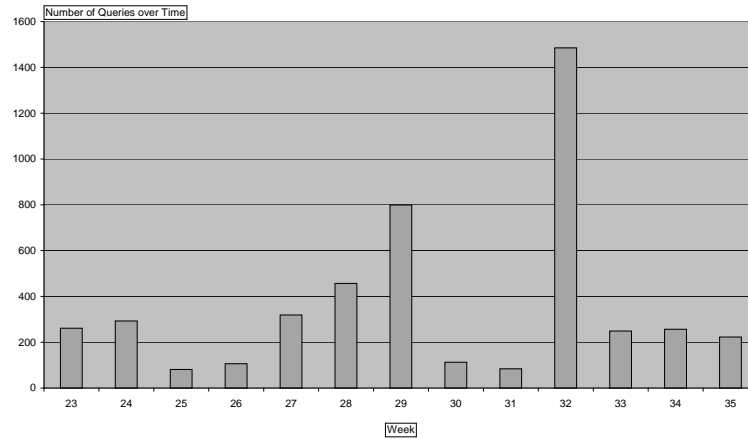


Figure 3.2: Distribution of Query Activity

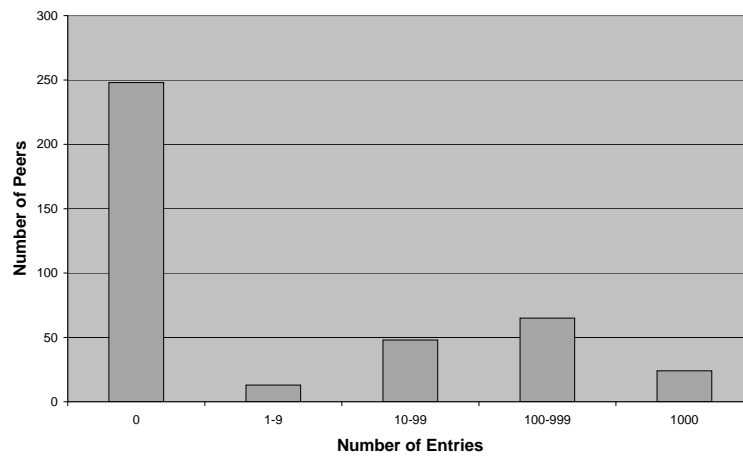


Figure 3.3: Distribution of Shared Content

Querying

The users performed a total of 3319 queries. The two ontologies (SWRC and ACM) were used for about half of the queries, whereas the other half used the generic search over all attributes. The SWRC ontology was mainly used for the purpose of querying for special

types of publications (e.g. for Articles) or for publications of a given author.

In 720 queries the users asked for topics of the ACM topic hierarchy (isAbout). Thereby it is likely that the users are accepting the ontology based searching capabilities and that there is a benefit for them in using these ontologies. Furthermore, the frequent queries for ACM topics would justify the chosen topic-based expertise model as the basis for peer selection.

With respect to the scope of the queries, the users mainly performed queries on their local peers and automatic search across the entire network. Only in few cases the queries were directed to a manually selected peer. This confirms the need for efficient peer selection algorithms.

Peer Selection

With respect to query routing and the use of the expertise based peer selection, we were able to reduce the number of query messages by more than 50 percent, while retaining the same recall of documents compared with a naive broadcasting approach. The precision of the peer selection process (the percentage of the reached peers that actually provided answers to a given query) was determined as follows: While the expertise based peer selection results in an almost constant high precision of 28%, the naive algorithm results in a lower precision decreasing from 22% after 1 hop to 14% after 4 hops⁴.

From the number of forwarded query messages sent per query, it can be seen that with an increasing number of hops, the number of messages sent with the expertise based peer selection is considerably lower than with the naive algorithm. Although we have shown an improvement in the performance, the results also show that with a network of the size as in the field experiment, a naive approach is also acceptable. On the other hand, with a growing number of peers, query routing and peer selection becomes critical: In simulation experiments with larger peer networks with thousands of peers, we have shown improvements in the order of one magnitude in terms of recall of documents and relevant peers [HSvH04].

Comparison with Results from Simulation Experiments

Overall, the results of the simulation experiments have been validated: We were able to improve the precision of the peer selection and thus reduce the number of sent messages. However, the performance gain by using the expertise based peer selection was not as significant as in the simulation experiments⁵.

This is mainly due to the following reasons:

⁴The decrease is due the redundancy of relevant peers found on different message paths: Only distinct relevant peers are considered.

⁵In terms of recall, there were no improvements at all, as even the naive algorithm generally was able to reach all relevant peers.

- *Size of the network* The size of the network in the field experiment was considerably *smaller* than in the simulation experiments. While the total number of participating peers was already fairly large (398), the number of peers online at a certain point in time was fairly small (order of tens).
- *Network topology* In the field experiment we built the semantic topology on-top of the JXTA network topology. Again, related to the small size of the network, the JXTA topology degenerates to a fully connected graph in most cases. Obviously, for these topologies, a naive algorithm yields acceptable results.
- *Distribution of the content* In the simulation experiments, we distributed the shared content according to certain assumptions (based on topics, conferences, journals). In real world experiments, the distribution is much more heterogeneous, both in terms of the expertise of the peers and the amount of shared content (c.f. Figure 3.3).

Summary

The Bibster system was used by 398 peers across Europe and the USA. Although free-riding was quite frequent, it did not hamper the overall system performance. The ontology based search functionality has been used by the users. We have been able to show that the usage of Semantic Web technologies and ontologies provide an *added value*.

Unsurprisingly, in small networks with *small user groups*, intelligent query routing is not a major issue. While it is beneficial to direct queries to specific peers known to the user (a trust issue!), advanced routing algorithms may only be beneficial for a much larger number of users in a network. Based on our experience we now conjecture that content-based routing and trust issues will have to converge for such larger networks, too.

3.2.3 Conclusion

The overall opinion of the users and the results of the system evaluations indicate that the ideas behind Bibster are promising but the current implementation and the low number of users discourage people to use it. It is striking to see that users do not seem to appreciate the topic hierarchy for searching but the log files of the usage did show that people really used it.

Due to the small size of the network, the advanced peer selection method cannot show its potential because broadcasting does not hinder. However, we can already see its advantage in the fact that the number of messages is halved.

The users liked the possibility to export the local repository to \LaTeX and also the search possibilities on specific attributes (author, title, etc.). The implementation of the

duplicate detection seems to work satisfactorily for the users. Many users were free-riders but this did not hamper the overall system performance.

Chapter 4

Dissemination

Due to the diversity of the SWAP consortium there were two broad lines of dissemination activities that we can point out. On the one hand, there was dissemination geared towards academic environments done primarily by the research partners with publications, presentations and co-organisation of main conferences, events and workshops related to the project. Likewise, there have been dissemination activities geared towards the industry participating in the more important events and mass media, at both the local and the international level.

The Continuous Dissemination Actions task has been going on throughout the entire project:

One of the deliverables was the design and continuous maintenance of the SWAP web page (<http://swap.semanticweb.org>), which has followed the project from its beginning to its official completion. It helped as a valuable means of dissemination for attracting the interest of the industry and research organizations in relation to IS Technologies compatible with those developed within the project, and potential users that could eventually incorporate the developed technologies into their organization schemes.

Marketing material has been produced to support the different events in which all the partners have participated during the project. A poster representing the project and leaflets in English and Spanish explaining the project have been prepared.

Many articles and papers have been published in specialized scientific/technical publications (and also in non-technical publications with strategic value for dissemination, such as those related to the user partners themselves or their related business/activity sectors: Public Administration, Industrial Associations, Student Associations, as well as the general Press).

The note following highlights (detailed by partner in the Individual Contribution Section of the deliverable D8.5, on the References web page of the project, and in the publications chapter of this final report). Academic publications: DIT 02, CIA 02, FQAS 02, KI 02, EKAW 02, AMKM 03, ISWC-03, WM 03, WIAMIS 03, SemPGRID, KMDAP

03, KIM 03, ENTER 03 ITC-IRST, XML Colloquium, OntoWeb, SIG 03, IJCAI-WS, WWW03, WWW-04, ECAI-04, ICSNW-04. Industrial and Mass Media Articles: Diario de Mallorca, Balearic TV, Innovation & Energy Bulletin (Balearic Gov.), MadrI+D website, 5DIAS, ComputerWorld, Computing.

The partners have attended appropriate forums for direct dissemination, such as specialized workshops, congresses and conferences, both national and European, related to Semantic Web and peer-to-peer information. These forums were not only in the EU, but also outside the EU: ITC-IRST, XML Colloquium, OntoWeb, WWW03 Workshop on P2P & Grid, SIG03, EKAW-2004, WWW2004, STAIRS2004, ECAI04, ESWS-2004, I-Know 2004, Meta4 International User Forum 2002 and 2003 First International KM Conference for Spanish Speaking Countries 2004, etc.

A Special Deliverable of the project is the D1.7 Book. Springer, a major publisher, will publish this book. The book summarizes the results of the project and will disseminate them to a broader audience. The title of the book will be “Semantic Peer-to-Peer for Knowledge Management”.

Concerning the contribution to standards, SWAP built upon existing standardization efforts in the field of Semantic Web and Ontologies. In fact, the two research institutes have been key players in creating the Semantic Web ontology standard OWL or the query language SeRQL, as well as in building and enhancing some of the most sophisticated existing Semantic Web tools (KAON, Sesame).

SWAP is part of the broader Semantic Web activity and has interacted closely with third parties (universities, EU projects, Consulting Firms) throughout the project. An extra deliverable has been produced in collaboration with Carnegie Mellon University to develop a case study. SWAP has also collaborated with Edamok, Meaning PADLR/ Edutella and Sewasie Projects and Aduna and Ontoprise Consultancy companies for tools.

Chapter 5

Exploitation

We can divide the actual outcome into two lines. On the one hand we have defined a range of new methods in the field of peer-to-peer systems on the other hand we have developed two applications for two different case studies. On the method side, the SWAP project has produced new methods in different areas and make significant contribution to the research of the Semantic Web. On the representation language level a query language for RDF and a context language for OWL were defined. We designed a meta model to capture different attributes of peers and the knowledge they share. This model is used by almost all other methods applied in the SWAP system. Ontology heterogeneity is inherent in ontology based peer-to-peer systems, since peers may or may not share the same ontology to model a certain domain. In the SWAP project we tackled the problem of ontology heterogeneity in two ways. Firstly we have developed a new mapping method, which generates mappings between ontologies much faster than existing methods and with higher accuracy. These mapping method are based on the combination of different similarity functions. Secondly we have defined a process which helps the participants of the network to find shared concepts of their ontologies. This reduces heterogeneity for the concepts most often used. Based on the research done on ontology heterogeneity we could also define a new method to provide recommendation to users. This method is used in the Bibster system to suggest new bibliographic entries to the user. A particular problem in peer-to-peer networks is the routing of queries to knowledgeable peers. In the course of the project two different routing algorithms were developed. The advertisement based approach facilitates the setup of the peer-to-peer network while the rating based approach can be used to fine tune the routing quality. Methods which are not yet implemented in any of the applications but point to future directions include new ways to process complex queries. This becomes interesting when the answer of a certain query must be combined from the results delivered by different peers.

On the application side, the SWAP project has produced a generic peer-to-peer platform to exchange data in RDF format. On top of this generic application two case study specific solutions were build. For the research community the Bibster system was developed.

XAROP, was optimized to fulfil the needs of knowledge workers in the tourism domain. XAROP enables its users to share files, bookmarks and emails with other participants of the network. In the XAROP case study the number of participants is small, thus routing is not a problem. The methods most significant to the XAROP platform are the security mechanisms, the automatic document classification, the similarity detection between folders, the distributed ontology engineering process, new visualization metaphors and the view generation of the query language.

The other solution, Bibster is currently used by more than 1000 participants with the number growing each day. Bibster allows to exchange metadata about Bibliographic entries. In this system methods for intelligent routing as well as for similarity detection are of particular interest. Similarity between bibliographic entries can be used to determine duplicates in a given answer or to suggest similar entries which might be interesting to the user. Bibster can be downloaded through <http://bibster.semanticweb.org>.

The Technological Implementation Plan summarizes the results of the project and describes the exploitation plans and the business plans of the different partners, the tangible results of the project.

Following the recommendations of the project Final Review Meeting, apart from on the eTIP web page, IST results have also been published on the CORDIS web page, in the Technology MarketPlace area. The two prototypes resulting from the project have been published: Xarop (RCN 33254) and Bibster (RCN 33192).

Each partner has implemented own exploitation plans to take advantage of both the knowledge acquired throughout the project and its tangible results.

Among the latter are the two prototypes, Xarop and Bibster, developed in the two case studies on the platform developed in SWAP, which are now in executable state.

Regarding the Xarop prototype, the traditional marketing channels if the industrial partners will be used with a licence model based on the number of peers or users, first on the installed base of Meta4 clients, taking advantage of the tourism environment and the relationships that IBIT has in the tourism environment.

To facilitate the dissemination of the results, the Bibster code has been classified open source and has been placed in SourceForce for download. The exploitation of this prototype, due to its focus, is being carried out among Researchers and academic institutions world-wide with an LGPL licence model, headed by AIFB and VU. The exploitation and system-roll out was organized in three phases: In the first phase (April to May) the system was delivered to the SWAP participants and their associated institutes. The purpose of this phase was to identify bugs and test the stability of the system. In the second phase (May to June) the system was rolled out and introduced to some selected institutions. During this phase about 30 - 40 users worked with the system. In the last phase (June to September) the system was propagated on several mailing lists and therefore delivered to the public. In this last phase we counted over 10000 visitors of our website (<http://bibster.semanticweb.org>) and more than 3000 downloads of the system.

