In: Proc. 2nd Int'l. Conf. on the Practical Application of Intelligent Agents and Multi-Agent Technology (PAAM '97), April 21-23, 1997, London, U. K.

Agent-Based Document Retrieval for the European Physicists: A Project Overview

Uwe M. Borghoff [1], Eberhard R. Hilf [2], Remo Pareschi[1], Thomas Severiens[2], Heinrich Stamerjohanns[2] and Jutta Willamowski[1]

[1] Rank Xerox Research Centre, Grenoble Laboratory
6, chemin de Maupertuis. F-38240 Meylan, France
{borghoff, pareschi, willamow}@grenoble.rxrc.xerox.com

[2] Department of Physics, Carl von Ossietzky University Oldenburg
D-26111 Oldenburg, Germany
{hilf, severien, stamer}@merlin.physik.uni-oldenburg.de

Abstract

Today in physics, more and more scientific documents are prepared electronically, are posted and archived. At the same time, an increasing number of heterogeneous Web-servers are dedicated to hosting such archives. While this trend makes available more relevant scientific information, it also complicates the task of finding, retrieving, processing and printing scientific documents. As a consequence physicists could benefit from an efficient user-friendly tool that integrates functionalities for document retrieval and distribution. In this paper, we present a solution to this problem. Our approach links two systems together: an agent-based search engine for document retrieval and a printing-on-demand service that handles printing, binding and shipping of retrieved documents to readers who request them. We discuss a prototype already available that enables document retrieval from a number of Web servers, including the global network of physics preprints, set up by P. Ginsparg, and the PhysDoc broker which integrates 1000 distributed European Web-servers maintained by local physics departments.

Although our motivation is the physics community, it goes without saying that the ideas and issues of this project apply in general.

1 Introduction

Scientific document handling is a usual practice in physics. In the past, documents were always handled in paper form: authors submitted their manuscripts to publishers and sent printed drafts around to colleagues. The publishers asked referees for comments, and prepared a printed version combining articles into a journal volume. They sold these journals to distributors who in turn sold them to university libraries where the journals could be borrowed for the purpose of reading or copying.

The coming of the electronic age in document handling has changed this practice, especially in physics because of competitive pressure, high mobility of researchers, and large and distributed groups of collaborators, many of whom are also skillful programmers.

1.1 Background

Electronic document handling can be seen as an integrated approach where different, formerly independent tasks are integrated to form a consistent and efficient whole.

However, as Figure 1 points out, for a full integration there are still some missing links. Some tasks, such as document management and information retrieval are already tightly linked together. Other links have still to be established or improved. For example, workflow management systems are not fully integrated with the document management and retrieval process. The same holds true for publishing/viewing and for electronic report distribution/output systems, such as a printing-on-demand system.

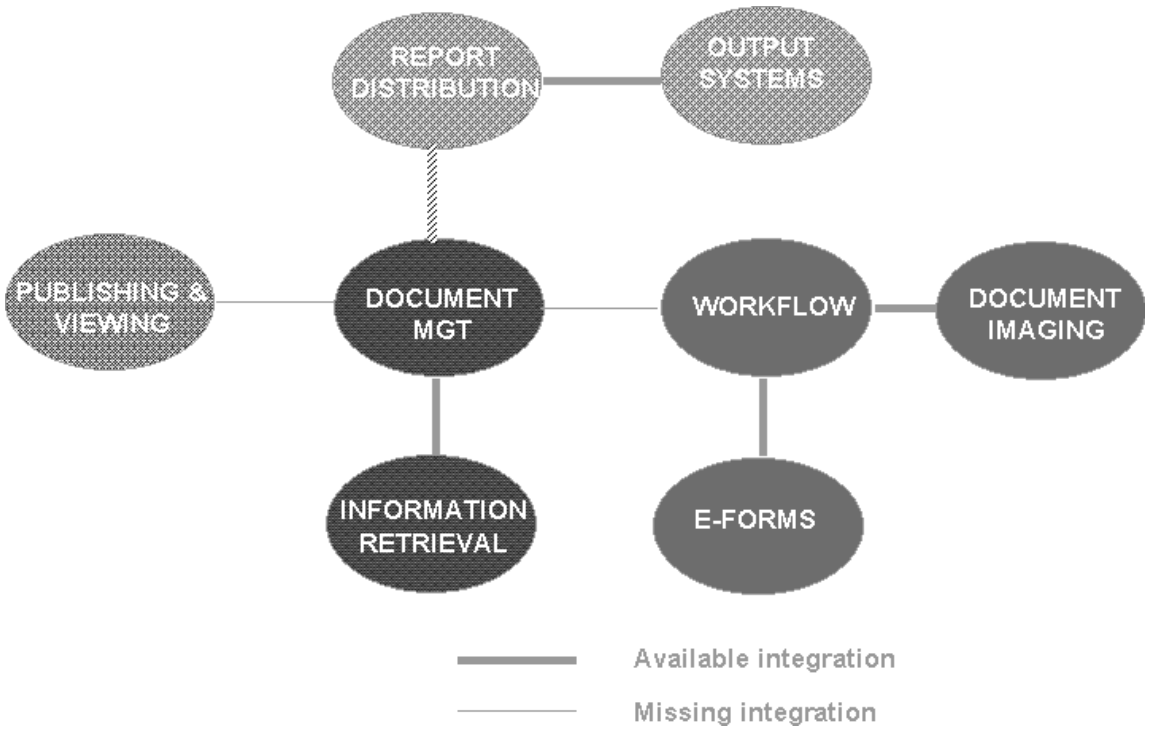

Figure 1: Islands of Integration (courtesy of PERSEO consulting). This paper addresses the integration of Document Management and Report Distribution, or more generally the integration of Document Management / Information Retrieval and Report Distribution / Output.

