Developing a European Technical Reference Digital Library

Antonella Andreoni, Maria Bruna Baldacci, Stefania Biagioni, Carlo Carlesi, Donatella Castelli, Pasquale Pagano, and Carol Peters

Istituto di Elaborazione della Informazione-CNR,Pisa, Italy
{andreoni,baldacci,biagioni,carlesi,castelli,pagano,carol@iei.pi.cnr.it}

Abstract. The development of a European digital library for grey literature is described. The aim has been to provide a digital library for scientists working in the areas of information science and applied mathematics and also to build a test-bed for research activities. The service has been implemented as part of NCSTRL (the US Networked Computer Science Technical Reference Library) and developed, extending the Dienst system used by NCSTRL, to meet the requirements of the European scientific community. The additional functionality is described and the difficulties encountered when trying to extend an existing architecture, protocol and system are discussed.

1 Introduction

The aim of the Digital Library Initiative of the European Research Consortium for Informatics and Mathematics (ERCIM) is to promote the development of digital library technology in Europe. Since 1996, a series of research-oriented activities, mainly sponsored by the DELOS working group, have thus been organised, e.g. workshops, conferences, collaborative studies on DL-related research issues. However, towards the end of 1997, ERCIM also decided to undertake an implementation activity by setting up its own digital library for documentation produced by its member institutes: the ERCIM Technical Reference Digital Library (ETRDL). The intention was two-fold:

- to assist ERCIM scientists in making their research results immediately available world-wide and provide them with appropriate facilities for accessing the technical documentation of others working in the computer science or related areas;
- to provide the ERCIM DL group with a test-bed for experimental activities such as the implementation of new functions or services.

ETRDL has thus received funding from ERCIM towards the setting up of the DL service and from DELOS for DL research activities.

The first prototype of ETRDL was released in 1998 [1],[2]; after a one-year period of testing and refining, the ETRDL service is now available for the ERCIM Librarians and scientists and for the general public at http://www.iei.pi.cnr.it/

S. Abiteboul, A.-M. Vercoustre (Eds.): ECDL '99, LNCS 1696, pp. 343–362, 1999.

DELOS/ETRDL. We are now working on developing and testing additional functionality that will be implemented in ETRDL-2.

In this paper, we describe the work involved in implementing this first version of ETRDL[1], discuss the reasons why we decided to base our digital library on an existing service (the US Networked Computer Science Technical Reference Library - NCSTRL) and adopt the same protocol (Dienst), explain why the services offered by NCSTRL did not entirely satisfy the requirements of a digital library for European research institutions, and illustrate the problems that arose when trying to implement additional functionality in an existing system. The paper is thus organised as follows: Section 2 outlines the services that should be provided by a European digital library for scientists and compares them with those offered by NCSTRL; Section 3 describes the basic functionality provided by Dienst and Section 4 its application and extension in ETRDL; Section 5 discusses the language needs of a multilingual community and the provisions we are making; the final section indicates our plans for future developments.

2 What services should a DL provide for the ERCIM scientific community?

The first step towards the development of ETRDL was a survey of the requirements of the ERCIM member institutions in order to determine how best to satisfy them. Our starting point was the library and an analysis of how the technical documentation produced by our institutes is traditionally processed and managed. This is because ETRDL is seen as an extension of the traditional institutional library services through the creation of an infrastructure connecting the separate technical collections of the single institutions and providing them with a gateway to related scientific collections.

A digital library has been defined as "an institution that performs and/or supports (at least) the functions of a library in the context of distributed, networked collections of information objects in digital form" [3]. This is the perspective taken by ETRDL. We intend that our digital library should provide far more than a dynamic remote search functionality; it should provide users with a complete service. While it is true that the presence of electronic rather than paper documents completely revolutionises information search and document access possibilities - by eliminating boundaries of space, time, and location - it is also true that, in order to exploit such capabilities, a number of tasks must be performed. Some of these are very similar to those of the traditional library even if their execution is different: the documents must be acquired, described bibliographically, catalogued, and retrieved. Others are new: for the librarian, they include issues such as the organisation and preservation of digital collections, security, copyright, control of versions, updating; for the users, the ways

[1] The ERCIM Technical Reference Digital Library (ETRDL) is a collaborative activity in which eight organisations (CNR, CWI, GMD, INRIA, SICS, INESC, SZTAKI, FORTH) currently participate. It is sponsored jointly by ERCIM, by the DELOS Working Group (ESPRIT LTR No. 21057), and by the participating institutions.

in which documents can be accessed and manipulated. Our intention is that ETRDL should provide a full set of functionality for the performance of such tasks.

The ETRDL system has thus been developed in terms of the services to be provided for two main user classes: the librarian, and the scientist. The librarian is responsible for the management of the information; the scientist is seen as both a potential seeker and provider of information. ETRDL aims at providing both the librarian and the scientist with an integrated work environment from which any of the DL services on the ERCIM collections can be accessed, in the preferred language of the user[2].