Aside from the prototypes, we have used the conceptual work and components developed during the project. In close cooperation with the company Aduna, the Vrije Universiteit Amsterdam has developed and extended some base technologies for accessing information in P2P networks that will be exploited in future research as well as commercial product development at Aduna. There are mainly three relevant technologies: Visual Front-end for ontology information in P2P networks, Sail Main Memory Implementation and SERQL Language Definition.

The Security architecture, SeRQL engine and the conceptual work done on distributed knowledge management will be used in the Emplois CM Solutions. Meta4 have also evaluated the possibility of integrating the technology developed during the project to develop a new release of the Meta4 KnowNet product, opening new lines of R & D, to satisfy the current market needs, which tend towards simplicity and collaboration.

Chapter 6

Summary

“SWAP combines two highly successful technologies, viz. Semantic Web and Peer-to-peer computing.”

This quote is taken from the original project proposal and the technical annex of the project. To summarize the concrete results as presented on the previous pages, we will now map these results to the original goals.

“1. SWAP will offer a Conceptual Analysis of the problem. The analysis comprises the use of semantic web technology in peer-to-peer computing, but it will also consider how emergent semantics may be constructed by overlapping knowledge structures of different peers (e.g. common categorizations of knowledge, common bookmark names of web pages).”

We have provided a thorough analysis of the involved paradigms. Different deliverables have looked at the paradigms themselves, as well as their combination for knowledge management. Based on this analysis we developed a methodological approach, which was then used for SWAP.

“2. SWAP will explore the conceptual architectures made possible by this combination. They constitute the foundations for the subsequent integration task as well as for further exploitation.”

An architecture has been developed considering various aspects and degrees of detail. The aspects ranged from high-level general architecture considerations, to very specific issues such as software and hardware requirements from user side. From this we developed a general understanding on the structure which the SWAP system was to be built on. It was reflected in the SWAP architecture.

“3. SWAP will investigate new methods to realize a Semantic Web based peer-to-peer system.”

A lot of innovative research has been carried out in the two and half years of the SWAP project. Many individual methods were output of this research. For a full list please refer to the published deliverables. The success was not limited to the project, but was spread across the research community. The list of academic publications can be seen as an indicator for this.

“4. The methods will be implemented in a set of tools.”

Theoretical concepts and methods alone are not enough for success. The methods were consequently transferred and implemented in a set of tools. All of them, except for the very latest, could be integrated into the SWAP system.

“5. The tools will be integrated based on the foundations from the conceptual architectures in a demonstrator that can be tested in case studies.”

Integration of very different tools into a system, which has to meet very different requirements from case study side, was a challenging task. Nevertheless, this task was met in the project very well. The results can be viewed in the two systems: Xarop and Bibster. Effectively, we did not only develop demonstrators, but tools which run across the world on a daily basis. For a demonstration of our systems we refer to the open-source Bibster (<http://bibster.semanticweb.org>), which can be easily downloaded, installed, and run.

“6. The tool environment will be evaluated in two case studies about Knowledge Management for investment analysts and about Knowledge Sharing between SMEs (virtual enterprises).”

The two SWAP systems have been thoroughly evaluated in two case studies, of which one had to be replaced during the project term due to the drop-out of one partner. Xarop was successfully run across six independent organizations. Bibster is still running around the world. The objective evaluation has shown mostly positive very interesting results. We have received extensive feedback from evaluators, and in the case of Bibster additionally from many users. From all this feedback we could extract valuable lessons learned. Further, our consortium was rewarded the do-IT software award (15.000 Euro) of the German state of Baden-Württemberg for the Bibster system.

The results of the project were constantly promoted, in past and current dissemination and exploitation activities. Dissemination through the webpage (<http://swap.semanticweb.org>) was in place shortly after project begin, it was constantly maintained, and will be kept online for the time after project termination. Numerous publications from both academia and industrial partners were presented at well-known research conferences as well as application oriented events. Exploitation will be continued by the commercial partners, which plan to use the gained technology in their business fields.

Open source efforts of the university partners will also continue, specifically we plan to exploit the basic software in both the projects SEKT and KnowledgeWeb.

As expressed by all partners, the project can be regarded as very successful. Partners already continue their effective collaboration in other projects, or they are actively planning on how to ensure future collaboration.

Appendix A

A.1 Deliverables

Deliverable No	Deliverable Name	Lead Participant
D1.1	Analysis P2P systems	VU
D1.2	Analysis Ontology-based KM solutions	VU
D1.3	Analysis O for P2P, P2P for KM	Meta4
D1.4	Analysis P2P for O	Unikarl
D1.5	First version of SWAP methodology	VU
D1.6	Refined SWAP methodology	VU
D1.7	Book	VU
D2.1	Conceptual architecture	Meta4
D2.2	Modular architecture	Meta4
D2.3	Technical architecture	Meta4
D2.4	Integration architecture	empolis Polska
D3.1	Method survey	Unikarl
D3.2	Methods design	Unikarl
D3.3	Methods implementation	Unikarl
D3.4	Test scenario	Unikarl
D3.5	Integrated Methods	Unikarl
D3.6	Refined Methods	Unikarl
D4.1	Platform selection	Unikarl
D4.2	Implementation of basic ontology-services	VU
D4.3	Tools - first versions	VU
D4.4	Tools - improved versions	VU
D4.5	P-Up	Unikarl
D4.6	P-Viz	VU
D4.7	Final tools	Unikarl
D5.1	Basic application platform	empolis Polska
D5.2	Integrated application platform	empolis Polska
D5.3	Instantiated application platform for KM case study	empolis Polska
D5.4	Instantiated application platform for virtual enterprise case study	Meta4

Deliverable No	Deliverable Name	Lead Participant
D6.1	Specification of use cases	Dresdner Bank
D6.2	Refined specification of use cases	Dresdner Bank
D6.3	Evaluation for KM case study	Dresdner Bank
D7.1	First User environment definition	IBIT
D7.2	Refined User Environment Definition	IBIT
D7.3	Virtual Enterprise SWAP prototype	IBIT
D7.4	Evaluation plan	IBIT
D7.5	Evaluation report	IBIT
D8.1	Public Web Page and Project presentation	Meta4
D8.2	Dissemination and Use Plan	Meta4
D8.3	Exploitation market survey	Meta4
D8.4	Technology Implementation Plan	Meta4
D8.5	Report on overall exploitation and dissemination actions	Meta4
D9.1	Report "Project Description"	UniKarl
D9.2	Report "1st Six-monthly Periodic Progress Report"	UniKarl
D9.3	Report "2nd Six-monthly Periodic Progress Report"	UniKarl
D9.4	Report "3rd Six-monthly Periodic Progress Report"	UniKarl
D9.5	Report "4th Six-monthly Periodic Progress Report"	UniKarl
D9.6	Report "5th Six-monthly Periodic Progress Report"	UniKarl
D10.1	Requirements analysis	UniKarl
D10.2	Evaluation plan	VU
D10.3	Integrated system SWAP-Bib	empolis Polska
D10.4	Evaluation report about SWAP-Bib	VU
D11.1	Authorisation requirements analysis and design	empolis Polska
D11.2	Integrated platform with authorization	empolis Polska
D11.3	SWAP-Oreng integration	empolis Polska
D11.4	SeRQL prototype implementation report	empolis Polska
	SWAP CMU Report	VU
	SWAP 2002 Annual Report	UniKarl
	SWAP 2003 Annual Report	UniKarl
	Research Agenda V.1	UniKarl and VU
	Research Agenda V.2	UniKarl and VU
	Final Report	UniKarl

All deliverables are available via the project webpage. A password might be required, if the document is confidential.

A.2 Publications

Publications are grouped by partner.

- T. Sollazzo, S. Handschuh, S. Staab, M. Frank, N. Stojanovic: “Semantic Web Service Architecture Evolving Web Service Standards towards the Semantic Web”. In: Proceedings of the 15th International FLAIRS Conference. Pensacola, Florida, May 16-18, 2002. AAAI Press
- Editorial; M. Frank, N. Noy, S. Staab (eds.). Semantic Web 2002 Proceedings of the Workshop at WWW-2002. Honolulu, Hawaii, May 7, CEUR Workshop proceedings, 2002.
- Christoph Tempich, Julien Tane, Steffen Staab, Marc Ehrig, Christoph Schmitz: “Towards Evaluation of Peer-to-Peer-based Distributed Knowledge Management Systems” at the conference for “Agent Mediated Knowledge Management03”
- Marc Ehrig, Christoph Tempich, Steffen Staab, Frank van Harmelen, Ronny Siebes, Marta Sabou, Jeen Broekstra, Heiner Stuckenschmidt: “SWAP: Ontology-based Knowledge Management with Peer-to-Peer Technology” at the “2nd Knowledge Management Conference (WM 2003)” in Lucerne
- U. Reimer, A. Abecker, S. Staab, G. Stumme (eds.). WM2003 Professionelles Wissensmanagement Erfahrungen und Visionen. Proceedings der Konferenz WM2003. Luzern, CH, April 2-4, 2003. Lecture Notes in Informatics. GI e.V. / Kllen Verlag, 2003.
- Marc Ehrig, Christoph Tempich, Steffen Staab, Frank van Harmelen, Ronny Siebes, Marta Sabou, Jeen Broekstra, Heiner Stuckenschmidt: “SWAP: Ontology-based Knowledge Management with Peer-to-Peer Technology” at the “4th European Workshop on Image Analysis for Multimedia Interactive Services” at the University of London
- Marc Ehrig, Christoph Tempich, Steffen Staab, Frank van Harmelen, Ronny Siebes, Marta Sabou, Jeen Broekstra, Heiner Stuckenschmidt: “SWAP: Ontology-based Knowledge Management with Peer-to-Peer Technology” at the “2nd Knowledge Management Conference (WM 2003)” in Lucerne
- Jeen Broekstra, Marc Ehrig, Peter Haase, Frank van Harmelen, Arjohn Kampman, Marta Sabou, Ronny Siebes, Steffen Staab, Heiner Stuckenschmidt, Christoph Tempich: “A Metadata Model for Semantic-Based Peer-to-Peer Systems” at the workshop on Semantics in Peer-to-Peer and Grid Computing at the 12th International World Wide Web Conference in Budapest, 2003.

- Marc Ehrig, Peter Haase, Ronny Siebes, Steffen Staab, Heiner Stuckenschmidt, Rudi Studer, Christoph Tempich: "The SWAP Data and Metadata Model for Semantics-Based Peer-to-Peer Systems" at the First German Conference on Multi-agent System Technologies, Erfurt, Germany, 2003
- Christoph Tempich, Raphael Volz: Towards a benchmark for Semantic Web reasoners - an analysis of the DAML ontology library. in Proceedings of Evaluation of Ontology-based Tools (EON2003) at 2nd International Semantic Web Conference (ISWC 2003), 20.10.2003, ed. Y. Sure, 2003 D7.4 Evaluation Plan EU IST-2001-34103 17
- M. Ehrig, C. Schmitz, S. Staab, J. Tane, C. Tempich. Towards Evaluation of Peer-to-Peer-based Distributed Knowledge Management Systems. In: L. van Elst et al. (eds) Agent-Mediated Knowledge Management. Post-proceedings of the AAAI Spring Symposium. LNAI 2926 Springer, 2004.
- Christoph Tempich, Steffen Staab, Rudi Studer: TIME2Research - Ein Wissenportal für den Unternehmensanalysten . WI2003, Dresden, 2003.
- Marc Ehrig, Peter Haase, Steffen Staab, Christoph Tempich: "SWAP - A Semantics-Based Peer-to-Peer System Demonstration" in Bergmann et. al. (eds.), Workshop Wissens- und Erfahrungsmanagement - FGWM-2003, GIWorkshopwoche LLWA, Karlsruhe, Okt. 6-8, 2003.
- Marc Ehrig, Christoph Tempich, Steffen Staab, Frank van Harmelen, Ronny Siebes, Marta Sabou, Jeen Broekstra, Heiner Stuckenschmidt: "SWAP: Ontology-based Knowledge Management with Peer-to-Peer Technology" at the Workshop Anwendungsentwicklung mit dem Peer-to-Peer Framework JXTA", Berlin, November 11, 2003
- S. Handschuh, S. Staab (eds.). Annotation for the Semantic Web. IOS Press, June 2003.
- A. Maedche, S. Staab. Services on the Move - Towards P2P-Enabled Semantic Web Services. In: Proceedings of the 10th International Conference on Information Technology and Travel & Tourism, ENTER 2003, Helsinki, Finland, 29th-31st January 2003, pp. 124-133.
- V. Pekar, S. Staab. Word classification based on combined measures of distributional and semantic similarity. In: Proc. EACL-2003, 10th Conference of the European Chapter of the Association for Computational Linguistics, April 12-17, Budapest, Hungary, 147-150.
- S. Staab, K. Aberer, D. DeRoure, O. Dousse, D. Fain, G. Flake, F. Heylighen, D. Pennock, W. Shen, P. Thiran. Neurons, Viscose Fluids, Freshwater Polyp Hydra and other Self-organizing Information Systems. IEEE Intelligent Systems, Trends & Controversies, 18(4), Jul/Aug 2003.