This paper addresses the improvement and integration of two of these "islands," namely the document management/information retrieval system "island" and the report distribution/output systems "island." An application to document handling in physics illustrates our approach.

1.2 Today's Practice

A new way to integrate electronic document handling is emerging in physics: most documents are prepared electronically, and passed via e-mail or file transfer to one of the Web-servers (i.e. document management systems) where they are made available on the internet to the physics community.

Two groups of servers can be distinguished: a) servers dedicated to high quality documents only; and b) servers created and maintained locally where all kinds of other not refereed ("gray") documents can be posted and found.

Authors post about 40000 scientific publications per year to the globally interconnected distributed preprint hosts belonging to the first group. Such hosts run at the physics laboratories *LANL*[1], home of their creator P. Ginsparg, *SLAC*, Brown University, as well as at the European mirror sites *SISSA* and Augsburg University[2]. They make the publications along with complete bibliographic data available through well-adapted retrieval engines. They offer add-on services such as links to documents cited by the author, and links to documents citing the document itself. They also give the authors a priority assuring entrance date, and provide world-wide access to the documents, as well as readability and archiving.

Other documents which their authors deem not suitable for distribution via preprint servers, such as laboratory annual reports, project status reports, or Ph.D. theses are often still accessible via local department Web-servers. Since the documents are not formally registered, each author individually decides on the format and the content of the information s/he makes available. Apart from the documents themselves, authors often also provide their publication list, references to older published papers, and links to full scanned documents.

Thus, electronic posting of documents is now a generally accepted practice in physics, but integrated automatic *searching*, *retrieval* and *printing/delivery* of the many repositories can be improved greatly. The search engines for preprint servers, publishers' servers and local department Web-servers are not connected to each other. The necessary bibliographic meta-information is, especially in department servers, not sufficiently marked-up. Search engines use proprietary retrieval languages and heterogeneous semantics/logics. There is no agreed standard for presenting the retrieved information. In some cases, just raw data is provided, which is difficult to interpret.

As a consequence, a user searching for documents on a certain topic must

1. know about the relevant servers,
2. connect separately to each one,
3. formulate the query in the appropriate retrieval language, and
4. inspect the results presented in different ways.

Afterwards, s/he needs to select and retrieve interesting documents individually, and, possibly, print them out. This process is further complicated by the fact that the material related to a document is not always stored at the same place as the document: authors submitting to the preprint hosts often provide pointers to additional material, e.g. graphics, on their home Web-server. Finally, documents made electronically have been authored through a vast variety of individually designed text-processing styles formats.

1.3 Project History

To tackle the issues of document *searching* and *retrieval*, the European Physical Society (EPS)[3] has initiated the *PhysDep*-service[4]. This service supports the use of the *Harvest*[5]

---

[1] http://xxx.lanl.gov

[2] http://physinfo.uni-augsburg.de/archiv/index

[3] http://www.nikhef.nl/www/pub/eps/eps.html.

[4] http://www.physik.uni-oldenburg.de/EPS/EurophysNet/PhysDep/query.pl.cgi

[5] http://harvest.cs.colorado.edu/

search engine over a network of about 1000 local web-servers maintained at different European physics departments. The documents to be searched only need to be deposited on the department servers which are regularly searched and indexed by the Harvest system.

Among the specific services provided (Hilf *et al.* 1996a-c) there are *PhysDoc*[6] and *PhysDiss*[7], supporting, respectively, the search of locally stored scientific documents and Ph.D. theses. But, due to the heterogeneity of the information provided by the authors, these services often provide rather cryptic output. Hence, the problem still remains of extracting the relevant data, and making them readable in some familiar user-friendly format. Under discussion is fixing a common set of metadata to be provided by the authors; such a uniform standard should help yield more homogeneous output from the search services.

In order to integrate the EPS services with other services, such as the preprint system cited above, the Constraint-Based Knowledge Broker (CBKB) system[8], currently under development at the Rank Xerox Research Centre in Grenoble, France, was chosen. The CBKB system supports efficient document retrieval, and allows uniform access to the different existing search engines. The first prototype was presented at the 1st PAAM Conference (Borghoff *et al*. 1996), as well as at the joint IuK-Conference, in Munich, Germany (IuK 1996).

The CBKB system acts as a front-end to a printing-on-demand system, originally called *Physicists Network Publishing System (PNPS*), which addresses the problems of document *printing* and *delivery*. It was developed during 1995 in a joint effort between the physicists at the University of Oldenburg, Germany and Rank Xerox Business Services, Germany. The initial design and an early prototype were presented at the DPG-Conference 1995, in Berlin, Germany.

The CBKB system forwards the retrieved documents to the printing-on-demand system. It thus bridges the gap between the information retrieval/document management system "island," represented by the Web-servers with their search interfaces, and the report distribution/output systems "island," represented by the printing-on-demand system.

1.4 Outline

In Section 2, we summarize existing approaches related to resource discovery on the World Wide Web, highlighting the trend from individual search engines and publishing systems towards integrated document retrieval and delivery systems. Section 3 illustrates our proposed retrieval strategy: a uniform meta-search interface with clear semantics, developed on top of different search engines. We handle complex queries where queries are decomposed, and submitted partially, and where results are recombined, and locally sifted using constraint solving techniques. Our initial results with applications to the global preprint server system and to the European distributed document system are presented. Problems due to the extreme heterogeneity will be discussed. Section 4 shortly describes our novel *printing-on-demand* service which offers printing, binding in book form, and mailing of the documents finally selected by the reader, once she/he has found the URL of the documents in question. A prototype has been set up and tested thoroughly for Rank Xerox Business Services (Stamerjohanns *et al.* 1995). Section 5 summarizes the work performed, and gives perspectives for the design

---

[6] http://www.physik.uni-oldenburg.de/EPS/EurophysNet/PhysDoc/query.pl.cgi.