This environment must take into account three important features of the collections of technical documentation produced by the ERCIM member institutions: i) for historical and cultural reasons, the boundaries between informatics and mathematics are not well defined in these collections; ii) the documents are written in many different languages; iii) the collections contain different types of documents - currently ranging from technical reports to pre-prints of journal articles - with different characteristics and different potential life-times.

These features have very much influenced the design of the services to be offered by ETRDL because they imply functionality that:

- provides users with local language interfaces
- searches across languages
- searches by subject
- selects sub-collections by date, language, type
- handles withdrawal and update as well as submission capabilities.

Another important requirement was also identified: we did not want to create a DL in *isolation* but to develop a tool that would encourage the dissemination of ideas and communication between researchers around the world working in computer science or applied mathematics. The obvious solution was to follow the direction of an initiative already underway in the US: NCSTRL - the Networked Computer Science Technical Reference Library [4], with a very similar core motivation: to improve early and detailed communication of research results across the community [5]. This initiative has extensive visibility, with a large number of European and North American participating institutions. This decision implied employing the Dienst system [6],[7], adopted by NCSTRL for disseminating, searching and accessing its documents; Dienst is an open system, independent of NCSTRL and can be extended to meet the needs of other applications. We thus decided to build our collection along the same lines as those established by NCSTRL and to adopt the same basic infrastructure.

However, there was not complete compatibility between the service offered by NCSTRL and the requirements of the ERCIM digital library. NCSTRL has focussed on the provision of an efficient search and retrieval functionality for online documentation that places the emphasis on the rapid dissemination of technical

[2] The ERCIM scientific community currently consists of 14 national institutions speaking 13 different major European languages.

documentation, whereas our aim with ETRDL is to provide a full set of integrated library services. Neither is there total homogeneity between the NCSTRL and ETRDL collections: NCSTRL contains computer science literature; ETRDL extends its scope to include applied mathematics, an important area of activity within ERCIM. The ERCIM digital library has thus been implemented as part of NCSTRL collection, but with its own distinguishing characteristics and services.

3 Dienst as Infrastructure for ETRDL

A logical consequence of our decision that ETRDL should form part of the larger NCSTRL collection was the adoption of the Dienst infrastructure. Dienst is the term used by its developers to refer to a conceptual architecture for digital libraries, a protocol for communication in the architecture, and a software system implementing the architecture. The model of the Dienst digital document (DD) has two components: the bibliographic description and the "body" of the document; each document has a globally unique name (URN).

Here below, we give a brief outline of the main features of the reference implementation that we adopted[3], before discussing in the following section some of the measures we have taken in order to specialise it to meet the requirements of ETRDL.

The Dienst distributed digital library services can be logically divided in four classes:

- A Repository Service that provides the mechanisms for storage of and access to the digital documents.
- An Index Service that provides the mechanisms for the discovery of DDs. It stores meta information about documents in the collection. Queries submitted to this service return a set that contains the URNs of the DDs that match the query.
- A Meta Service that provides a directory of locations of all other services. This service provides the mechanisms to identify the index servers of a distributed digital library collection and to manage meta-information about these servers.
- A User Interface Service that provides a human front-end to the other services.

Each of these services is accessible via a well-defined open protocol - a set of service requests - that defines the public interface to the service. The service requests are specified by a signature and a semantics. The signature consists of a description of the request and the result formats. The service requests are implemented as protocol messages whose format is given in terms of i) the name of the service that is to handle the message, ii) the service request and iii) one or more, optional, arguments. All messages are encoded into URLs.

[3] The activity described here began in 1997 with Dienst version 4.1.9 and all observations refer to this version; both NCSTRL and Dienst have progressed since then.

The "openness" of this protocol means that it is possible to add new services to an existing architecture, to add new service requests to existing services, to specialise service requests and to replace an existing service implementation by alternative implementations.

A service is implemented by a server. There are no constraints on the way in which a server can be implemented. The only obligation is that it implements the functionality of the corresponding service, i.e. it accepts the service requests allowed by the protocol and processes them correctly with respect to the protocol semantics.

The open-architecture also allows the creation of different *federated digital library instances*. An instance is the result of aggregating a collection of servers, that communicate via the protocol. The functionality of the digital library instance is a result of the union of the service requests from the aggregated servers.