- D. Fensel, S. Staab, R. Studer, F. van Harmelen, J. Davies. A Future Perspective: Exploiting Peer-to-Peer and the Semantic Web for Knowledge Management. In: Davies, N.J., Fensel, D. & van Harmelen, F. (eds), *Towards the Semantic Web Ontology-based Knowledge Management*, Wiley, London, UK, 2003.
- C. Tempich, S. Staab, A. Wranik. REMINDIN: Semantic Query Routing in Peer-to-Peer Networks based on Social Metaphors. In: *Proceedings of the 13th International World Wide Web Conference, WWW 2004*, New York, USA, May, 2004. ACM Press.
- S. Pinto, S. Staab, C. Tempich. DILIGENT: Towards a fine-grained methodology for Distributed Loosely-controlled and evolvInG Engineering of oNTologies. In *Proc of ECAI-2004*, Valencia, August 2004.
- P. Cimiano, A. Hotho, S. Staab. Comparing conceptual, partitional and agglomerative clustering for learning taxonomies from text. In *Proc of ECAI- 2004*, Valencia, August 2004.
- Sofia Pinto, Steffen Staab, York Sure, Christoph Tempich: OntoEdit: Supporting DIstributed, Loosely-controlled and IntelliGent Engineering of oNTologies. 1st European Semantic Web Symposium in Heraklion, Greece, May 10-12, 2004
- Marc Ehrig, York Sure: Ontology Mapping: An Integrated Approach: 1st European Semantic Web Symposium in Heraklion, Greece, May 10-12, 2004 D7.4 Evaluation Plan EU IST-2001-34103 18
- Jeen Broekstra, Marc Ehrig, Peter Haase, Frank van Harmelen, Maarten Menken, Peter Mika, Bjrn Schnizler, Ronny Siebes: Bibster A Semantics- Based Bibliographic Peer-to-Peer System. The Second Workshop on Semantics in Peer-to-Peer and Grid Computing at the 12th World Wide Web Conference, New York, USA, May 18, 2004
- Christoph Tempich, Sofia Pinto, Steffen Staab, York Sure: A case study in supporting DIstributed, Loosely-controlled and evolvInG Engineering of oNTologies (DILIGENT), IKNOW 2004
- S. Pinto, S. Staab, Y. Sure, C. Tempich. OntoEdit empowering SWAP: A case study in supporting Distributed, Loosely-controlled and evolvInG Engineering of oNTologies (DILIGENT). *Proc. of 1st European Semantic Web Symposium*. Crete, May 10-12, 2004.
- Karl Aberer, Philippe Cudr-Mauroux, A. Ouksel, T. Catarci, M.-S. Hacid, A. Illarramendi, V. Kashyap, M. Mecella, E. Mena, E. Neuhold, O. De Troyer, T. Risse, M. Scannapieco, F. Saltor, L. De Santis, S. Spaccapietra, S. Staab, R. Studer. *Emergen Semantics Principles and Issues*. In Y. Lee et al. (Eds.) *Proceedings of Database Systems for Advances Applications, 9th International Conference, DASFAA 2004*, Jeju Island, Korea, March 17-19, 2004, LNCS 2973 Springer. pp. 25-38.

- Christoph Schmitz, Steffen Staab, Christoph Tempich: Socialisation in Peer-to-Peer Knowledge Management, IKNOW 2004
- E. Onaindia, S. Staab (eds.). Proceedings of STAIRS-2004, the Starting AI Researchers Symposium. IOS Press, 2004.
- S. Staab, R. Studer (eds.). on Ontologies. International s on Information Systems, Springer Verlag, 2004.
- R. Volz, S. Handschuh, S. Staab, L. Stojanovic, N. Stojanovic. Unveiling the Hidden Bride: Deep Annotation for Mapping and Migrating Legacy Data to the Semantic Web. Journal of Web Semantics: Science, Services and Agents on the World Wide Web. Elsevier. Appears 2004.
- Peter Haase, Ronny Siebes, Frank Harmelen, Peer Selection in Peer-to-Peer Networks with Semantic Topologies, In International Conference on Semantics of a Networked World: Semantics for Grid Databases, 2004, Paris. 2004.
- Peter Haase, Marc Ehrig, Andreas Hotho, Björn Schnizler: Personalized Information Access in a Bibliographic Peer-to-Peer System, In Proceedings of the AAAI Workshop on Semantic Web Personalization, 2004. July 2004.
- Philipp Cimiano, Alexander Pivk, Lars Schmidt-Thieme, Steffen Staab. Learning Taxonomic Relations from Heterogeneous Evidence. Proc. of ECAI- 2004 workshop on Ontology Learning and Population, Valencia, Spain, August 2004.
- Marc Ehrig, Steffen Staab. Efficiency of Ontology Mapping Approaches. Proc. of Semantic Intelligent Middleware for the Web and Grid. Workshop at ECAI- 2004, Valencia, August 2004
- Marc Ehrig: Ontology Mapping Experiments: I3CON workshop at the Performance Metrics for Intelligent Systems Workshop (PerMIS 04), Gaithersburg, MD, USA, August 2004
- Marc Ehrig, Steffen Staab. QOM Quick Ontology Mapping (Extended Abstract). Workshop Informationsfusion at the Informatik 2004 in Ulm, September 21, 2004. D7.4 Evaluation Plan EU IST-2001-34103 19
- Marc Ehrig, Steffen Staab: QOM - Quick Ontology Mapping. 3rd International Semantic Web Conference (ISWC 2004), Hiroshima, November 2004
- Peter Haase, Jeen Broekstra, Andreas Eberhart, Raphael Volz: A Comparison of RDF Query Languages. ISWC 2004, Hiroshima, November 2004
- Peter Haase, Jeen Broekstra, Marc Ehrig, Maarten Menken, Peter Mika, Mariusz Olko, Michal Plechawski, Pawel Pyszlak, Björn Schnizler, Ronny Siebes, Steffen Staab, Christoph Tempich: Bibster A Semantics-based Bibliographic Peer-to-Peer System. ISWC 2004, Hiroshima, November 2004

- Ronny Siebes, Frank van Harmelen: Ranking Agent Statements for Building Evolving Ontologies Proceedings of the AAAI-02 Workshop on Meaning Negotiation
- Heiner Stuckenschmidt: Adapting Communication Vocabularies using Shared Ontologies, Proceedings of the Workshop on Ontologies and Agent Systems at AAMAS02
- Heiner Stuckenschmidt: Exploiting Partially Shared Ontologies for Multi-Agent Communication, Proceedings of the International Workshop on Cooperative Information Agents CIA02.
- H. Stuckenschmidt Query Processing on the Semantic Web in KI - Special Issue on the Semantic Web, 2003.
- H. Stuckenschmidt and F. van Harmelen and F. Giunchiglia Query Processing in Ontology-Based Peer-To-Peer Systems. Technical Report DIT-02-096, Informatica e Telecomunicazioni, University of Trento, 2002.
- H. Stuckenschmidt and F. van Harmelen Approximating Terminological Queries Proceedings of the Fifth International Conference on Flexible Query Answering Systems. October 27 - 29, 2002, Copenhagen, Denmark
- H. Stuckenschmidt Approximate Information Filtering for the Semantic Web Proceedings of the 25th German Conference on Artificial Intelligence
- Paolo Bouquet, Fausto Giunchiglia, Frank van Harmelen, Luciano Serafini, Heiner Stuckenschmidt C-OWL: Contextualizing Ontologies. International Semantic Web Conference, Sanibal Island, FL 2003.
- Arthur Ouwerkerk, Heiner Stuckenschmidt: Visualizing RDF Data for P2P Information Sharing. Proceedings of the K-CAP03 Workshop on Visualization for Interactive Knowledge Engineering, Sanibal Island, FL, 2003.
- Jeen Broekstra, Marc Ehrig, Peter Haase, Frank van Harmelen, Maarten Menken, Peter Mika, Bjorn Schnizler, Ronny Siebes "Bibster - A Semantics- Based Bibliographic Peer-to-Peer System" The Second Workshop on Semantics in Peer-to-Peer and Grid Computing at the Thirteenth International World Wide Web Conference 18 May 2004, New York, USA
- Peter Haase, Ronny Siebes, Frank van Harmelen "Peer Selection in Peer-to-Peer Networks with Semantic Topologies" at the International Conference on Semantics of a Networked World, June 2004, Paris, France
- Gergely Adamku: Implementation and Evaluation of Distributed RDF Processing. Masters Thesis. AI Department, Vrije Universiteit Amsterdam 2004.

- Heiner Stuckenschmidt, Richard Vdovjak, Geert-Jan Houben and Jeen Broekstra: Data Structures and Algorithms for Querying Distributed RDF Repositories. In Proceedings of the 13th International World Wide Web Conference, New York, 2004.
- H. Stuckenschmidt, F. van Harmelen, L. Serafini, P. Bouquet and F. Giunchiglia, Using C-OWL for the Alignment and Merging of Medical Ontologies. Proceedings of the KR-MED workshop at KR 2004.
- H. Stuckenschmidt, M. Sabou and M. Klein; The Semantic Web Bringing meaning to distributed systems. Invited feature article. Distributed Systems Online. 2004.
- Paolo Bouquet, Fausto Giunchiglia, Frank van Harmelen, Luciano Serafini, Heiner Stuckenschmidt Contextualizing Ontologies, Journal on Web Semantics, to appear 2004.
- H. Stuckenschmidt and R. Vdovjak and Geert-Jan Houben and Jeen Broekstra: Towards Distributed Processing of RDF Path Queries. International Journal of Web Engineering and Technology, to appear 2005.
- S. Staab, H. Stuckenschmidt (eds.) . Semantic Web and Peer-to-peer. Springer Verlag, to appear March 2005.
- Pedro Vaquero. “Una red ideada para mquinas “. In the economic newspaper 5DIAS, Pedro Vaquero, explained in a personal interview that our objective is to advance in semantic web, talking about Meta4s European projects.
- Pedro Vaquero. ” La semntica abre nuevos retos a la gestin del capital humano” (Semantics open new challenges in Intellectual Capital Management) appeared in ComputerWorld software magazine on 3 April 2003 (Spanish edition) This article exposes the most important goals of the SWAP Project.
- Marc Sabagh: “El sistema de RRHH constituye el corazn de muchas aplicaciones” In Computing magazine on 30 June 2004, Marc Sabbagh Meta4 R&D VP, explains how Meta4 is investing in Web Semantic, for its KM and HR applications.
- Oktober issue of “Teleinfo” magazine contains interview with Mariusz Olko and an article exclusively devoted to SWAP project, its objectives and solutions proposed.
- Two papers were submitted to conferences (WI2003 and Iknow 2003 respectively). Both cover aspects of knowledge management applications using ontologies, information extraction, and natural language processing. The SWAP project is described as one major building block for a “minimal invasive” knowledge management as proposed by our group. In short words, the main issue of this approach is to technically enable knowledge management with as little impact on daily routine as possible.

- In Q4 2003 a book on Text Mining and Its Applications (Zanasi, A. (Edt). Publ. Date: 2003: “Text Mining and Its Applications”. -Computational Mechanics- ISBN: 1-85312-995-X) will be published containing a chapter on “Text Mining based Knowledge Management in Banking“. This chapter was written in Feb/March 2003 and builds on the SWAP framework as one building block of a Knowledge Management Workbench“.
- “Diario de Mallorca” (newspaper), 01.03.2003
- “Televisin Espaola regional” (Spanish TV regional connexion), 4.3.2003
- “Quaderns dInnovaci i Energia, num 7”
- (Innovation & Energy Bulletin. Balearic Government publication), April 2003
- “Quaderns dInnovaci i Energia, num 8” , Innovation & Energy Bulletin. Balearic Government publication, May 2003
- Distribution list “Bitcora”, num 37, June 2003

A.3 XAROP — A Report on Introducing a Decentralized Semantics-based Knowledge Sharing Application

A.3.1 Abstract

Knowledge management solutions relying on central repositories sometimes have not met expectations, since users often create knowledge ad-hoc using their individual vocabulary and using their own individual IT infrastructure (e.g., their laptop). To improve knowledge management for such decentralized and individualized knowledge work, it is necessary to, first, provide a corresponding decentralized IT infrastructure and to, second, deal with specific problems such as security and semantic heterogeneity. In this paper, we describe the technical peer-to-peer platform that we have built and summarize some of our experiences applying the platform in case study for cooptitioning organizations in the tourism sector.

A.3.2 Introduction

Knowledge management solutions relying on central repositories sometimes have not met expectations, since users often create knowledge ad-hoc using their individual vocabulary and using their own individual IT infrastructure (e.g., their laptop). [BBT02] provide an explanation for the failed cases. They argue that traditional knowledge management systems take on a traditional managerial control paradigm, while subjectivity and sociality are essential features of knowledge creation and sharing. From that point they overhaul the traditional architecture of knowledge management systems towards a more decentralized architecture.

This general observation is supported by our own case study, which we pursued in the course of the SWAP project¹. The case study is in the tourism domain of the Balearic Islands. The needs of the tourism industry there are best described by the term ‘cooptition’. On the one hand the different organizations *compete* for customers against each other. On the other hand, they must *cooperate* in order to provide high quality for regional touristic issues like infrastructure, facilities, clean environment, or safety — that are critical for them to be able to compete against other tourism destinations. Working based on traditional, centralized knowledge management systems is infeasible, since no single organization can control all processes.