[7] http://www.physik.uni-oldenburg.de/EPS/EurophysNet/PhysDis/query.pl.cgi.

[8] http://www.rxrc.xerox.com/research/ct/prototypes/cbkb

strategies of the development of the broker system, and for the necessary automatization of adaptation work in order to integrate and connect further brokers, or other on-line repositories and databases. Finally, we discuss the necessary "society" of meta-brokers extracting information from existing distributed information sources.

2 Related Work

2.1 Overview

Well-established, electronic publishing systems, as the World-Wide Web (WWW) or the earlier *Gopher*, provide a seamless information space in the internet. Index and search subsystems appeared hand in hand with the rapid growth in the amount of information and in the number of users having specific needs (Obraczka *et al.* 1993, Schwartz *et al.* 1992).

The *Wide-Area Information Servers* (WAIS) (Kahle and Medlar 1991), for instance, provide a Z39.50-based search and retrieval interface. Another famous approach is *Archie* (Emtage and Deutsch 1992) which periodically contacts a set of registered ftp-servers to index the file names.

At the University of Karlsruhe, A.-C. Achilles supports a well-known application on top of *Glimpse* (Manber and Wu 1994), namely the sophisticated search facility for a large collection of computer science bibliographies[9].

Based on Glimpse, Hardy *et al.* (1994-1996) developed *Harvest*, in which they separated a gatherer and a broker. Using special indexing-techniques, Harvest reduces network load.

Although unspecific undedicated multi-source index/search subsystems have already been built for Gopher, with *Veronica*[10], and for the WWW, e.g. with *Alta Vista*[11], *Lycos, Yahoo*, retrieval engines or retrieval support systems for heterogeneous information are an active research field (Barbara 1993).

The current search tools can be roughly divided into four groups. The first group of unspecific undedicated brokers tries to surpass each other in their coverage of html-pages (e.g., *Alta Vista*). They index "everything" their spiders find. They have a very efficient, but rather simple keyword-based interface allowing keyword grouping with and and or. However, the user is then left to extract the desired specific answer out of a possibly huge number of hits, say the few scientific publications out of the thousands of retrieved user group annotations. Even if the advanced query interface of Alta Vista for example provides means to rank the retrieved documents, this is generally not enough to relieve the user from filtering out the interesting documents by himself.

The second group, the dedicated ("homogenous") brokers, search on a defined coherent set of hopefully well marked up documents with a correspondingly well adapted search engine. Examples include brokers for books in a library, or the abstract data bases of commercial hosts. The user starts by choosing a data repository (e.g. Library of Congress or some university's digital library), gets a corresponding form showing what she/he can ask for there, and, finally, performs the search at that repository gaining a well marked up set of answers.

---

[9] http://liinwww.ira.uka.de/bibliography/index.html

[10] gopher://gopher.unr.edu/11/veronica

[11] http://altavista.digital.com

The third group of dedicated heterogeneous brokers normally just take a simple keyword query and pushes it to several distributed brokers and search engines, independently and in parallel. The user receives a list of answers with comments where the hits were found and which tool was successful in finding them. Recent examples in physics are a broker by M. Bischoff[12], searching in about 50 distributed libraries in the Darmstadt area in Germany, and the *One Stop In Physics*, searching in the preprint databases of major high-energy research facilities and in the LANL repository. However, these systems fall short of the ideal requirements due to restricted functionality, possibly unspecific and excessive querying of the addressed brokers, and to poor precision. Also, over all they still provide only poor recall.

The final group includes offline Web agents designed for asynchronous browsing-support, such as

- *Folio Web Retriever V. 2.*0 from Folio Corp.
- *FreeLoader V. 2.0* from Freeloader, Inc.
- *Metz Netriever V. 1.1* from Metz Software, Inc.
- *The PointCast Network V. 1.1* from Pointcast, Inc.
- *Smart Bookmarks V. 2.0* from FirstFloor, Inc.
- *WebEx V. 1.01* (formerly MilkTruck) from Traveling Software, Inc.
- *Web Whacker V. 2.0* from ForeFront

What is missing, is distributed joint-functionality among the different data repositories, and a sophisticated way of processing the search results.

To answer this need we have designed and implemented a *heterogeneous* broker, which logically combines heterogeneous information retrieved from other brokers, on-line repositories, and databases (Borghoff and Schlichter 1996), in reply to a physicist's search query.

Intelligent agents (CACM 37:7 1994, Wooldridge and Jennings 1995) or knowledge brokers (Barbara and Clifton 1992) also address this problem. A related thrust in knowledge brokerage has been defined by the consortium of the *KRAFT* project[13].

One basic law governing the future society of brokers negotiating and querying each other automatically is that the querying broker be at least as "intelligent" (as far as query semantics and result parsing are concerned) as the queried ones. Otherwise, the user should be passed directly to the more "intelligent" broker.

This is why we specify that referred specific, and heterogeneous brokers be at least as "intelligent" as the queried ones. The brokers provide added-on value by intelligent combining and evaluating the heterogeneous data. We define as task: creating a search service on the Web where the user can formulate a complex, well-specified request, including interdependencies among the attribute-based entries of different, heterogeneous data sources. This service should offer on-line as well as off-line help for putting together the request, and should exploit the full search semantics of the search engines involved. In addition to that, the tool could be enhanced by local constraint solving techniques to extend the search capabilities in

---

[12] bischoff@iap.physik.th-darmstadt.de

[13] http://www.csd.abdn.ac.uk/~apreece/Research/KRAFT

order to allow local filtering and sifting of information. We will give examples in the later paper. Finally, we envision the pure search service integrated into the networked publishing system, as outlined by Borghoff *et al.* (1996b).