The version of Dienst on which we based our implementation had the following organisation for its servers (see Figure 1):

1. Master Meta Server (MMS) and Regional Meta Server (RMS). These servers constitute a distributed instance of the Meta Service. In particular:
   - the MMS stores information on all the institutions participating in a collection and on the servers that act as regional metaservers. It also stores a set of collection views, i.e. perspectives on collections specific to a region; each collection view identifies the list of index servers that should be used by the Dienst Standard Site servers in that region to process a query.
   - the RMS supplies its regional servers with the list of publishing institutions that form part of a collection, the list of servers implementing the index server to be contacted to process the queries, the list of servers that implement the repository service to be contacted for document retrieval. The metadata collected by the RMS are retrieved through regular calls to the MMS.
2. Dienst Standard Site (DSS). This server instances the functionality of the Repository Service, the User Interface Service and the Index Service for its own digital documents. Each DSS refers to just one RMS. For this reason, each DSS belongs to a single, specific region. When implementing the functionality described above, a DSS can only communicate with the servers in its own region. The only exception is when processing queries directed to authorities outside its own region; in this case, the DSS directs its query to its MIS.
3. Merged Indexes Server (MIS). This server instances the functionality of the Index Server for all the repository servers that are outside a given region. The meta-information collected is used to process queries directed to authorities not belonging to that region. The MIS communicates with the RMS in order to obtain the meta-information regarding its own region.

Dienst is the foundation for NCSTRL but its developers state that it can be extended and/or specialised to meet the needs of other applications. In the development of ETRDL we have been able to test this claim.

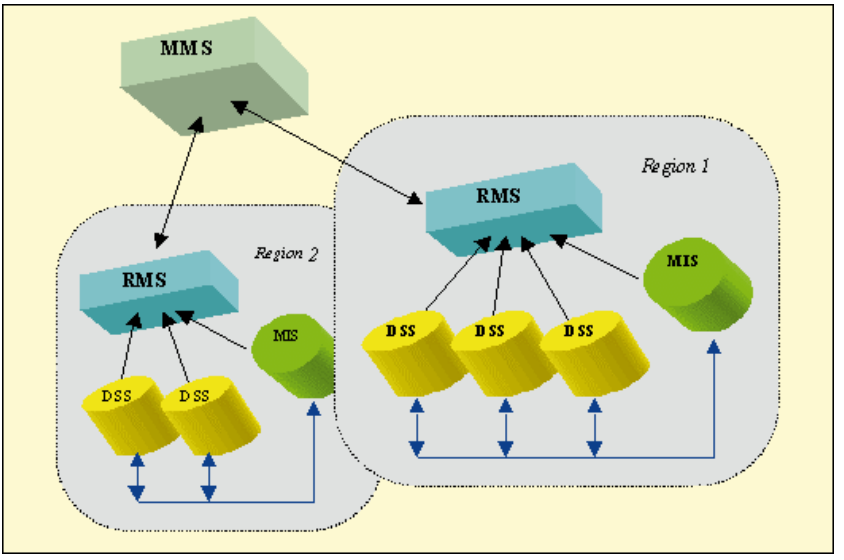

Fig. 1. Dienst Architecture

Our aim has been to implement the additional functionality required by the ETRDL services while maintaining compatibility with Dienst, in order to keep our status as a partner in NCSTRL. Thus, exploiting the Dienst characteristic of "openness", we have:

1. specialised the existing Dienst services;
2. extended the Dienst protocol, including a number of new service requests and specialising some of the existing ones;
3. extended the Dienst architecture, adding a new service.

4 Implementing ETRDL

Lagoze in [8], defines a DL as "a managed *collection* of digital objects (content) and services (functionality) associated with the storage, discovery, retrieval, and preservation of these objects". In this section, we discuss how the ETRDL collection and its services have been designed and implemented in the ERCIM digital library. It should be remembered that we had two objectives when constructing ETRDL: on the one hand we wanted to satisfy the requirements of the ERCIM users; on the other we wanted to ensure our *non-isolation* from the rest of our scientific community.

4.1 The ETRDL Collection

A DL collection has been logically defined as: "a set of criteria for selecting resources from the broader information space" [8]. Following this definition, we

have characterised the ETRDL collection by the following criterion: the set of documents whose publishing institutions have names that begin with the string "ercim". Obviously, a necessary assumption is that all the institutions participating in ETRDL have named their own collections according to this rule.

In accordance with our requirement of non-isolation, this collection must be visible both to ETRDL and to NCSTRL users. Furthermore, ETRDL users must be able to access the specific ETRDL services. The intuitive way to achieve this appeared to be to implement the ETRDL collection as a specialised sub-collection of NCSTRL. However, Dienst only recognises as sub-collections sets of documents belonging to a single publishing institution. The problem was that the ERCIM users had decided not to concentrate their collections into a single publishing institution for several reasons: i) each member institute wanted to manage its own documents, ii) a distributed management facilitates local specialisation, e.g. of the interface, submission policies, administration services, etc. It was thus necessary to find a way to add a collection to NCSTRL that was distributed over several publishers.

The solution we have adopted satisfies all but one of these conditions: the ETRDL collection is not visible as such to NCSTRL users who access from non-ERCIM sites. These users can access documents of the ETRDL collection, using the NCSTRL discovery tools, but they see them as belonging just to the local ERCIM publishing institutions, not as part of the larger ETRDL collection. On the other hand, the ETRDL interface provides the ERCIM user with a choice between three collections: the NCSTRL collection, the ERCIM collection and the collection of the local ERCIM institution (see Figure 2).