Although the need for decentralized knowledge management solutions is obvious, the field of solutions is still very limited. Two technologies are currently emerging as candidate solutions. On the communication level peer-to-peer (P2P) networks provide the means to connect the different participants (*cf.* [BBD⁺04]) while ontologies can provide the necessary expressivity on the representation level [D. 98].

¹<http://swap.semanticweb.org/>

As for peer-to-peer networks one can distinguish three separate levels: *viz.* infrastructure, application and community level(*cf.* [SF03]). The P2P infrastructure provides basic mechanisms to communicate, security mechanisms, resource identification and peer identification. The P2P application provides services to the users supporting their process needs. P2P communities comprise the social activities which are enabled by the P2P paradigm.

Our use of ontology-based knowledge representation lies orthogonal to the three levels of the P2P network. The use of semantic descriptions of data sources stored by peers and indeed of semantic descriptions of peers themselves alleviates current problems on all three levels as we will show in the remainder.

In this paper we describe our experiences in introducing a peer-to-peer based knowledge management application. The application is built around and takes advantage of ontologies at the three different levels mentioned before. Before we start to explain this interaction, we briefly introduce the organizational setting of our case study (Section A.3.3). From the case study we have derived several technical requirements (Section A.3.4) for our application. The requirements led us first to an architecture for a semantically enriched peer-to-peer application (Section A.3.5), and second to a number of new methods (Section A.3.6). We finalize the paper with a summary of the case study (Section A.3.7) and lessons learned (Section A.3.7), a reference to the related work (Section A.3.8) and the conclusion (Section A.3.9).

A.3.3 Organizational Context

The case study we consider here is based in the tourism sector of the Balearic Islands. A number of organizations participating in the case study want to collaborate on some regional issues. Therefore they now collect and share information about *indicators* reflecting the impact of growing population and tourist fluxes in the islands, their environment and their infrastructures. Moreover, these indicators can be used to make predictions and help planning. For instance, organizations that require *Quality & Hospitality management* use the information to better plan, for example, their marketing campaigns. As another example, a governmental agency, a Balearic Government's co-ordination center of telematics, provides the local industry with information about *new technologies* that can help the tourism industry to better perform their tasks.

Due to the different working areas and objectives of the collaborating organizations, it proved impossible to set up a centralized knowledge management system or even a completely centralized ontology. The case study partners asked explicitly for a system without a central server, where knowledge sharing is integrated into the normal work, but where very different kinds of information could be easily shared with others.

A.3.4 Requirements

From a technical point of view, the different organizations can be seen as one or many independently operating nodes within a “knowledge” network. Nodes can join or disconnect from the network at any moment and can live or act independently of the behavior of other nodes in the system. A node may perform several tasks. The most important one is that it acts as a peer in the network, so it can communicate with other nodes to achieve its goals. But apart from that it may act as an interface to interact with the human user of the network, or it may access knowledge sources to accomplish its tasks. One node may have one or more knowledge sources associated with it. These sources contain the information that a peer can make available to other peers. Examples are a user’s filesystem and mail folders or a locally installed database.

A node must be designed to meet the following requirements that arise from the task of sharing information from the external sources with other peers: *Integration*: Each piece of knowledge requires metadata about its origin. To retrieve external information, the metadata needs to capture information about where the piece of information was obtained from. This information will allow to identify a peer and locate resources in its repositories. *Information heterogeneity*: As each peer may use its own local ontology, the distributed information is inherently heterogeneous. Mappings may be required, e.g. to overcome the heterogeneous labelling of the same objects. However, in most cases some of the defined structures are very similar to each other. A general process is needed to identify commonalities and make them explicit.

Security: Some information may be of private nature and should not be visible to other peers. Other information may be restricted to a specific set of peers.

Presentation: In a peer-to-peer network queries are forwarded to different peers. Due to different network latencies and resources on the answering machines, answers can come at any time. Hence, the interface must help the user to distinguish between recent and old results, must update itself from time to time and should visualize where results come from.

Network efficiency: In peer-to-peer systems a general problem is, to distribute the queries in the network. Since, the number of messages increases exponentially with the number of hops a query is allowed to travel, intelligent query routing algorithms are required when the size of the network grows.

A.3.5 The SWAP platform

In the SWAP project, we have build the generic platform SWAPSTER to account for the general need of sharing semantic-based information in P2P fashion². Based on SWAPSTER we have developed a semantic and P2P based knowledge management solution appropriate for the case study sketched above. The latter is called XAROP i.e. Catalan

²Bibster [HBE⁺04] is another solution for a different case

for syrup.

XAROP nodes wrap knowledge from their local sources (files, e-mails, etc.). Nodes ask for and retrieve knowledge from their peers. For communicating knowledge, XAROP transmits RDF structures [KC03], which are used to convey conceptual structures (e.g., the definition of what an indicator for airtravel is) as well as corresponding data (e.g., data about the number of arrivals by plane). For structured queries as well as for keyword queries, XAROP uses SeRQL, an SQL-like query language that allows for queries combining the conceptual and the data level and for returning newly constructed RDF-structures.

In the following we describe only the XAROP components that we refer to later in this document.

Knowledge Sources: Peers may have local sources of information such as the local file system, e-mail directories, local databases or bookmark lists. These local information sources represent the peer's body of knowledge as well as its basic vocabulary. These sources of information are the place where a peer can physically store information (documents, web pages) to be shared on the network.

Knowledge Source Integrator: The Knowledge Source Integrator is responsible for the extraction and integration of internal and external knowledge sources into the Local Node Repository. This task comprises (1) means to access local knowledge sources and extract an RDF(S) representation of the stored knowledge, (2) the selection of the RDF statements to be integrated into the Local Node Repository and (3) the annotation of the statements with metadata.

Local Node Repository: The local node repository stores all information and its meta information a peer wants to share with remote peers. It allows for query processing and view building. The repository is implemented on top of Sesame [BKv02].

User Interface: The User Interface of the peer provides individual views on the information available in local sources as well as on information on the network. The views can be implemented using different visualization techniques (topic hierarchies, thematic maps, etc). One part of the user interface is the *Edit* component. The *Edit* component allows the user to supervise the mapping process and enables light weight ontology engineering.

Communication Adapter: This component is responsible for the network communication between peers. Our current implementation of the Communication Adapter is build on the JXTA framework [Gon01].

Information and Meta-information. Information is represented as RDF(S) statements in the repository. The SWAP meta model³ (cf. [EHvH⁺03]) provides meta-information about the statements in the local node repository in order to memorize where the statements came from and other meta-information.

Besides the SWAP meta data model the SWAP environment builds on the SWAP

³<http://swap.semanticweb.org/2003/01/swap-peer#>

common ontology⁴. The SWAP common model defines concepts for *e.g.* File and Folder. Purpose of these classes is to provide a common model for information usually found on a peer participating in a knowledge management network.

Querying for Data. SeRQL[Bro03] is an SQL like RDF query language. The main feature of SeRQL is the ability to define structured output in terms of an RDF graph that does not necessarily coincide with the model that has been queried. This feature is essential for defining personalized views in the repository of a XAROP peer.

A.3.6 Method description on the three P2P levels

In this section we explain some of the methods that are most crucial to the XAROP knowledge management solution at the infrastructure, application or community level. Because of space restriction we skip over parts that are only of longer term interest, such as scalability towards a larger P2P network, which we have explore in another case study in somewhat more depth [HBE⁺04].

Infrastructure level: A distributed Security framework

Security considerations are of particular concern within a peer-to-peer framework. Users of a P2P system want to be sure that they give access to local resources only to trusted persons. Furthermore, they want to define different access control levels to their local knowledge, since they might trust different participants to various extent. While these consideration in themselves are difficult to handle, the security mechanism in a P2P system must be straightforward to define.

For authentication we use a public-key infrastructure (PKI) infrastructure with certificate authorities established within XAROP. A certain XAROP node acts as a root certificate authority for the XAROP system, all other peers will configure this node as trusted root certificate authority. In small networks, this certificate authority will issue certificates directly for users, whereas in large networks, it is possible to build a hierarchy of certificate authorities. The certificate creation will be done offline, on the configuration level. Certificates themselves will not be transmitted within the standard SWAP user interface.

The access control model has to be based on rules, since we cannot demand from users that they will enumerate privileged users for each local resource. The rules have to base on strict facts with proven origin (*i.e.* signed by the peer that generates a fact). A simple rule gives access for a single document (*eg.* SWAP.doc) to a single, fixed person (*eg.* Esteve). More complex rules are based on knowledge about both people and resources (All people involved in SWAP project can download all documents from my SWAP folder). All access control rights have to be explicit. Otherwise we assume that access is not allowed.

Further, the right to decide about the access properties of peers can be delegated to

⁴<http://swap.semanticweb.org/2003/01/swap-common#>

other peers. A person that we delegate the right to can be fixed or described by a similar pattern, forming a chain of trust (*eg.* All people, about which Esteve said that they are involved in SWAP, and Mariusz said that they work for empolis, can download all my documents about SWAP). The access properties have to be digitally signed by the peer who assigned them.

Example for IBIT IBIT needs access control based on organization boundaries. On the other hand, it is not acceptable for an average user to be forced to create and maintain organizational information by herself. Using the described security model for the IBIT case an average user has to define (1) their administrator and (2) access control rules for his local resources. Administrators, define and maintain information about (1) organizational structures and (2) user membership to organizations.

Application level: Dealing with Semantic Heterogeneity and Visualization

Heterogeneity Semantic mapping between ontologies is a necessary precondition to establish interoperation, i.e. overcoming the heterogeneity between peers using different ontologies. When we tried to apply existing tools [NM03, DDH03] to our scenario, we found that existing mapping methods were not suitable for the ontology integration task at hand. They have laid focus exclusively on improving the effectiveness and neglected efficiency, which becomes an issue already with the ontologies required in the context of Xarop. It is not sufficient to provide its user with the best possible mapping, it is also necessary to answer his queries within a few seconds - even if the two peers use two different ontologies and have never encountered each other before.

We briefly introduce the process that our approach follows [ES04]. It is started with two ontologies, which are going to be mapped onto one another, as its input. Mapping one ontology onto another means that for each entity (concept, relation, or instance) in ontology, we try to find a corresponding entity, which has the same intended meaning, in ontology. *Feature engineering* transforms the initial representation of ontologies into a format digestible for the similarity calculations such as RDFS. In a naive approach all entities of the first ontology are compared with all entities of the second ontology (*search steps*). However, in our approach we use heuristics to lower the number of candidate mappings. In specific, this means to compare pairs with similar labels or pairs for which neighboring entities have been assigned a mapping. The *similarity computation* measures are needed to compare the features of ontological entities. The features need to be extracted from extensional and intensional ontology definitions such as URIs, RDF/S primitives such as *subclass*, or domain specific features. Extremely costly features, in terms of runtime complexity, are replaced by less costly features. Then “String Similarity” measures the similarity of two strings on a scale from 0 to 1 based on Levenshtein’s edit distance [Lev66]; “SimSet” is used to compare sets of entities based on measures used for multidimensional scaling [CC94]. In general, there are several similarity values for a candidate pair of entities. These must be aggregated into a single aggregated similarity value. This is achieved through a summarization and normalization of adjusted similarity values. A

similarity value above a certain threshold finally implies a mapping (*interpretation*). Several *iterations* of similarity calculations are needed to receive meaningful results. The returned output is a mapping table.

In first evaluation runs we have shown that our approach for identifying mappings between two ontologies is on a par with other good state-of-the-art algorithms concerning the quality of proposed mappings, while outperforming them with respect to efficiency. Using an approach combining many features to determine mappings clearly leads to significantly higher quality mappings. The here used approach is faster than standard prominent approaches by a factor of 10 to 100. This makes the presented method a valid approach for the Xarop scenario.

Visualization We have developed the Cluster Map[FSvH03], for visualizing populated, light-weight ontologies as used for XAROP. It visualizes the instances of a number of selected classes from a hierarchy, organized by their classifications. Figure A.1 shows an example Cluster Map, visualizing documents, classified according to topics discussed in those documents. The dark gray spheres represent ontology classes (the topics), with an attached label stating their name and cardinality.