The *Constraint-Based Knowledge Broker* system (CBKB) fullfills these specifications. Constraints have been introduced to flexibly manage the search space of broker agents, as well as to flexibly adapt user requests and answers from information providers. Andreoli *et al.* (1995, 1996) present the theoretical background of the CBKB system. Protocol issues within the CBKB system are addressed by Arcelli *et al.* (1995) and by Borghoff *et al.* (1997).

Fikes *et al.* (1995) also use logic-based models to capture the domain of expertise of information brokers. Instead of using constraints, their modeling language is based on a predicate logic with contexts. The *Tsimmis* project (Chawathe *et al.* 1994) takes a different approach using a self-describing object model for the internal representation of information and queries.

The proposed service for integrated document delivery in physics in Europe combines the strengths of the search capabilities of the *Constraint-Based Knowledge Broker* model with the advantages of the *Physicists Network Publishing System*. At the backend, the service will connect, among others, to the Harvest servers within the European Physical Society, and the global physics preprint network. At the frontend, the user interacts with the service via her/his preferred Java-enabled Web-browser.

2.2 Features of the Constraint-Based Knowledge Brokers

The key features of the *Constraint-Based Knowledge Brokers* are:

1. *Structured Information* — The brokers parse unstructured information and return it in a structured format as if it were a database record. But there is more to it than just disguising unstructured information repositories as (pseudo)-databases. In fact, the parsed information can be passed to constraint solvers that extend backend query engines with additional filtering capabilities. In this way, the filtering mechanism can be tuned to return information which is relevant and accurate for the user (for instance, the information returned by a bibliographic backend query engine could be filtered through such constraints as *year of publication* > *1994*, *publisher != IEEE*)
2. *Concurrent Asynchronous Searches* — A broker search engine can launch many concurrent searches that in principle could go on forever, be checked, refined and re-launched periodically. This is a radically different user paradigm (and a complementary one) than the “one query at a time” kind of interaction with which Web search engines support today.
3. *Incremental Refinements* — The brokers do not distinguish between “queries” and “answers”. Both cases are treated as “sets of (information) constraints”. Under this view, an answer is simply a refinement of the information initially contained in the query, and can be the input to another search. A broker environment can take advantage of this feature in several ways:
    - Queries can be arbitrarily complex, and can reflect the user context.
    - Users can interrupt a running search, inspect and edit the current constraint set, and then resume querying. Users can also manipulate and re-submit a modified set.

- By leveraging the concurrency of the broker search engine, multiple users work together in a query session, and collaboratively refine queries/answers.

4. *Knowledge Combination* — The broker search engine allows combination of information returned by different information sources. Thus, this feature introduces a form of "joining," similar to that provided by database systems in the context of information gathering from multiple sources in distributed domains. From the point of view of the user, knowledge combination can be thought as a way of combining and refining complex information until it becomes "knowledge" that can be used for problem solving and decision making, as well as for consolidating existing knowledge.

There are other features which are relevant to the CBKB framework, for instance the caching mechanism for information re-use, however the four above are deemed as the most important in the current context.

3 Current Implementation

The CBKB system already serves as a testbed for research activities in the area of digital libraries at RXRC Grenoble. Our second project was to respond to the physics community's need for integrated services by connecting PhysDoc, Physdis, and the Augsburg mirror of LANL.

These repositories are distinct and heterogeneous enough to test the ability of CBKB to integrate new backends.

The objectives of the Oldenburg server are different from those of the Augsburg preprint server. The Oldenburg server aims to provide uniform access to as large as possible a set of documents available at European physics departments. It provides only basic information about the heterogeneously marked-up and distributed documents.

The Augsburg server gives access to a large number of formally registered scientific documents with complete bibliographical data. It provides detailed information about each registered document. Thus the possibilities to formulate a query and the format and complexity of the results obtained from these servers differs greatly. Also, until the CBKB integrated these two services, the Web query-forms to these services consisted only of a textfield for query specification (see Figure 2).

With the Oldenburg search services, the original Web search form allowed formulating queries either using a large number of possible search fields, such as document-type, author, or title, or alternatively via fulltext search. However, the possibility to retrieve a document via specific search fields depends on the mark-up/meta-information provided by the author. For example, if the author field is not explicitly marked up, it cannot be indexed by the Harvest engine and can therefore not be used as search field to retrieve the document. Due to the non-uniformity of the mark-up and the provided bibliographic meta-information, the results returned by the Oldenburg server only contain the URL of the matched document, the corresponding host and path, a one-line-description, and the lines of the document containing a match to the query.

In order to homogenize indexed documents and query results information, a common set of meta-information fields for all the connected department servers is currently being defined.