Users selecting the NCSTRL collection will access the standard NCSTRL search and browse functionality; those selecting, the ETRDL collection will access the specialised ETRDL services, while users accessing their local collection may have extra services available (e.g. an interface in the local language, additional discovery mechanisms, specialised administrative features).

In order to be able to constitute the ETRDL collection (and overcome the constraints imposed by Dienst), we had to choose between two alternative implementation strategies:

1. Introduction of a new specialised ERCIM Meta Service (EMS) in the architecture through the implementation of dedicated ETRDL meta-servers in each region. This service would provide mechanisms to identify the ETRDL index servers from the other index servers and to manage meta-information about these servers. Each EMS would retrieve information from the RMS of its region and simulate the behaviour of the RMS with respect to the ETRDL DSSs. The EMS must be transparent to the non ERCIM DSSs which continue to refer to the RMS. This approach implies implementing a number of extra servers, potentially one for each region. If, for example, an ERCIM partner registers in a region that does not contain other ETRDL servers, a new EMS must be introduced in that region. For this reason, this approach was judged impractical as it would considerably increase the work load.

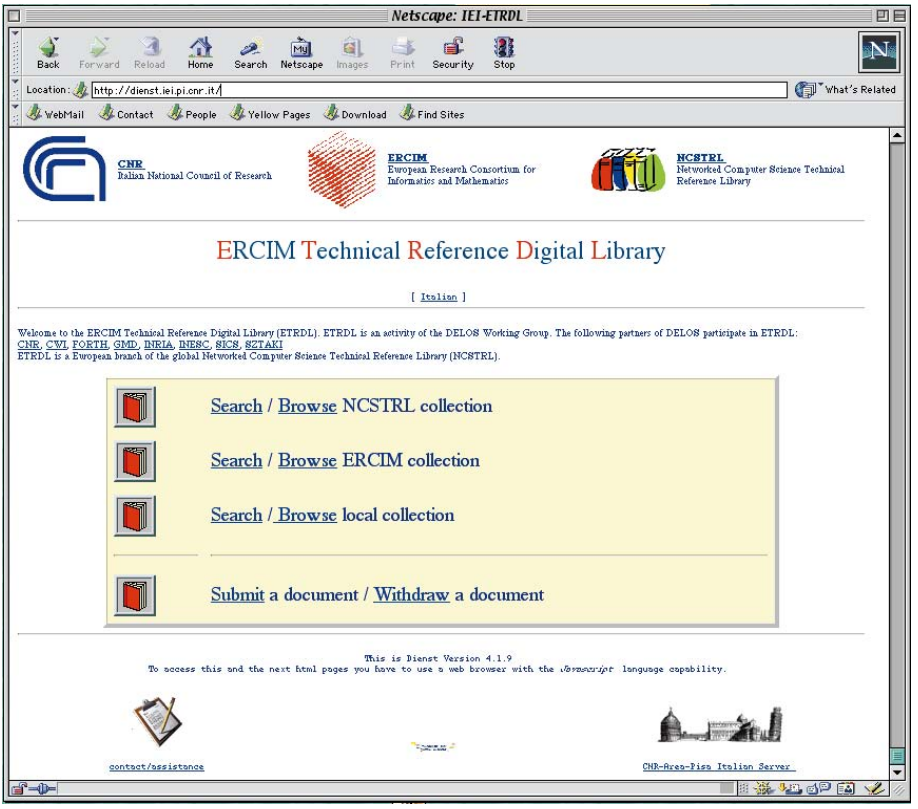

Fig. 2. The ETRDL home page

2. All the ETRDL servers must belong to the same region and communicate with the same RMS. The mechanism to identify the ETRDL index servers would thus be automatically guaranteed by the architecture. This solution requires that the ETRDL servers currently distributed throughout Europe would have to migrate to a single region and refer to a single RMS. This solution would appear to be in conflict with the principle underlying the implementation of the regional meta servers, created to guarantee connectivity. However, as there is generally an acceptable level of connectivity in Europe and as this solution is practically cost free, we have decided to adopt this strategy.

The adopted solution, which is based on the current static state of the regions, could result unsatisfactory if there are future developments of the architecture with respect to a dynamic management of regional membership in function of connectivity. We are thus investigating the feasibility of creating a mechanism for auto-identification of ERCIM index servers. Unlike the above approaches that

impose constraints on the system architecture, this solution would only involve modifying the ETRDL DSS.

In the next section, we describe how we have implemented the set of specialised ETRDL services.

4.2 The ETRDL Services

There are three main classes of ETRDL services:

1. search and retrieval
2. submission/withdrawal of documents
3. DL administration

As we have already stated, the aim of ETRDL is to provide its users with a digital library service satisfying their requirements. We have thus made the following extensions with respect to NCSTRL. The ETRDL search and browse service offers additional functionality by implementing subject searching and browsing, and providing users with local language interfaces. The submit/withdraw service is new, and aims at assisting the authors by providing facilities to classify their documentation (using classification schemes for both computer science and the mathematics) quickly, easily and correctly. The administration service is also new, and assists the librarians by providing mechanisms to manage the digital documentation efficiently.