When a subclass relation holds between two classes, they are connected by a directed edge. The light yellow spheres represent instances. Balloon-shaped edges connect instances to the class(es) they belong to. Instances with the same class membership are grouped in clusters. Our example contains two clusters, one of them showing overlap

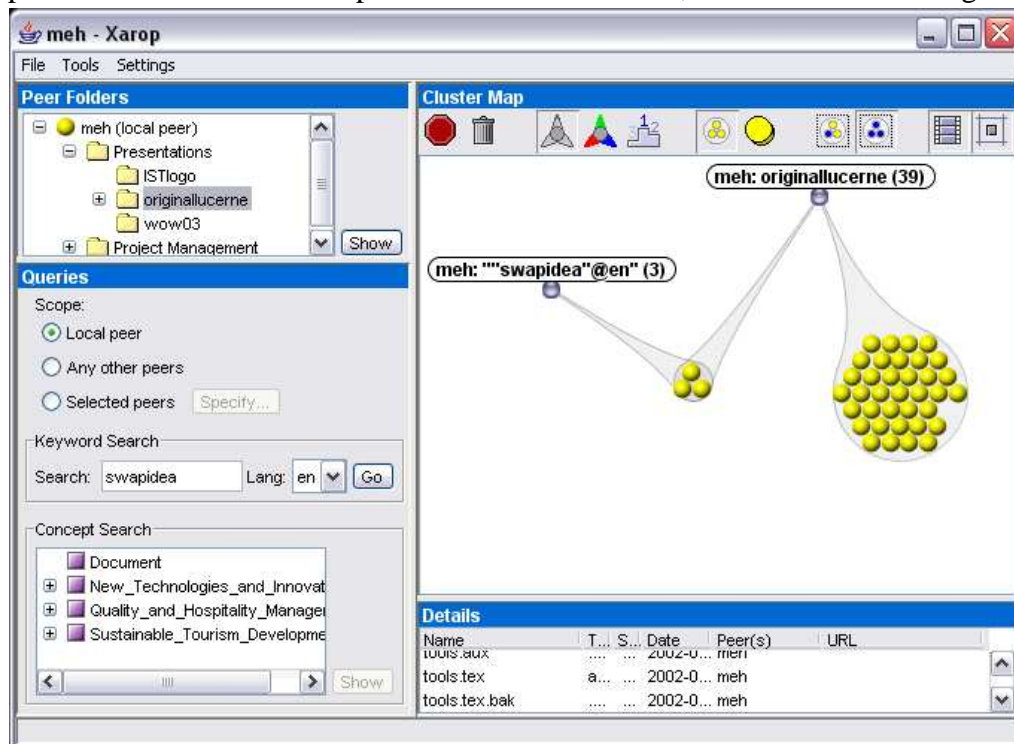


Figure A.1: Cluster Map

between the two classes. Cluster Maps contain a lot of information about the instantiation of the classes, specifically exploiting the overlaps between them. For example, figure A.1 shows that the “original lucerne” folder class has a significant overlap with the “swap idea”. Such observations can trigger hypotheses about the available information and the domain in general. The graph layout algorithm used by the Cluster Map is a variant of the well-known family of spring embedder algorithms. Its outcome results in the geometric closeness of objects indicating their semantic closeness: classes that share instances are located near each other, and so are instances with the same or similar class memberships.

The Cluster Map is embedded in a highly interactive GUI, which is designed for browsing-oriented exploration of the populated ontology. Users can subsequently create visualizations of a number of classes by marking the check boxes in the class tree on the left pane. The software can animate the transition from one visualization to the next, showing how the instances are regrouped in clusters. Through interaction a user can also retrieve information about the specific documents that are contained in a class or cluster. Further the visualization can be fine-tuned in several ways, in order to support certain tasks, or improve scalability. In the context of XAROP it was important to account for the particularities of P2P systems. Hence, the results are marked with the peer name they are coming from. Results are added to the Cluster Map incrementally, since not all peers answer at the same time. New results are highlighted. The search can be stopped when the user is satisfied. Thus it provides the usability requirements needed for the XAROP system.

Community level: The Distributed Ontology Engineering Process

Every participant in the XAROP network is allowed to structure his knowledge according to his needs. However, as we found in the case study, people working on the same issue have very similar ways of structuring information. Hence, a first step towards community building is to raise awareness about the existing commonalities within the group. From an ontological perspective that is equivalent to the agreement on a shared ontology. To enable the detection and building of shared ontologies we have defined a new ontology engineering process template *viz.* DILIGENT. It is important to note that the purpose of this process is not to agree on a conceptual model for the entire domain, but to find the subset of that model which is implicitly already agreed on. We introduce a board in charge of analyzing local ontologies and defining shared ones. This means that the participants can and should change the shared ontology after its publication. The DILIGENT process focuses in contrast to known ontology engineering methodologies available in the literature [GPFLC03] on distributed ontology development involving different stakeholders, who have different purposes and needs and who usually are not at the same location.

We will now describe the general process, roles and functions in the DILIGENT process (cf. [PSST04]). It comprises five main activities: (1) **build**, (2) **local adaptation**, (3) **analysis**, (4) **revision**, (5) **local update** (cf. figure A.2). The process starts by having *domain experts, users, knowledge engineers* and *ontology engineers* **building** an

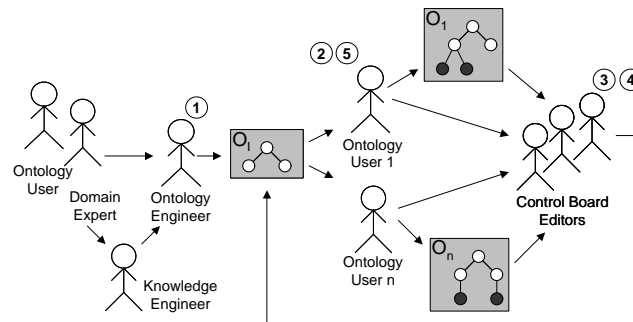


Figure A.2: Roles and functions in distributed ontology engineering initial ontology.

The team involved in building the initial ontology, *viz.* the board, should be relatively small, in order to more easily find a small and consensual first version of the shared ontology. Moreover, we do not require completeness of the initial shared ontology with respect to the domain. On the first sight it seems contradictory that the case study partners do not want to share a common infrastructure but a shared ontology. However, the existence of a shared domain, overlapping or related competencies within organizations and the need for carrying out certain functions in cooperation suggests that it is possible to develop ontologies shared by the different sub-communities.

Once the product is made available, users can start using it and **locally adapting** it for their own purposes. Typically, due to new business requirements, or user and organization changes, their local ontologies evolve in a similar way as folder hierarchies in a file system. In their local environment they are free to change the reused shared ontology. However, they are not allowed to directly change the ontology shared by all users. Furthermore, the control board collects change requests to the shared ontology.

The board **analyzes** the local ontologies and the requests and tries to identify similarities in the users' ontologies. Since not all of the changes introduced or requested by the users will be introduced,⁵ a crucial activity of the board is deciding which changes are going to be introduced in the next version of the shared ontology. The input from users provides the necessary arguments to underline change requests. A balanced decision that takes into account the different needs of the users and meets user's evolving requirements⁶ has to be found. The board should regularly **revise** the shared ontology, so that local ontologies do not diverge too far from the shared ontology. Therefore, the board should have a well-balanced and representative participation of the different kinds of participants involved in the process.

Once a new version of the shared ontology is released, users can **update** their own **local** ontologies to better use the knowledge represented in the new version. Even if the differences are small, users may rather reuse *e.g.* the new concepts instead of using their previously locally defined concepts that correspond to the new concepts represented in

⁵The idea in this kind of development is not to merge all user ontologies.

⁶This is actually one of the trends in modern software engineering methodologies (see Rational Unified Process).

the new version.

A.3.7 The case study

More concretely four organizations including 21 peers took part in the case study and it is expected that the total number of organizations will grow by 7 to a total of 28 peers until September 2004.

Early experiences In a collaboration effort the KEEEx [BBD⁺04] system was distributed among the participants in an early stage of the project. In the distributed version of KEEEx users could share file, email and bookmark folders and the corresponding documents with the participants. Additionally to a keyword based query mechanism they could look at the contents of other participants folders. Furthermore they were able to provide manual mappings between their own folders and remote folders. This made it possible to query for remote documents by means of the local folder structure. From this early experiment we could draw two main conclusions which led us to our current solution. Our first conclusion was, that keyword based querying is not sufficient in the context of the case study, since three different languages were in use and the search suffered from low precision. The performance of the search was improved by the use of a folder based approach. Our second conclusion was, that some users are willing to provided mappings between their structures and remote structures, but the additional effort to keep them updated is prohibitive.

Current status With XAROP users can share different kinds of folders and the corresponding documents. In the setup phase of the application users must define which information they want to share with whom. Additionally to the security infrastructure we introduced a small but shared ontology, to reduce the cost of mapping provision. Local documents can be associated automatically and manually with the concepts of the shared ontology. Participants can thus query for the content of others also with the concepts of the ontology. The shared ontology can be extended by the participants individually. The manual association task between documents and concepts was eased in two ways. Firstly, users could associate entire folder sub trees with certain concepts. Secondly, users could import their local folder structures and use the mapping tool to search automatically for correspondences between their local structures and the shared ontology. The mapping tool can also be used to search for correspondences between local and remote structures. When all participants had worked with the system for several days the board – comprising two domain experts and two ontology engineers – came together to analyze the changes made to the common ontology. The board could identify several concepts which are shared implicitly by all users. After some discussion the board decided to slightly change the shared ontology and redistribute a new version.

Lessons learned The case study helped us to better comprehend the use of ontologies in a peer-to-peer environment in general. The first lesson we learn is that security is the single most important issue in inter-organizational knowledge exchange. Without appro-

appropriate security mechanism users are not willing to share any information. Additionally to the implemented security mechanisms the participants would also like to know, who downloads which information. The possibility to grant access to certain documents on a 'per-query' base, was also requested. They could thus decide on query time which documents to share with whom.

On the application level we made the observation that the combination of keyword based and concept based search helps users to refine their query according to their information needs. At this point the ease of use of our user interface is very important. However, we must still work on the system and make it more light weight since not all participants work with the latest available computer equipment. Our mapping algorithm provides first suggestions for possible mappings very early, although these might be revised when the algorithm proceeds. This behavior was appreciated by most users.

Regarding the community level, the shared ontology was quickly accepted. Even so our users did understand the ontology mainly as a classification hierarchy for their documents and did not create instances of the defined concepts. Our expectation that the collaborative ontology engineering effort raised community awareness was met. Some of the participants changed their own folder structures in order to adhere to the commonly defined shared ontology. Other participants extended and changed it a bit. Later on this was the input for a refinement of the original ontology.

In spite of the technical challenges, user feedback was very positive since (i) the tool was integrated into their daily work environment and could be easily used and (ii) the tool provided very beneficial support to perform their tasks.

A.3.8 Related work

Knowledge management in Peer-to-Peer systems is the topic of various active research projects. Edutella [N⁺02] provides an RDF-based infrastructure for exchanging metadata in P2P applications. The Edutella Query Service is intended to be a standardized query exchange mechanism for RDF metadata stored in distributed RDF repositories. The Edutella project focuses on the education community. The Edamok project [BBD⁺04] also deals with distributed knowledge management in Peer-to-Peer systems and provides advanced facilities for mapping knowledge structures. However, neither of the systems is accompanied with an overall process for decentralized knowledge management on the community level. [J. 03] presents a commercial P2P solution. They emphasize the organizational difficulties and the security concerns of the participants when introducing their P2P application. The system does not provide semantics-based representation of knowledge, and thus allows only for keyword-based queries.

A.3.9 Conclusion

In this paper we have described a solution for the recently recognized problem of decentralized knowledge management. We have developed a semantics based P2P knowledge sharing platform. In this platform, we introduced an appropriate security mechanism on infrastructure level. On the application level we introduced new mapping and visualization techniques. To foster community building we have defined a distributed ontology engineering process. We have described how the different methods have been applied in a concrete case study in the tourism domain. The results of our early experiments in introducing the application in the case study show the promises of our work. We could derive several lessons learned which will drive the future development of XAROP.

While we have addressed a broad range of problems in decentralized knowledge management, there are some issues to be addressed as part of future work. For example, query routing is currently not an issue within the XAROP system since the number of users is still small and simple query routing mechanism work effectively. However, in bigger networks, as they are found in the Bibster case study (the second system build on Swapster) efficient query routing is crucial. The interested reader may find some ideas of how to deal with such scalability issues in the description of the Bibster system[HBE⁺04].

A.4 Bibster — A Semantics-Based Bibliographic Peer-to-Peer System

A.4.1 Abstract

This paper describes the design and implementation of Bibster, a Peer-to-Peer system for exchanging bibliographic data among researchers. Bibster exploits ontologies in data storage, query formulation, query routing and answer presentation: When bibliographic entries are made available for use in Bibster, they are structured and classified according to two different ontologies. This ontological structure is then exploited to help users formulate their queries. Subsequently, the ontologies are used to improve query routing across the Peer-to-Peer network. Finally, the ontologies are used to post-process the returned answers in order to do duplicate detection. The paper describes each of these ontology-based aspects of Bibster. Bibster is a fully implemented open source solution built on top of the JXTA platform.

A.4.2 Introduction

The advantages of Peer-to-Peer architectures over centralized approaches have been well advertised, and to some extent realized in existing applications: no centralized server (thus avoiding a bottleneck for both computational performance and information update), robustness against failure of any single component, scalability both in data volumes and the number of connected parties.

However, besides being the solution to many problems, the large degree of distribution of Peer-to-Peer systems is also the cause of a number of new problems: The lack of a single coherent schema for organizing information sources across the Peer-to-Peer network hampers the formulation of search queries, duplication of information across the network results in many duplicate answers to a single query, and answers to a single query often require the integration of information residing at different, independent and uncoordinated peers. Finally, query routing and network topology (which peers to connect to, and which peers to send/forward queries to) are significant problems.

The research community has recently turned to the use of semantics in Peer-to-Peer networks to alleviate these problems [B⁺03], [N⁺02], [CFM⁺03]. The use of semantic descriptions of datasources stored by peers and indeed of semantic descriptions of peers themselves helps in formulating queries such that they can be understood by other peers, in merging the answers received from other peers, and in routing queries across the network. In particular, the use of ontologies and of Semantic Web technologies has been identified as promising for Peer-to-Peer systems.

The scenario that we have envisioned is that researchers share bibliographic metadata in a community with a Peer-to-Peer system. The data may have been obtained from local

BibTeX files or from bibliography servers like the DBLP database⁷ or CiteSeer⁸. As one may easily recognize, this scenario (like some others that we do not elaborate upon in this paper) exhibits two characteristics that strongly require a semantics-based Peer-to-Peer system.