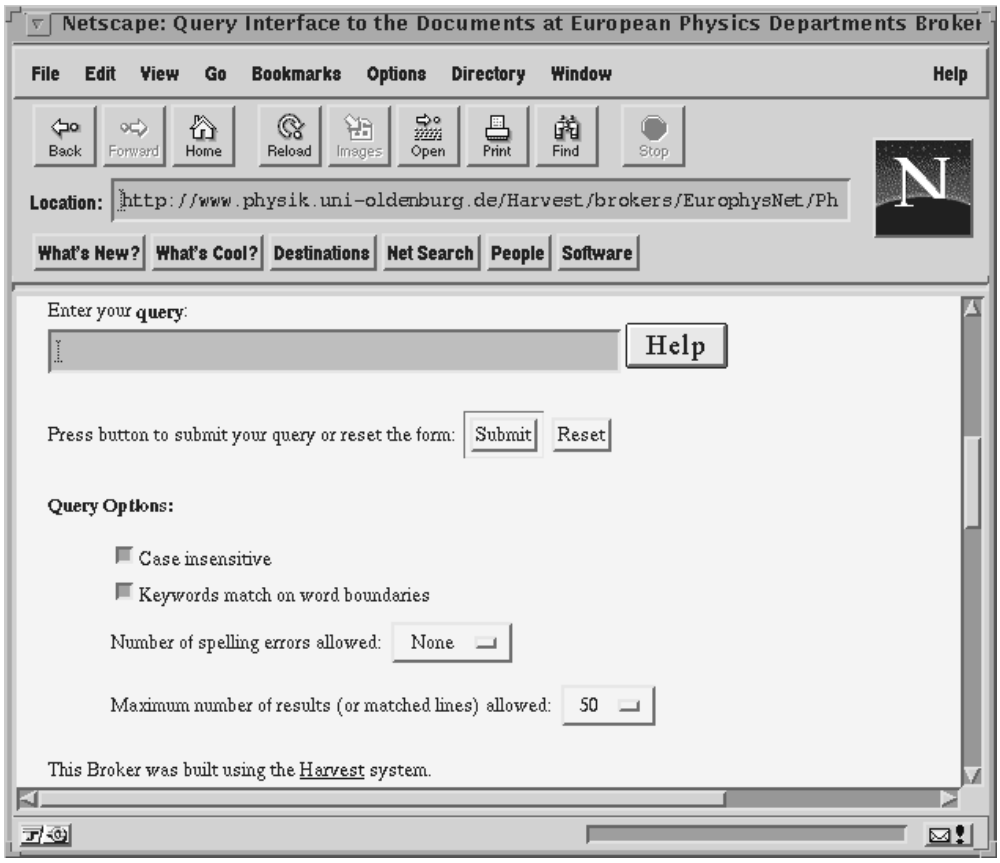

Figure 2: The search form for the PhysDep service only provides a textfield to enter the search term.

With the Augsburg search service, a query always triggers a fulltext search. The search keywords are matched against a WAIS index covering the author, the title and the abstract of the registered documents. Thus through the service's web search form, one cannot, for example, explicitly search for articles containing specific search keywords in the title: all articles containing the keyword in the abstract are then also listed. On the other hand, the filtering capabilities of the CBKB-system can filter hits. In contrary to the Oldenburg server, the results returned from the Augsburg server contain the abstract and complete bibliographical data as well as the URL of the matching documents. These result fields can also be constrained in order to further filter out the most relevant documents.

The goal of the Constraint-Based Knowledge Broker system (CBKB) is to provide a uniform search interface to a number of the existing backends, no matter which search engine or indexing tool they use. The CBKB interface exploits all the capabilities of the underlying backends but provides a uniform formalism for formulating queries and to present the results obtained from the different servers. Several relevant fields, such as author, title etc., are proposed to the user for specifying queries (Borghoff *et al.* 1996). The final output can be displayed in any of a number of formats selectable by the user: readable on screen as bibliographical data (authors, title, status, link, abstract); in summary form (ranking or sorting the results alphabetically by title or author for example); or displaying the complete information found about each matched document.

Figure 3 shows CBKB in action: the system is managing several queries concurrently (in the foreground), and displays query results (in the background).