Search and Retrieval Search operations rely on matching between user queries and document descriptions; the richer the descriptions the more successful the search. The standard metadata supported by NCSTRL (*author*, *title*, *abstract*) was not sufficient for the implementation of the more powerful search and browse functionality requested by the ERCIM users. The first step was thus to define the set of metadata to be associated with the documents in the ETRDL collection. We have selected a metadata format which represents an extension of the basic NCSTRL metadata set, in order to ensure NCSTRL to ETRDL interoperability, and which is also compatible with the Dublin Core metadescription standard [9]. The decision to comply with the Dublin Core standard was made so that integrated queries over different digital libraries using the DC metadata would be possible in the future.

The new fields introduced into the ETRDL bibliographic record are *subject*; *type*, *date* and *language*; *local language abstract*.

The inclusion of subject fields makes it possible to overcome a serious defect in most current retrieval systems. In recent years, the diffusion of powerful search engines capable of indexing full document text has lead to the idea that the task of subject classification, which requires much intellectual work by librarians, is perhaps obsolete and can be substituted by free keyword searching on indexes of document terms. The disadvantages of this type of searching, however - e.g. the false coordination of terms, the risk of very low recall values - are well known.

Therefore we decided that document contents should also be explicitly represented, by the authors, using descriptors from specialised controlled vocabularies: the ACM (for computer science) or the AMS (for mathematics) classification schemes. We also allow our authors to apply free terms, each consisting of one or more keywords, for those topics for which classification codes have not yet been introduced. It has been found that the frequent use of certain free keywords acts as a stimulus for their introduction into a controlled vocabulary.

The document type, date and language fields have been added to allow the selection of particular sets of results, i.e. to refine the results of a search operation. The user can restrict the display of results to given document types (e.g. technical reports, proceedings, pre-prints, theses, etc.), to documents published in a given year, or to documents in a given language (one of the languages of the ERCIM member institutions). This is done by selecting the desired type, date and language values from a set of pop-up menus.

An English abstract is mandatory for all documents. Documents in languages other than English will also have an abstract in the local language. The local language abstract field is used for these documents.

The Browse Service. The Dienst protocol provides two ways to browse the collection:

- by author (all authors, a range of authors, authors whose names begin with a given letter).
- by year (all years, a range of years, a given year)

In the traditional library, the user can browse through the subject catalogues in order to be acquainted with the material contained and to see what is available for a given argument. We wanted to provide a similar facility in ETRDL to give the users a starting point to investigate the contents of the collection and thus improve the precision of their queries. We thus added a new function: browse by subject terms (all terms, a range of terms, or terms beginning with a given character).

The browse function is implemented in the User Interface (UI) Service of the Dienst protocol. The introduction of the new service has implied modifying the Index Server, in order to create and index the subject terms, and extending the UI Service, to access the functions of browse by ACM and AMS classification codes and browse by keywords.

The Search Service. The Dienst protocol provides two functions for querying a collection: Simple Search and Fielded Search. The Simple Search returns all documents whose author, title, or abstract contain any of the terms entered. The Fielded Search takes as input a complex condition. Logically, it can be seen as decomposed in two parts: the set of publishing institutions on which the search is to be performed and the condition to be imposed. This is defined in terms of the author, title and abstract search fields, and by a boolean operator "AND" or "OR" linking the simple conditions specified on each field.

In designing the ETRDL search service, we have attempted to create a *homogenous work environment* for all users. Thus a user accessing the ERCIM digital library can choose to search the entire NCSTRL collection or only the ETRDL sub-collection. Both provide certain services and present them to the user in a certain way. If the service is to be efficient, then it is important that the user enters a familiar work environment, independently of the collection he has selected. For this reason, we have maintained the division introduced by NCSTRL between Simple Search e Fielded Search, even though we have modified them.

However, the search functions provided by NCSTRL were insufficient with respect to the requirements of the ERCIM users. In particular, for the field search we wanted:

1. to have a richer set of access points in order to raise the level of recall and precision;
2. to be able to build sets of results in order to support an incremental formulation of the query;
3. to be able to express more complex search conditions.

The first of these requirements derives mainly from the need to be able to search by subject - generally the primary need of a user of a library and, in our opinion, also of the digital library user. The second reflects the need to allow users to refine the results of a query, in order to achieve a higher precision. The third arises from the fact that with Dienst queries the relationship between the conditions imposed in different fields must always be the same, either "AND" or "OR", never a combination of the two. For example, the query: (Author = "Sebastiani" AND (Title $\cong$[4] "maintenance" OR Abstract $\cong$ "maintenance")) is not acceptable. This limitation is made more restrictive by the fact that it is not possible to maintain a result set for a first query, on which to formulate a second query. Thus a query of this type cannot be executed even in two steps. Furthermore, it is also impossible to apply field selectors (e.g. on the language of the documents to be retrieved, or on their date, etc.) using the Dienst search mechanism.