First, a centralized solution does not exist and cannot exist, because of the multitude of informal workshops that researchers refer to, but that do not show up in centralized resources such as DBLP. Any such centralized resource will only cover a limited scientific community. For example, DBLP covers a lot of Artificial Intelligence, but almost no Knowledge Management, whereas a lot of work is being done in the overlap of these two fields. At the same time, many individual researchers are willing to share their resources, provided they do not have to invest work in doing so.

Second, the use of Semantic Web technology is crucial in this setting. Although a small common-core ontology of bibliographic information exists (title, author/editor, etc), much of this information is very volatile and users define arbitrary add-ons, for example to include URLs or abstracts of publications.

In this paper we will describe the design of the Bibster system and emphasize the semantic components and their use: semantic extraction of bibliographic metadata in section A.4.5, semantic querying in section A.4.6, peer selection using semantic topologies in section A.4.7, and semantic duplicate detection in section A.4.8. Furthermore, Bibster has been fully implemented and we present evaluation results from a field experiment in section A.4.9.

A.4.3 Major Bibster Use Cases

Bibster is aimed at researchers that share bibliographic metadata. Requirements for Bibster must include capabilities that support their daily work. Researchers may want to:

1. query a single specific peer (e.g. their own computer, because it is sometimes hard to find the right entry there), a specific set of peers (e.g. all colleagues at an institute) or the entire network of peers (to obtain the maximal recall at the price of low precision).
2. search for bibliographic entries using simple keyword searches, but also more advanced, semantic searches, e.g. for publications of a special type, with specific attribute values, or about a certain topic.
3. integrate results of a query into a local knowledge base for future use. Such data may in turn be used to answer queries by other peers. They may also be interested in updating items that are already locally stored with additional information about these items obtained from other peers.

⁷<http://dblp.uni-trier.de/>

⁸<http://citeseer.org/>

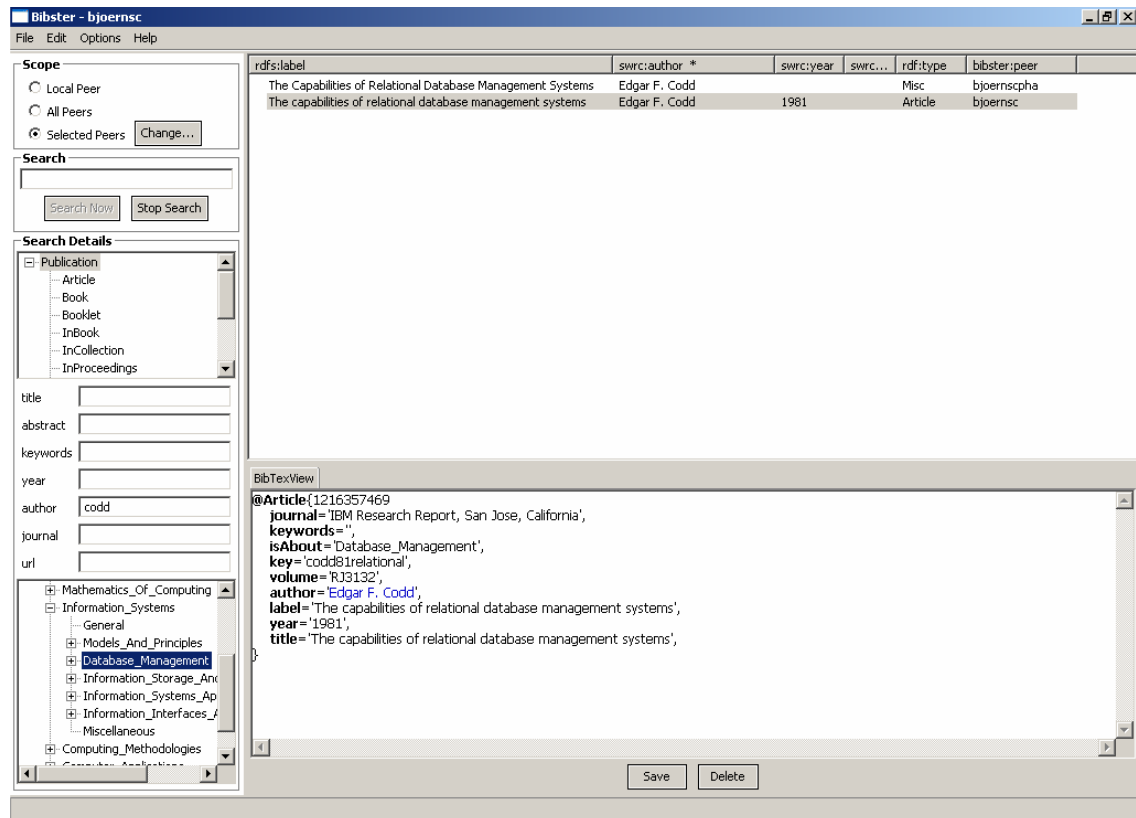


Figure A.3: Searching for publications about database management authored by Codd

The screenshot in figure A.3 partially indicates how these use cases are realized in Bibster. The *Scope* widget allows for defining the targeted peers, the *Search* and *Search Details* widgets allow for keyword and semantic search; *Results Table* and *BibtexView* widgets allow for browsing and re-using query results. The query results are visualized in a list grouped by duplicates. They may be integrated into the local repository or exported in formats such as BibTeX and HTML.

A.4.4 Design of Bibster

Ontologies in Bibster Ontologies are crucial throughout the usage of Bibster, viz. for importing data, formulating queries, routing queries, and processing answers.

Firstly, the system enables users to import their own bibliographic metadata into a local repository. Bibliographic entries made available to Bibster by a user (cf. section A.4.5) are automatically aligned to two common ontologies: The first ontology (SWRC⁹) describes different generic aspects of bibliographic metadata (and would be valid across

⁹<http://www.semanticweb.org/ontologies/swrc-onto-2001-12-11.daml>

many different research domains), the second ontology (ACM Topic Hierarchy¹⁰) describes specific categories of literature for the Computer Science domain.

Secondly, queries are formulated in terms of the two ontologies: Queries may concern fields like author, publication type, etc. (using terms from the SWRC ontology) or queries may concern specific Computer Science terms (using the ACM Topic Hierarchy).

Thirdly, queries are routed through the network depending on the expertise models of the peers describing which concepts from the ACM ontology a peer can answer queries on. A matching function determines how closely the semantic content of a query matches the expertise model of a peer. Routing is then done on the basis of this semantic ranking.

Finally, answers are returned for a query. Due to the distributed nature and potentially large size of the Peer-to-Peer network, this answer set might be very large, and contain many duplicate answers. Because of the semistructured nature of bibliographic metadata, such duplicates are often not exactly identical copies. Ontologies help to measure the semantic similarity between the different answers and to remove apparent duplicates as identified by the similarity function.

Bibster Architecture and Modules The Bibster system has been implemented as an instance of the SWAP System architecture as introduced in [B⁺03]. Figure A.4 shows a high-level design of the architecture of a single node in the Peer-to-Peer system. We will now briefly present the individual components as instantiated for the Bibster system.

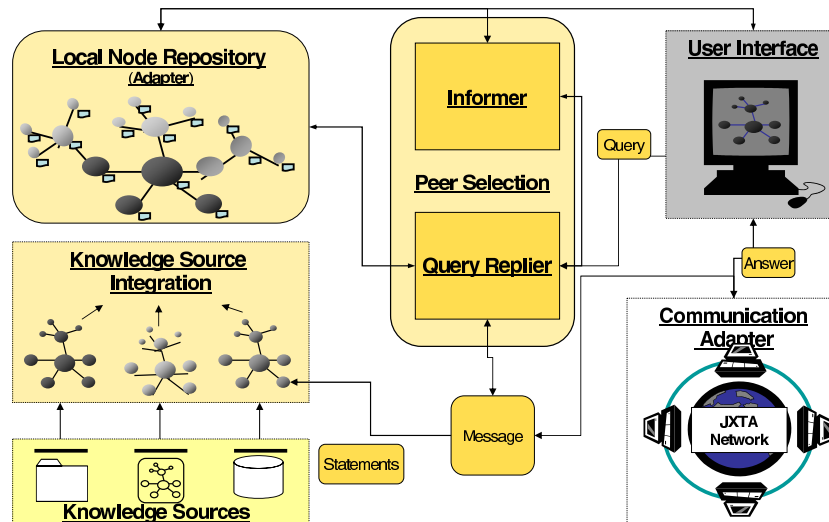


Figure A.4: SWAP System Architecture

Communication Adapter: This component is responsible for the network communication between peers. It serves as a transport layer for other parts of the system, for sending and forwarding queries. It hides and encapsulates all low-level communication details

¹⁰<http://www.acm.org/class/1998/>

from the rest of the system. In the specific implementation of the Bibster system we use JXTA as the communication platform.

Knowledge Sources: The knowledge sources in the Bibster system are sources of bibliographic metadata, such as BibTeX files stored locally in the file system of the user.

Knowledge Source Integrator: The Knowledge Source Integrator is responsible for the extraction and integration of internal and external knowledge sources into the Local Node Repository. In section A.4.5 we describe the process of semantic extraction from BibTeX files. In section A.4.8 we explain how the knowledge of local and remote sources can be merged, i.e. how duplicate query results are detected.

Local Node Repository: In order to manage its information models and views as well as information acquired from the network, each peer maintains a Local Node Repository providing the following functionality: (1) Mediate between views and stored information, (2) support query formulation and processing, (3) specify the peer's interface to the network, and (4) provide the basis for peer ranking and selection. In the Bibster system, the Local Node Repository is based on the RDF-S Repository Sesame [BKvH01]. The query language SeRQL is used to formulate semantic queries against the Local Node Repository, as described in section A.4.6.

Informer: The task of the Informer is to proactively advertise the available knowledge of a peer in the Peer-to-Peer network and to discover peers with knowledge that may be relevant for answering the user's queries. This is realized by sending advertisements about the expertise of a peer. In the Bibster system, these expertise descriptions contain a set of topics that the peer is an expert in. Peers may accept – i.e. remember – these advertisements, thus creating a semantic link to the other peer. These semantic links form a semantic topology, which is the basis for intelligent query routing (cf. section A.4.7 for details).

Query Replier: The Query Replier is the coordinating component controlling the process of distributing queries. It receives queries from the User Interface or from other peers. Either way it tries to answer the query or distribute it further according to the content of the query. The decision to which peers a query should be sent is based on the knowledge about the expertise of other peers.

User Interface: The User Interface (Figure A.3) allows the user to import, create and edit bibliographic metadata as well as to easily formulate queries.

A.4.5 Semantic Extraction of Bibliographic Metadata

Many researchers have accumulated extensive collections of BibTeX files for their bibliographic references. However, these files are semi-structured and thus single attributes may be missing or may not be interpreted correctly. For interchanging bibliographic data in a semantically based Peer-to-Peer network it has to be represented in a structured and formal way. *BibToOnto* is a component of Bibster for extracting explicit knowledge of

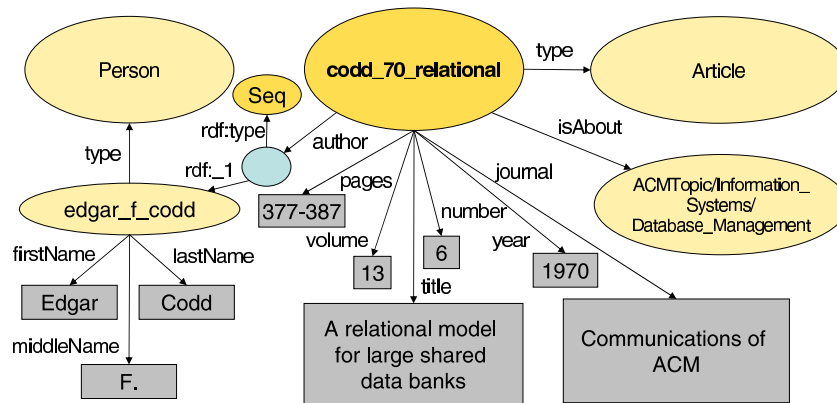


Figure A.5: SWRC Sample Metadata

bibliographic items. Plain BibTeX files are transformed into an ontology based knowledge representation.

The target ontology is the Semantic Web Research Community Ontology (SWRC), which models among others a research community, its researchers, topics, publications, tools, and properties between them. The SWRC ontology defines a shared and common domain theory which helps users and machines to communicate concisely and supports the exchange of semantics.

BibToOnto automatically classifies bibliographic entries according to the ACM topic hierarchy. Additionally, it is possible to reclassify the entries manually in the user interface of Bibster. The ACM topic hierarchy is a standard schema for describing and categorizing computer science literature. It covers 1287 topics of the computer science domain. In addition to the sub- and supertopic relations, it also provides information about related topics.

The following example shows a transformation of a BibTeX entry to a SWRC ontology based item. The result¹¹ is represented as an RDF graph in figure A.5.

```
@ARTICLE{codd70relational
  author    = "Edgar F. Codd",
  year      = "1970",
  title     = "A relational model for large shared data banks",
  journal   = "Communications of ACM",
  volume    = "13",
  number    = "6",
  pages     = "377--387" }
```

¹¹For better readability we used a concatenation of the author name and the title of the publication as a URI in this example. In the Bibster system however we calculate hash codes over all attribute values to guarantee the uniqueness of URIs.

A.4.6 Semantic Querying

Each peer node in the Bibster system manages a local RDF repository with bibliographic data extracted by BibToOnto or integrated from other peers. The query language interface to the local RDF repository is SeRQL [BK04].