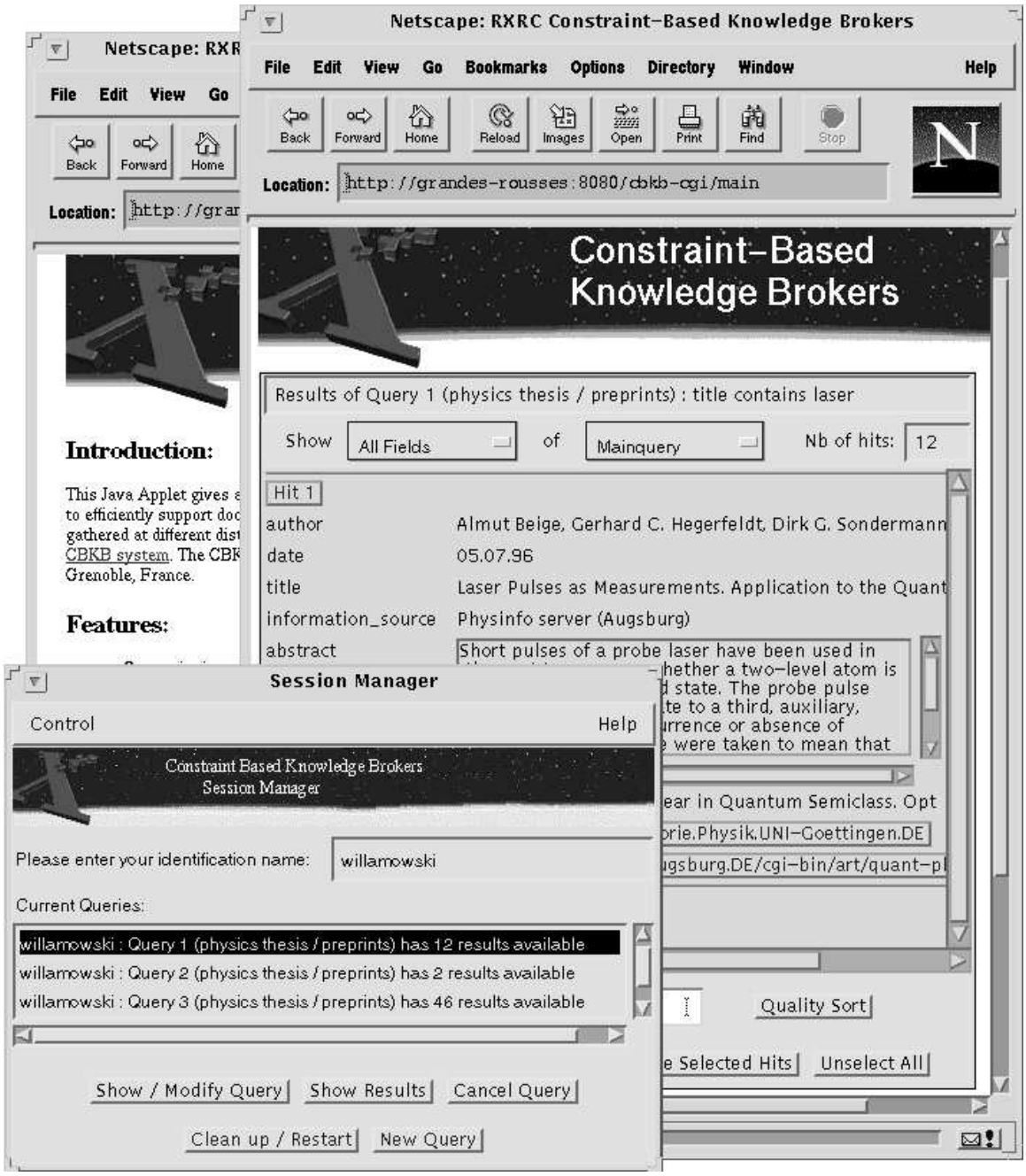

Figure 3: The graphical user interface of the Constraint-Based Knowledge Broker system presenting results to a query about documents containing *laser* in the title.

In the example, Query 1 requests documents about lasers. Twelve hits were found in the global preprint service (Augsburg mirror) and the Oldenburg server altogether. The user can dynamically customize this presentation: choose format, and specificity (e.g. display only the information responding to the main query or display information corresponding to possible subqueries of a complex query). The user can then either directly inspect the retrieved documents (by clicking on the corresponding URL) or decide to select relevant hits for saving or sending to the printing-on-demand service (see Section 4).

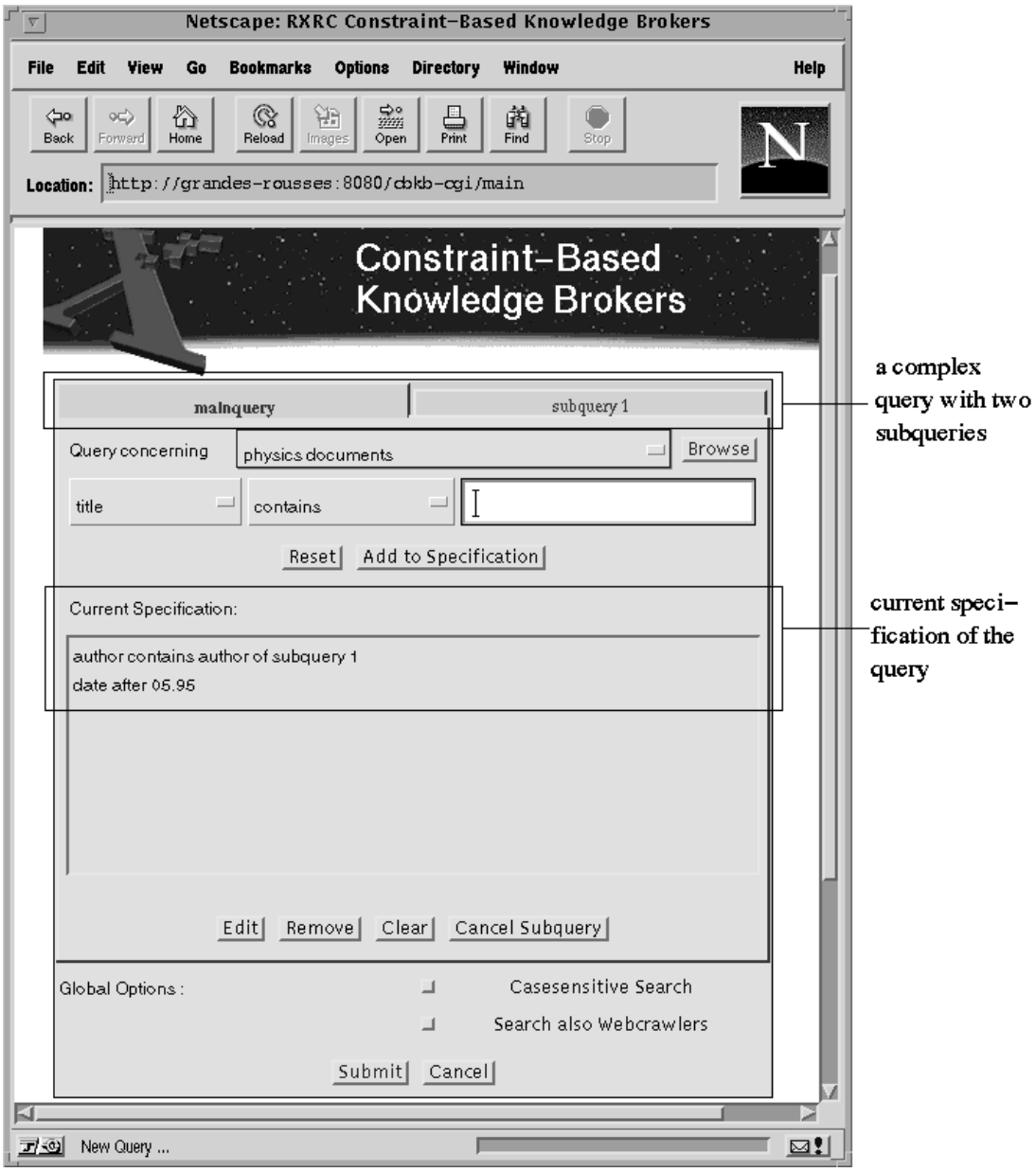

Figure 4: The graphical user interface of the Constraint-Based Knowledge Broker during complex query formulation. In this example the user's query is composed of two subqueries; the *mainquery* is displayed in this screenshot. The user is searching for documents written *after may 1995*, by the same *author* as the documents retrieved with *subquery 1*. Subquery 1 could e.g. be the query for an article for which the user knows the title but not the author.