Figure 3 shows the User Interface for the ETRDL Fielded Search. ¿From the figure it can be seen that Fielded Search has three logical components: the bibliographic fields (Author, Title, Abstract, Subject) and two radio buttons to specify whether the values entered in the fields should be "ANDed" or "ORed"; three selectors to filter documents according to Type, Date, Language; a menu to select one or more collections on which to perform the search; and a check box to select all collections.

In order to meet the requirements specified above, we have modified both the metadata associated with the ETRDL document with respect to NCSTRL and the Dienst protocol.

In doing this it was essential to maintain interoperability with NCSTRL. The service requests sent from servers external to ETRDL had to be accepted and the semantics of the Dienst protocol had to be respected.

[4] By $\cong$ we intend a matching of the input with indexed strings in the bibliographic field that contain the keyword entered as substring, in whatever position.

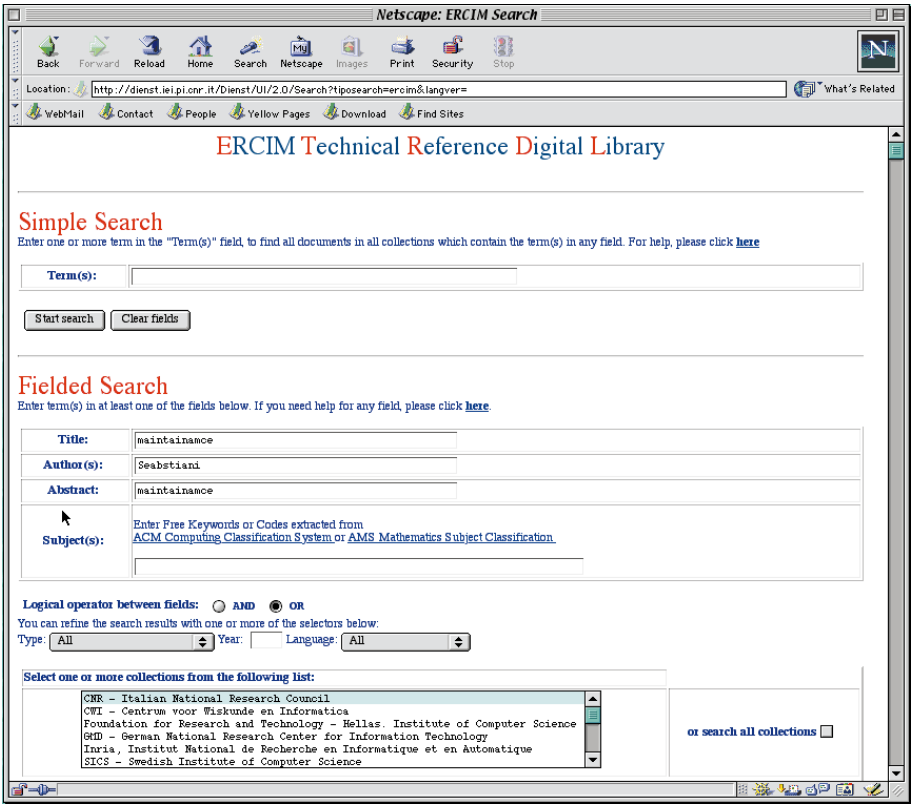

Fig. 3. The Fielded Search

To obtain this goal, we have modified Dienst respecting a principle that we have called *specialisation*. Given a service request $SR(p_1, \ldots, p_n)$, where SR is the name of the service request and p_1,..., p_n are its parameters, that returns a result R and has a behaviour B, we define as the specialisation of SR a service request that satisfies the following conditions: - has name SR - has parameters p_1 ,... , p_n, p_{n+1}, ... , p_k where p_{k+1}, ... , p_k are optional - the type of result returned is a subtype of R - the new behaviour has the same effect as B when the service request is invoked by the parameters p_1,..., p_n.

For example, the Dienst service request that we have modified most extensively while respecting the above principle is the SearchBoolean. This is a service request of the Index Service. In its original version, it takes as input parameters a boolean operator and a list containing one or more of the author, title and abstract fields, each of which may have an associated value. A list of records is returned, each record contains the title, author, date, and the URN of a document satisfying the conditions expressed by the service request. We have specialised the service request by extending the list of input parameters with additional fields: *subject*, *type*, *year* and *language*. These last three can be employed as selectors

on the result set deriving from a search imposed using the other parameters. The behaviour of the specialised service request is the following:

- if none of the additional fields has been specified, it is equal to the behaviour of the original SearchBoolean
- if the subject field has been specified but no selector is indicated, it is similar to the behaviour of the original SearchBoolean with the difference that the condition imposed as value of the subject field is also evaluated
- if at least one selector has been specified, it first evaluates conditions imposed on the *author*, *title*, *abstract*, *subject* fields; a temporary transparent result set is constructed to which the conditions specified by the selectors are applied.