SeRQL (Sesame RDF Query Language) is an RDF/RDF-S query language that was developed in the context of the SWAP project to address practical requirements that were not sufficiently met by other query languages.

SeRQL Design Principles. Several characteristics are urgently required in the context of Bibster (but also of many other systems) for an RDF/RDF-S query language like SeRQL. In particular, it must:

1. be functional such that each query returns a RDF graph, which may be integrated into the local repository or queried again,
2. be aware of the (optional) schema,
3. let the user formulate path expressions for navigating the RDF graph, e.g. the combination of SWRC and ACM topic hierarchy,
4. be able to deal with *optional* values, e.g. a publisher field may be given or not.

Without showing all capabilities of SeRQL in full detail¹², we briefly show how SeRQL queries are composed, and how tasks in the Bibster system are performed using SeRQL.

(1) SeRQL uses a `select-from-where` or `construct-from-where` filter, where the `select` or `construct` clauses specify projections, the `from` clause specifies a graph match template (by means of path expressions), and the `where` clause allows the definition of additional boolean constraints on matched values in the path expressions.

(2) SeRQL takes the RDF schema into account by mapping from the given graph to its formal model.

(3) When navigating the RDF graph, SeRQL exploits the formal semantics of path labels. For example, `<rdfs:subClassOf>` is interpreted as a reflexive transitive relation and upward inheritance of instances is interpreted (through the `<rdf:type>` relation (cf. [BK04] for full details).

(4) Bibtex entries may be incomplete. SeRQL allows to distinguish between optional and required elements in the query and, hence, is flexible enough to deal with these circumstances.

A Querying Scenario. In our running example, a researcher is querying for journal articles written by the author Codd about database management. Internally, this request is formulated as a SeRQL query that looks as follows:

¹²See <http://www.openrdf.org/doc/SeRQLmanual.html> for a complete overview.

```

construct distinct
  {s} prop {val};
  <rdf:type> {t};
  <swrc:author> {x} <rdf:type> {<rdf:Seq>};
  <rdfs:member> {author} prop_author {val_author}
from
  {s} <serql:directType> {t};
  <rdf:type> {<swrc:Article>};
  prop {val};
  <swrc:isAbout> {<acm:ACMTopic/Information_Systems/Database_Management>};
  <swrc:author> {x} <rdfs:member> {} <swrc:lastName> {lname},
  [{x} <rdfs:member> {author} prop_author {val_author} ]
where prop != <rdf:type> and lname like "Codd"
using namespace
  swrc = <!http://www.semanticweb.org/ontologies/swrc-onto-2001-12-11.daml#>,
  acm = <!http://daml.umbc.edu/ontologies/classification#>

```

Compare the structure of the from-clause to the representation of the RDF graph given in figure A.5. The from-clause retrieves not only the identifier for the particular journal entry (“codd_70_relational”, matched by s), but also the graph structure surrounding it, which essentially gives the entry its meaning: the name of the author, the type of publication, the year it was published, the number of pages, etc. Also, if the first and middle names of an author are known, the query retrieves those (but it does not fail if these are not known).

The use of schema-awareness is evident in the use of typing information on s : s need not only be of type `swrc:Article`, we also retrieve its *specific* (or *direct*) type. Being functional plays a role as well: A graph transformation is used to create a query result that can be easily processed to be given back to the user through the GUI.

A.4.7 Expertise Based Peer Selection

The scalability of a Peer-to-Peer network is essentially determined by the way how queries are propagated in the network. Peer-to-Peer networks that broadcast all queries to all peers do not scale – intelligent query routing and network topologies are required to be able to route queries to a relevant subset of peers that are able to answer the queries.

Modern routing protocols like Chord [SMK⁺01] and CAN [RFH⁺01] allow for sophisticated query routing based on distributed indices. More recently, in the Semantic Web context, schema based Peer-to-Peer networks such as the one described in [N⁺03] have emerged based on complex, extensible semantic descriptions of resources. They allow for complex queries against these metadata instead of simple keyword-based queries. Another semantic-based approach is pSearch [TXM02], a decentralized non-flooding Peer-to-Peer information retrieval system. pSearch distributes document indices through the Peer-to-Peer network based on document semantics generated by Latent Semantic Indexing (LSI). The search cost (in terms of nodes searched and data transmitted) for a given query is thereby reduced, since the indices of semantically related documents are likely to

be co-located in the network. Here we give an overview of the model of expertise based peer selection as proposed in [HSvH04] and how it is used in the Bibster system. In this model, peers use a shared ontology to advertise semantic descriptions of their expertise in the Peer-to-Peer network. The knowledge about the expertise of other peers forms a semantic topology, independent of the underlying network topology. If the peer receives a query, it can decide to forward it to peers about which it knows that their expertise is similar to the subject of the query. The advantage of this approach is that queries will not be forwarded to all or a random set of known peers, but only to the ones that have a good chance of answering it.

Semantic Description of Expertise. Peers. The Peer-to-Peer network consists of a set of peers P . Every peer $p \in P$ has a Local Node Repository, which stores the bibliographic metadata.

Common Ontology. The peers share an ontology O , which is used for describing the expertise of peers and the subject of queries. In our case, O is the ACM topic hierarchy that contains a set of topics T .

Expertise. An expertise description is an abstract, semantic description of the Local Node Repository of a peer based on the shared ontology O . The expertise E of a peer is thus defined as $E \subseteq 2^T$, where each $e \in E$ denotes a set of ACM topics, for which a peer provides classified instances.

Advertisements. Advertisements $A \subseteq P \times E$ are used to promote descriptions of the expertise of peers in the network. An advertisement $a \in A$ associates a peer p with an expertise e . Peers decide autonomously, without central control, whom to promote advertisements to and which advertisements to accept. This decision is based on the semantic similarity between expertise descriptions.

Matching and Peer Selection. Queries. Queries $q \in Q$ are posed by a user and are evaluated against the local node repositories of the peers. First a peer evaluates the query against its local node repository and then decides which peers the query should be forwarded to.

Subjects. A subject is an abstraction of a given query q expressed in terms of the common ontology. The subject specifies the required expertise to answer the query. In our scenario, the subjects of queries are defined as $S \subseteq 2^T$, each s is the set of ACM topics that are referenced in the query. E.g., the extracted subject of the query in example A.4.6 would be *Information Systems/Database Management*.

Similarity Function. The similarity function $Sim : S \times E \mapsto [0, 1]$ yields the semantic similarity between a subject $s \in S$ and an expertise description $e \in E$. An increasing value indicates increasing similarity. If the value is 0, s and e are not similar at all, if the value is 1, they match exactly. Sim is used for determining to which peers a query should be forwarded. In Bibster, the similarity function Sim_{Topics} is based on the idea

that topics which are close according to their positions in the topic hierarchy are more similar than topics that have a larger distance. For example, an expert on the ACM topic *Information Systems/Information Storage and Retrieval* has a higher chance of giving a correct answer on a query about *Information Systems/Database Management* than an expert on a less similar topic like *Hardware/Memory Structures*.

To be able to define the similarity of a peer's expertise and a query subject, which are both represented as a set of topics, we first define the similarity for individual topics. [LBM03] have compared different similarity measures and have shown that for measuring the similarity between concepts in a hierarchical structured semantic network, like the ACM topic hierarchy, the following similarity measure yields the best results:

$$sim_{Topic}(t_1, t_2) = \begin{cases} e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} & \text{if } t_1 \neq t_2, \\ 1 & \text{otherwise} \end{cases}$$

Here l is the length of the shortest path between topic t_1 and t_2 in the graph spanned by the *SubTopic* relation. h is the level in the tree of the direct common subsumer from t_1 and t_2 . $\alpha \geq 0$ and $\beta \geq 0$ are parameters scaling the contribution of shortest path length l and depth h , respectively. Based on their benchmark data set, the optimal values are: $\alpha = 0.2$, $\beta = 0.6$.

Peer Selection Algorithm. The peer selection algorithm returns a ranked set of peers, where the rank value is equal to the similarity value provided by the similarity function. Therefore, peers that have an expertise more similar to that of the subject of the query will have a higher rank. From this set of ranked peers one can, for example, select the best n peers, or all peers whose rank value is above a certain threshold. In the Bibster system we select the best n peers that have not yet received the query along the message path, where n can be specified. A maximum number of hops (set to 4 for the field experiment) further limits the forwarding of queries.

Semantic Topology

The knowledge of the peers about the expertise of other peers is the basis for a semantic topology. Here it is important to state that this semantic topology is independent of the underlying network topology. At this point, we do not make any assumptions about the properties of the topology on the network layer.

The semantic topology can be described by the following relation:

$$Knows \subseteq P \times P, \text{ where } Knows(p_1, p_2) \text{ means that } p_1 \text{ knows about the expertise of } p_2.$$

The relation *Knows* is established by the selection of which peers a peer sends its advertisements to. Furthermore peers can decide to accept an advertisement, e.g. to include it in their registries, or to discard the advertisement. The semantic topology in combination with the expertise based peer selection is the basis for intelligent query routing.

A.4.8 Semantic Duplicate Detection

When querying the Bibster network one receives a large number of results with an often high number of duplicates. This is due to the fact that we do not have a centralized but many distributed local repositories. Furthermore, the representation of the metadata is very heterogeneous and possibly even contradicting. To enable an efficient and easily usable system Bibster presents query results grouping duplicates together. Duplicates in Bibster are bibliographic entries which refer to the same publication, person, or organization in the real world, but are modelled as different resources. Bibster uses specific similarity functions to recognize two resources as being duplicates.

Similarity Function. A similarity function for RDF resources R of the local node repository is a function $sim : R \times R \rightarrow [0..1]$.

For each resource type (publication, person, organization), we have compiled a set of specific features used to assess the similarity between two of its instances. For instance, publications are assessed based on their titles, publication types, authors, years, ACM topics, etc. For each of the features we use different *individual similarity functions*, which are grouped as follows:

The *data value level* focuses on comparisons of data values, which in RDF are represented as typed literals. For example, to determine the similarity of data values of type string (e.g. to compare the last names of persons) we use the *syntactic similarity* of [Mae01]. At the *graph structure level* we check how resources are related to each other. For example, a publication resource is linked to person resources, e.g. authors. Thus we can compare two publications on the basis of the similarity of the sets of authors. This feature alone is not deciding, but it supports the hypothesis of having duplicate entries.

Similarity measures at the *ontology level* extend the ones at the graph structure level by ontology specific characteristics. To determine the similarity of two publications based on their topics we make use of the ACM topic hierarchy. We apply the hierarchical similarity function as presented earlier in this paper [RMBB89].

Applying background *knowledge about a specific domain*, we can define more appropriate similarity functions. For example, in the SWRC domain ontology there are many subconcepts of publications: articles, books, and technical reports to just name a few. Unknown publication types are often provided as Misc. We can thus define a function that returns a value of 1 if the publication type is identical, a value of e.g. 0.75 if Misc is one of it, and 0 otherwise.

From the variety of individual similarity functions, an overall value is obtained with an aggregated similarity function, using a weighted average over the individual functions. For Bibster, the weights have been assigned based on experiments with sample data. More precisely, several duplicates were detected manually. From these training duplicates the weights were adjusted to achieve a maximal f-measure (combination of precision and recall) value.

The Duplicate Relation. As duplicates we consider those pairs of resources whose similarity is larger than a certain threshold

$$t \in [0..1] : D_t := \{(x, y) | sim(x, y) \geq t\}$$

If we assume that the duplicate relation is transitive, we can define the transitive closure as:

$$TC(D_t) := \{(x, z) | (x, y) \in D_t \wedge (y, z) \in D_t\}$$

This transitive closure essentially represents clusters of semantically similar resources.

Resource Merging. Instead of presenting all individual resources of the query result, duplicates are visualized as one, merged, resource. The merged resources comprise the union of statements of the individuals identified as duplicates. In the case of conflicting property values, we apply heuristics for the merging of resources (e.g. for booktitles to select the most detailed value with the least abbreviations).

A.4.9 Results

The Bibster system that implements the methods presented in this paper has been evaluated by means of a public field experiment. The user actions and system events are continuously logged and analyzed to evaluate the user behavior and system performance. We have analyzed the results for a period of two months (June and July 2004) and have obtained the following interesting results: A total of 146 peers from various organizations spread mainly over Europe and North America used the Bibster system. The users shared more than 70000 bibliographic entries. While seventeen peers shared more than 1000 items each, accounting for 84% of the total content, a lot of peers provided only little content or were “free-riding”.

The users performed a total of 1782 queries. The SWRC ontology was used for about half of all queries, mainly for the purpose to search for special types of publications (e.g. only for articles) or for publications of a given author. In 348 queries the users asked for topics of the ACM topic hierarchy. Thereby it is obvious that the users are accepting the ontology based searching capabilities and that there is a benefit for them in using these ontologies.

With respect to query routing, with the expertise based peer selection we were able to reduce the number of query messages by about 50 percent, while retaining the same recall of documents compared with a naive broadcasting approach. Figure A.6 shows the number of forwarded query messages sent per query and the precision of the peer selection (the percentage of the reached peers that actually provided answers to a given query). Although we have shown an improvement in the performance, the results also show that with a network of the size as in the field experiment, a naive approach is also acceptable. On the other hand, with a growing number of peers, query routing and peer selection becomes critical: In simulation experiments with larger peer networks with thousands of

peers, we have shown improvements in the order of one magnitude in terms of recall of documents and relevant peers [HSvH04].