For the moment, the user can only enter simple flat queries, although the user interface and the brokers are able to manage complex ones, i.e. queries that must be decomposed into interdependent subqueries. For example, the results of a first subquery could be used as search pattern in a second subquery (see Figure 4). More concretely, one could search for articles written by co-authors of a specific author. This capability cannot be exploited at the present time because the results obtained from the different connected backends are still too heterogeneous and some result fields would require further natural language analysis. For instance, if the contents of the result-field author is to be used in a subsequent query, its content should be split into several fields if it contains several names. Furthermore other backends have to

be connected in order to obtain sufficient coverage to formulate and respond to complex queries.

Connecting an external server to the system is done by analyzing its search interface, then writing a wrapper for it. This wrapper receives the description of the constraints corresponding to the query, translates them into the query-string required by the search script, verifies that the indicated fields are accepted by the server and provides default values for required fields not specified by the user. It then queries the server and receives the results in html-format. Finally it parses the results and translates them into the constraint format accepted by the CBKB system. Additional filtering and sifting of results according to user-constrained field entries that are not properly handled due to restrictions in the backend search capabilities is managed by the local constraint solvers of the CBKB system. Problems such as homogenizing the presentation of results from different backends, e.g. presenting information always in the same order, are solved by the CBKB user interface.

The work required to integrate a new server depends solely on the complexity of the query interface the server provides. While writing of the corresponding code can be done relatively quickly, it could be accelerated if the backend servers automatically provided the format specification for the query, enumerating the query fields they accept, and the format in which they send back results, listing the possible result-fields. Currently result parsing is done using the html-tags used to separate the items; thus, whenever the tag is changed on the server side, the interface program must be modified accordingly.

One field always provided as result by the connected backends is the URL of the matched document. This allows to retrieve the document for printing. The URL together with other relevant information out of the result set can then be forwarded to the document delivery subsystem, described in the following.

4 Document Delivery

Printing a document found by a broker can be a bothersome problem for the reader.

The global physics document server solves this by storing the original LaTeX documents, tested as compiling to a postscript file. The attached graphics elements are not guaranteed however, as they are archived at the local server of the author. In addition, these attachments are often very large. It is frustrating for the reader to download these attachments over small bandwidth connections and then to have problems printing them out. Printing-on-demand services can remove these burdens from the user.

Writing a script to funnel a given URL to a local printer is easy. However, for a professional (and thus commercial) printing-on-demand service, the transfer is more complex. A prototype has been written which comprises an initial Web-form for the order, customer identification, and includes the URLs of the documents in question. The prototype phase of the transfer operation passes the printing requests to Oldenburg, then downloads the document from the archive where it resides, checks for printability, and processes the document. In addition, a job ticket transfers the necessary information to the provider of the service. Finally, the document is printed, bound (when requested), and sent by surface mail to the customer. Although this process was time consuming and took several days for delivery, the customer got - with seemingly no effort - a perfectly bound book with professional color print-out on his desk. The prototype system was successfully tested in late 1995 and received a positive response from many countries all over the world. The launch of the respective

commercial service by Rank Xerox is still pending. Different business models are currently being evaluated.

The printing-on-demand service can be seamlessly integrated as one backend option of the broker. An other printout service is offered by some large University Libraries as well as the British National Library. Here the document has to be available at the respective library as a paper document. If a document is ordered by e-mail, fax or letter, the library scans in the respective document picked from the shelf, and delivers it by fax (outside Germany), or, in some cases, by e-mail to the reader. While this is cumbersome for large documents it certainly fills a gap.

The real benefit to a reader, however, is to have a single user-friendly search interface, and to receive the document readable on his desk, with no effort on his/her side. Thus, all the different parts of such a service should be smoothly embedded into the future envisioned Physicists Network Publishing System.

5 Future Work and Summary

The main deficiency in the electronic information handling of scientific physics documents is the complex process of (1) finding documents in a heterogeneous set of individually styled and formatted local servers, (2) retrieving them, and finally (3) obtaining a high quality printout. In this paper we have presented the Constraint Based Knowledge Broker System which allows users to search seamlessly heterogeneous Web-servers and to forward retrieved interesting documents to a printing-on-demand system.

At the present time, each Web-server to be searched must be connected manually to a centralized system. However, a centralized solution for registration and adaptation of the many different data formats used within the search/index subsystems will eventually fail. Therefore, we plan to provide a semi-automized wrapper-builder which allows rapid connection of new servers/repositories to the existing service. Once, a critical mass has been reached and the service is well-established, the open design of the integrated document retrieval and delivery system will enable remote sites to register their repositories in a do-it-yourself style. Changes to the remote data format or modifications in the search interface could then be notified by the data owners themselves. This will greatly improve the consistency and coherence of the service.

While it is straightforward to coherently search through ordinary Web-servers, a *society* of collaborating heterogeneous brokers has yet to emerge in the near future. The CBKB framework addresses the problem of extracting a maximum amount of information from other distributed brokers, repositories and document servers, by logically combining the heterogeneous information these provide. Thus, it reduces the load on the user, and, in some cases, on the net as well. Care has to be taken to transfer the full search semantics an individual broker offers to the CBKB system. If an underlying broker is searched less specifically than necessary, too much information must be downloaded and the underlying broker will be used more intensively than necessary. The overall reputation of this “meta-broker” might suffer. As a first step, a CBKB operator has to negotiate the access modes among the individual brokers.

Acknowledgments

We gratefully acknowledge funding by Rank Xerox Business Services (XBS) in 1995, and continuous constructive discussions with H. Karch of Rank Xerox Germany Systems Opera-

tions (RXG-SO). We also wish to thank Natalie Glance and the anonymous referees for their comments on earlier versions of this paper.