It must be observed that the function implemented in this way does not satisfy our initial requirements as it still does not allow us to specify certain significant boolean structures, such as queries in which the relationship between the conditions imposed in author, title, subject and abstract search fields are a combination of the "AND" and "OR" boolean operators. This problem cannot be overcome with the simple specialisation of the Dienst reference implementation. In the new version of ETRDL now under implementation we are attempting to resolve this problem by extending the protocol.

The implementation of the ETRDL search service has also necessitated changes to the Index Service in order to support the indexing of the new fields and their management.

Submission and Withdrawal of Documents In the Dienst reference implementation used by ETRDL, in order to submit a new document to the collection, the user must:

- fill in a Web-based bibliographic record for the document
- submit the document in a file via FTP

This means that the user is obliged to use separate applications (the web browser and the FTP client) for a single submission operation, and the administrator has to maintain an ftp site for new document files and must manage document insertion in the collection by a shell command line.

In ETRDL we wanted to offer our information providers with a single work environment for the compilation of the bibliographic record and the submission of the document in the chosen format (PS, PDF, TXT, HTML, TIFF). The main features of this work environment are a simple and controlled access and a guided compilation of the bibliographic document.

We have thus introduced a new service (Administrator and User Interaction - AUI service) that supplies mechanisms for *submit* and *withdraw* of documents and for the DL *administration*. This service has been implemented by a set of modules that interact with the services provided by the architecture through a set of calls to service requests of the Dienst protocol. These modules are web-based.

The submit procedure guides the user when inserting the metadata and file associated with the document. The system will perform an automatic check on the formal correctness of the contents of the obligatory fields. During compilation of the submit form the user can access directly the Computing Classification System (CCS) and Mathematics Subject Classification (MSC) codes/descriptors (via a hypertextual link to the ACM and AMS sites). This service operates in two stages: the author submits the compiled bibliographic record and the document file to the DL; the administrator is responsible for its approval, and the actual insertion of the document and its associated metadata in the collection.

The withdraw procedure guides the user when deleting his documents from the ETRDL. This service has been included to support the inclusion in ETRDL of temporary documents, e.g. pre-prints that must be removed when the article in its definite form is published elsewhere. Similarly to submit, this service operates in two stages: the author communicates the document to be withdrawn and the motivation to the DL administrator; the administrator is responsible for the actual withdraw/delete. The author/administrator communication takes place through the automatic generation of e-mail messages.

Additional services provided by the submit and withdraw procedures are: a contextual online help to explain the syntax and semantics of each field; an e-mail access point to contact directly the librarian or system administrator if extra help is needed; a bilingual user interface in English and the local language.

Access to the submit and withdraw procedures is controlled and only authorised users can insert new documents or ask for the withdrawal of existing ones.

DL Administration As mentioned in Section 4.1, the ETRDL collection consists of a set of independent subcollections; each of which consists of the documents produced by a single publishing institution. Each publishing institution is responsible for the administration of its own collection. The Dienst reference implementation used by ETRDL provides basic utilities to help the System Administrator to manage the collections, e.g. when inserting or removing a document from the local repository. These utilities require that the operator is directly connected to the server system and that he/she knows the Unix command language.

However, in most of the ERCIM institutions the librarians will be responsible for verifying the formal correctness of the documents and bibliographic records submitted and will assign the identification number. As the librarian is not usually a systems expert, the ETRDL administration service had to satisfy the following requirements:

- the interface had to be platform independent - the administrator could access the system via a Web browser
- the administrator is able to check that the documents and bibliographic records submitted were formally correct
- the administrator is able to communicate with the information provider via e-mail, if necessary.

We have thus introduced a new administration procedure, part of the AUI Service, that supplies mechanisms to administer the DL by means of an Web-based integrated work environment. As with the submit and withdraw procedures, access to the administration procedure is controlled and limited to authorised users. This environment is very similar to that designed for the information providers and seekers.

The administration procedure allows the librarian to insert or delete documents from collections to which he/she has access. Whenever an information provider submits a new document or requests the withdraw of an existing document, the procedure notifies the librarian through the automatic generation of an e-mail.

In this work environment, the administrator can i) insert new documents in the collections he/she is responsible for; ii) eliminate outdated documents; iii) limit access to bibliographic information only if the document has been published elsewhere; in this case, reference to the publication is given.

4.3 The ETRDL User Interface

The extensions we have made to the basic Dienst service described above have clearly affected the interface design decisions. The ETRDL collection is a specialised sub-collection of NCSTRL, the user interfaces must reflect this. In this section, we mention briefly the main issues that have been considered when developing the ETRDL user interfaces and outline the difficulties we encountered (for a more detailed discussion, see [10]).