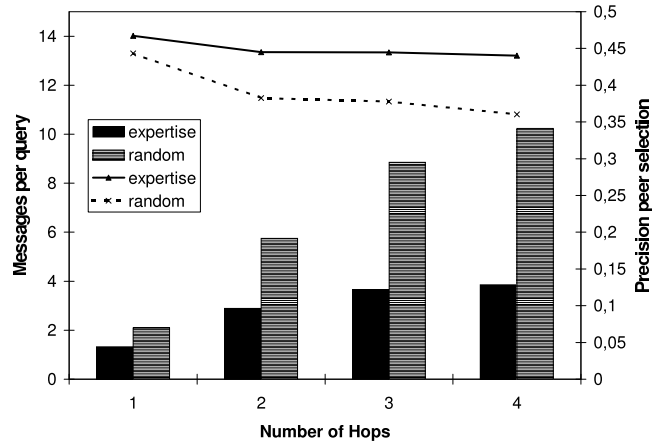


Figure A.6: Number of Messages and Precision of Peer Selection

Lessons Learned

This section summarizes some experiences we have gained from the development and application of Bibster.

For Bibster and similar applications the usage of Semantic Web technologies and ontologies provide an *added value* — in fact it is almost a strict requirement given its semi-structured, volatile data structures. Semantic structures serve important user concerns like high quality duplicate detection or comprehensive searching capabilities.

Unsurprisingly, in small networks with *small user groups*, intelligent query routing is not a major issue. While it is beneficial to direct queries to specific peers known to the user (a trust issue!), advanced routing algorithms may only be beneficial for a much larger number of users in a network. Based on our experience we now conjecture that content-based routing and trust issues will have to converge for such larger networks, too.

A.4.10 Related Work

In the previous sections, related work on the individual aspects of semantics-based Peer-to-Peer technology has already been discussed. Therefore in this section our study of related work focuses on complete systems.

Edutella (*cf. eg.* [N⁺03]) is a Peer-to-Peer system based on the JXTA platform, which offers similar base functionality as the SWAP system. The Edutella network focuses on the exchange of learning material. They use super-peer based topologies, in which peers are organized in hypercubes to route queries. In contrast to their work, Bibster is embedded in the general SWAP architecture and a running application.

[ANS02] describes the design of a Peer-to-Peer network for open archives, where data providers, i.e. research institutes, form a Peer-to-Peer network which supports distributed search over all the connected metadata repositories. This scenario which is similar to our bibliographic Peer-to-Peer scenario, however, their system has not been implemented up to this point.

P-Grid [AMD⁺03] is a structured, yet fully-decentralized Peer-to-Peer system based on a virtual distributed search tree. It aims at providing load-balancing and fault-tolerance, assuming that peers fail frequently and are online with low probability. P-Grid also considers updates with an update algorithm based rumor spreading.

The DFN Science-to-Science (S2S) [Wer03] system enhances content based searching by using peer-to-peer technology to make locally generated indexes accessible in an ad hoc manner. Whereas Bibster is fully distributed, S2S uses a kind of super peers (Search Hubs) to route queries and cache information.

Various systems address the issue of heterogeneity in Peer-to-Peer systems on the schema level, such as the Piazza peer data management system [TIM⁺03], which allows for information sharing with different schemas relying on local mappings between schemas.

A.4.11 Conclusion

In this paper, we have described the design and implementation of Bibster, a semantics-based Peer-to-Peer system for the exchange of bibliographic metadata between researchers. For this purpose, Bibster exploits lightweight ontologies, expressed in RDF Schema in all its crucial aspects: data-organisation, query formulation, query routing and duplicate detection. To our knowledge, Bibster now constitutes the first ontology-based Peer-to-Peer systems ready for fielded deployment.

In general, there are interesting alternatives for each of the different aspects (e.g., [N⁺02] for querying or [AMD⁺03] for query routing) and, actually, we are still exploring multiple approaches to optimize the overall system (e.g., [TSW04] for query routing). In practice, however, it constitutes a major challenge to integrate these different components into a coherent system like Bibster.

The next steps in the development of Bibster are, *(i)* its optimization (e.g., manual query optimization), *(ii)* its spreading to further user groups and, *(iii)* the extension of Bibster to better account for personalized semantic structures, based on the two common core ontologies, e.g. peer-local extensions of the ACM topic hierarchy.

The reader may find further reading material on Bibster, its underlying technologies and related material in the open available project deliverable documentation at <http://swap.semanticweb.org/> and <http://bibster.semanticweb.org/>.

Bibliography

- [AMD⁺03] K. Aberer, P. Cudr Mauroux, A. Datta, Z. Despotovic, M. Hauswirth, M. Puceva, and R. Schmidt. P-Grid: a self-organizing structured p2p system. *ACM SIGMOD Record*, 32(3):29–33, 2003.
- [ANS02] Benjamin Ahlborn, Wolfgang Nejdl, and Wolf Siberski. OAI-P2P: A peer-to-peer network for open archives. In *Workshop on Distributed Computing Architectures for Digital Libraries - ICPP2002*, 2002.
- [B⁺03] J. Broekstra et al. A metadata model for semantics-based peer-to-peer systems. In K. Aberer et al., editor, *Semantics in Peer-to-Peer and Grid Computing - SemPGRID, collocated with the 2003 WWW Conference*, Budapest, 2003.
- [BBD⁺04] M. Bonifacio, P. Bouquet, A. Danieli, A. Donà, G. Marni, and M. Nori. Keex: A peer-to-peer solution for distributed knowledge management. In K. Tochtermann and H. Maurer, editors, *Proceedings of the 4th International Conference on Knowledge Management (I-KNOW'04)*, pages 43–52, Graz, Austria, JUN 2004. *Journal of Universal Computer Science (J.UCS)*.
- [BBT02] Matteo Bonifacio, Paolo Bouquet, and Paolo Traverso. Enabling distributed knowledge management: Managerial and technological implications. *Novatica and Informatik/Informatique*, III(1), 2002.
- [BK04] J. Broekstra and A. Kampman. Serql: An rdf query and transformation language, 2004. Submitted to the International Semantic Web Conference, ISWC 2004.
- [BKv02] J. Broekstra, A. Kampman, and F. van Harmelen. Sesame: A generic architecture for storing and querying RDF and RDFSchemas. 2002.
- [BKvH01] J. Broekstra, A. Kampman, and F. van Harmelen. Sesame: An architecture for storing and querying rdf data and schema information, 2001.
- [Bro03] J. Broekstra. SeRQL: Sesame RDF query language. In M. Ehrig et al., editors, *SWAP Deliverable 3.2 Method Design*, pages 55–68. 2003.

<http://swap.semanticweb.org/public/Publications/swap-d3.2.pdf>.

- [CC94] T. Cox and M. Cox. *Multidimensional Scaling*. Chapman and Hall, 1994.
- [CFM⁺03] A. Castano, S. Ferrara, S. Montanelli, E. Pagani, and G.P. Rossi. Ontology-addressable contents in p2p networks. In *Proceedings of the WWW'03 Workshop on Semantics in Peer-to-Peer and Grid Computing*, 2003.
- [D. 98] D. OLeary. Using AI in knowledge management: Knowledge bases and ontologies. *IEEE Intelligent Systems*, 13(3):34–39, May/June 1998.
- [DDH03] A. Doan, P. Domingos, and A. Halevy. Learning to match the schemas of data sources: A multistrategy approach. *VLDB Journal*, 50:279–301, 2003.
- [EHvH⁺03] M. Ehrig, P. Haase, F. van Harmelen, R. Siebes, S. Staab, H. Stuckenschmidt, R. Studer, and C. Tempich. The SWAP data and metadata model for semantics-based peer-to-peer systems. In *Proceedings of MATES-2003. First German Conference on Multiagent Technologies*, LNAI, Erfurt, Germany, September 22-25 2003. Springer.
- [ES04] Marc Ehrig and Steffen Staab. QOM - quick ontology mapping. In Frank van Harmelen, Sheila McIlraith, and Dimitris Plexousakis, editors, *Proceedings of the Third International Semantic Web Conference (ISWC2004)*, LNCS, pages 683–696, Hiroshima, Japan, 2004. Springer.
- [FSvH03] Christiaan Fluit, Marta Sabou, and Frank van Harmelen. Supporting user tasks through visualisation of light-weight ontologies. In S. Staab and R. Studer, editors, *Handbook on Ontologies in Information Systems*. Springer-Verlag, 2003.
- [Gon01] Li Gong. Project jxta: A technology overview. Technical report, Sun Microsystems Inc., 2001.
- [GPFLC03] A. Gómez-Pérez, M. Fernández-López, and O. Corcho. *Ontological Engineering*. Advanced Information and Knowledge Processing. Springer, 2003.
- [HBE⁺04] P. Haase, J. Broekstra, M. Ehrig, M. Menken, P. Mika, M. Plechawski, P. Pyszlak, B. Schnizler, R. Siebes, S. Staab, and C. Tempich. Bibster - a semantics-based bibliographic peer-to-peer system. In *Proc. of the 3rd Int. Semantic Web Conf. (ISWC 2004)*, Nov 2004.
- [HSvH04] Peter Haase, Ronny Siebes, and Frank van Harmelen. Peer selection in peer-to-peer networks with semantic topologies. In *International Conference on Semantics of a Networked World: Semantics for Grid Databases, June 2004, Paris*, 2004.

- [J. 03] J. Schmücker and Wolfgang Müller. Praxiserfahrungen bei der Einführung dezentraler Wissensmanagement Lösungen. *Wirtschaftsinformatik*, 45(3):307–311, 2003.
- [KC03] G. Klyne and J. J. Carroll. Resource Description Framework (RDF): Concepts and abstract syntax. <http://www.w3.org/TR/rdf-concepts/>, December 2003.
- [LBM03] Yuhua Li, Zuhair A. Bandar, and David McLean. An approach for measuring semantic similarity between words using multiple information sources. *Transactions on Knowledge and Data Engineering*, 15(4):871–882, July/August 2003.
- [Lev66] I. V. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 1966.
- [Mae01] Alexander Maedche. Comparing ontologies - similarity measures and a comparison study. Technical report, Forschungszentrum Informatik, Karlsruhe, Germany, 2001.
- [N⁺02] W. Nejdl et al. EDUTELLA: A P2P networking infrastructure based on RDF. In *Proc. of the 2002 WWW Conference*, pages 604–615, Hawaii, USA, May 2002.
- [N⁺03] W. Nejdl et al. Super-peer-based routing and clustering strategies for rdf-based peer-to-peer networks. In *Proc. of the 12th World Wide Web Conference*, Budapest, Hungary, 20–24 May 2003. ACM.
- [NM03] Natalya F. Noy and Mark A. Musen. The PROMPT suite: interactive tools for ontology merging and mapping. *International Journal of Human-Computer Studies*, 59(6):983–1024, 2003.
- [PSST04] S. Pinto, S. Staab, Y. Sure, and C. Tempich. OntoEdit empowering SWAP: a case study in supporting DIstributed, Loosely-controlled and evolvInG Engineering of oNTologies (DILIGENT). In C. Bussler, J. Davies, D. Fensel, and R. Studer, editors, *First European Semantic Web Symposium, ESWS 2004*, volume 3053 of *LNCS*, pages 16–30, Heraklion, Crete, Greece, May 2004. Springer.
- [RFH⁺01] Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, and Scott Shenker. A scalable content-addressable network. In *Proc. of ACM SIGCOMM '01*, 2001.
- [RMBB89] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. In *IEEE Transactions on Systems, Man and Cybernetics*, pages 17–30, 1989.

- [SF03] D. Schoder and K. Fischbach. Peer-to-Peer Netzwerke für das Ressourcenmanagement. *Wirtschaftsinformatik*, 45(3):313–323, 2003.
- [SMK⁺01] Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek, and Hari Balakrishnan. Chord: A scalable peer-to-peer lookup service for Internet applications. In *Proceedings of the ACM SIGCOMM '01*, 2001.
- [SMM⁺] Ronny Siebes, Peter Mika, Maarten Menken, Peter Haase, and Björn Schnizler. Swap project deliverable d10.4 evaluation report. <http://km.aifb.uni-karlsruhe.de/projects/swap/public/Publications/swap-d10.4.pdf>.
- [SMMH] Ronny Siebes, Peter Mika, Martin Menken, and Peter Haase. Swap project deliverable d10.2 evaluation plan. <http://km.aifb.uni-karlsruhe.de/projects/swap/public/Publications/swap-d10.2.pdf>.
- [TIM⁺03] Igor Tatarinov, Zachary Ives, Jayant Madhavan, Alon Halevy, Dan Suciu, Nilesh Dalvi, Xin Dong, Yana Kadiyska, Gerome Miklau, and Peter Mork. The piazza peer data management project. *SIGMOD Record*, 32(3), 2003.
- [TSW04] C. Tempich, S. Staab, and A. Wranik. REMINDIN': Semantic query routing in peer-to-peer networks based on social metaphors. In *Proc. of the 13th Int. World Wide Web Conference, WWW 2004*, 2004.
- [TXM02] Chunqiang Tang, Zhichen Xu, and Mallik Mahalingam. pSearch: Information retrieval in structured overlays. In *ACM HotNets-I*, October 2002.
- [Wer03] Roland Wertlen. Dfn science-to-science: Peer-to-peer scientific research. In *Proceedings of the Terena Networking Conference (TNC 2003)*, Zagreb, Croatia, 2003.