References

Andreoli, J.-M., Borghoff, U. M., Pareschi, R., Schlichter, J. H. (1995). Constraint Agents for the Information Age. *J. Universal Computer Science* **1**:12, 762-789. Electronic version available at http://www.iicm.edu/jucs.

Andreoli, J.-M., Borghoff, U. M., Pareschi, R. (1996). The Constraint-Based Knowledge Broker Model: Semantics, Implementation and Analysis. *J. Symbolic Computation* **21**:4, 635-667.

Arcelli, F., Borghoff, U. M., Formato, F., Pareschi, R. (1995). Tuning Constraint-Based Communication in Distributed Problem Solving. In *Proc. 1st Int. Workshop on Concurrent Constraint Programming (CCP '95),* Venice, Italy.

Barbara, D., Clifton, C. (1992). Information Brokers: Sharing Knowledge in a Heterogeneous Distributed System. *Technical Report MITL-TR-31-92*, Matsushita Information Technology Lab., Princeton, NJ.

Barbara, D. (1993). Extending the Scope of Database Systems. *Technical Report MITL-TR-44-93*, Matsushita Information Technology Lab., Princeton, NJ.

Borghoff, U. M., Pareschi, R., Arcelli, F., Formato, F. (1997). Constraint-Based Protocols for Distributed Problem Solving. *Science of Computer Programming*, to appear in Oct. 1997.

Borghoff, U. M., Chevalier, P. Y., Willamowski, J. (1996). Adaptive Refinement of Search Patterns for Distributed Information Gathering. In *Proc. Int. Conf. EuroMedia/WEBTEC*, London, pp. 5-12. Electronic version available at http://www.rxrc.xerox.com/research/ct/publications/home.html.

Borghoff, U. M., Pareschi, R., Karch, H., Nöhmeier, M., Schlichter, J. H. (1996). Constraint-Based Information Gathering for a Network Publication System. In *PAAM '96*, London, pp. 45-59.

Borghoff, U. M., Schlichter, J. H. (1996). On Combining the Knowledge of Heterogeneous Information Repositories. *J. Universal Computer Science* **2**:7, 512-532. Electronic version available at http://www.iicm.edu/jucs.

CACM **37**:7 (1994). Special Issue on Intelligent Agents. *Communications of the ACM*.

Chawathe, S., Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J., Widom, J. (1994). The Tsimmis Project: Integration of Heterogeneous Information Sources. In *Proc. IPSJ Conf.*, Tokyo, Japan. Los Alamitos, CA: IEEE Comp. Soc. Press.

Emtage, A., Deutsch, P. (1992). Archie: An Electronic Directory Service for the Internet. In *Proc. Usenix Conf. Winter '92*, pp. 93-110, Sunset Beach, CA. Berkeley, CA: Usenix Association.

Fikes, R., Engelmore, R., Farquhar, A., Pratt, W. (1995). Network-Based Information Brokers. In *Proc. AAAI Spring Symp. Series on Information Gathering from Distributed Heterogeneous Environments*, Stanford, CA.

Hardy, D. R., Schwartz, M. F., Wessels, D. (1994-1996). *Harvest User's Manual*.

Hilf, E. R., Boswell, P. G., Laloe, F. (1996). Many Opportunities for Collaboration. *Europhysics News* **27**, 77. Electronic version available at http://epswww.epfl.ch/ene/ene_apr_paris_text.html.

Hilf, E. R., Rohen, G., Severiens, T. (1996). Electronic Information Management in Physics. In: Gasteiger, J.; GDCh (ed.): *Software-Entwicklung in der Chemie* **10**, 89-96. Electronic version available at http://schiele.organik.uni-erlangen.de/tagung/10_cic/hilf/index.html.

Hilf, E. R., Diekmann, B., Stamerjohanns, H., Curdes, J. (1996). Integrated Information Mangement for Physics. In: Dubois, J.-E., Gershon, N. (eds.*): The Information Revolution: Impact on Science and Technology*, pp. 186-189, Berlin, Heidelberg: Springer-Verlag. Also In: *Data and Knowledge in a Changing World,* September, 1994, Chambery, France. Electronic version available at http://www.physik.uni-oldenburg.de/documents/UOL-THEO3-95-2/codata7/codata7.html.

IuK-Conference "Neue Medien für die Wissenschaft" of the GI, GDCh, DPG, DMV in April 1996, Munich, Germany. On-line program: http://www11.informatik.tu-muenchen.de/proj/Medoc1/IuK96/IuKProgramm.html.

Kahle, B., Medlar, A. (1991). An Information System for Corporate Users: Wide Area Information Servers. *Connexions: The Interoperability Report* **5**:11, 2-9.

Manber, U., Wu, S. (1994). Glimpse: A Tool to Search Trough Entire File Systems. In *Proc. Usenix Conf. Winter '94*, pp. 23-32, San Francisco, CA. Berkeley, CA: Usenix Association.

Obraczka, K., Danzig, P. B., Li, S.-H. (1993). Internet Resource Discovery Services. *IEEE Computer* **26**:9, 8-22.

Schwartz, M. F., Emtage, A., Kahle, B., Neuman, B. C. (1992). A Comparison of Internet Resource Discovery Approaches. *Computing Systems* **5**:4, 461-493.

Stamerjohanns, H., Diekmann, B., Hilf, E. R. (1995). Drucken aus dem Internet: ein neuer Dienst. *Internal report for Rank Xerox Business Services*. http://alice.physik.uni-oldenburg.de/rxp/.

Wooldridge, M., Jennings, N. R. (1995). Intelligent Agents: Theory and Practice. *Knowledge Engineering Review* **10**:2, 115-152.