A first major decision regarded the system Home Page(s), i.e. the initial access points. ETRDL is an integrated work environment, and thus the Home Page must provide access to various types of functionality. We have also had to allow for different "views" on our collection: public vs. restricted; centralised vs. local. The contents of the ETRDL collection can also be accessed by NCSTRL users, but such users can only access the search and browse functionality of the separate participating institutions; the ETRDL information provider and administrator services are transparent to them. However, for the ERCIM user, ETRDL is a distributed collection, consisting of the set of the local ERCIM collections. The local collections are maintained on the local servers of each partner institution. This has comported the implementation of two levels of Home Pages. A centralised access point has been provided to the system through the DELOS Web site, whereas a local home page is installed on each local server. The "views" provided by these two different Home Pages respect the needs of the potential users at each site (centralised and local) and thus provide different points of entry.

The *Centralised Home Page* is in English only and has been designed for IT information users in general, not necessarily from ERCIM (see Figure 4). For this reason, it provides links to pages that describe the objectives of the ETRDL, to on-line documentation, and to other relevant Web sites. It allows the user to access the ETRDL through one of the local servers. Clicking on the logo of a given institution will open the relevant local home page interface. Our

initial intention was to provide direct access to the ETRDL collection (with the extended set of functionality) from the centralised Home Page. However, it was decided that this was not realistic; it implied maintaining a centralised server as well as the local ones. The user is thus informed that in order to search the ERCIM DL collection, he should select one of the local servers. At the same time, he is given a choice of language as each local server will maintain interfaces in English and in the local language (see Section 5 below).

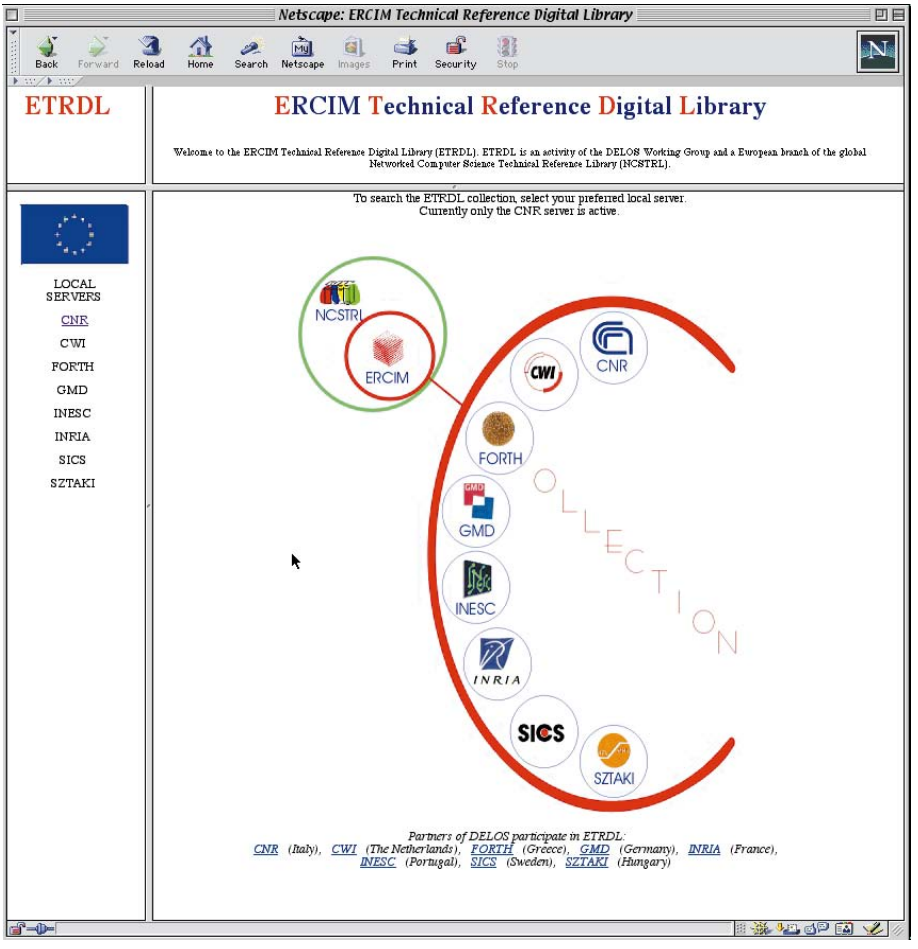

Fig. 4. The Centralised Home Page

The *Local Home Page* interface caters simultaneously for two user classes: information users and information providers by offering two options: search/browse any collection; submit/withdraw a document to/from a local collection. From the local home pages, the search and browse functions can be activated over the en-

tire NCSTRL collection, over the ERCIM collection, or over the collection(s) of the local institution (see Figure 2). In each case, the user is not only accessing a different collection (or sub-collection), but is provided with a different perspective on the information, depending on the functions that have been implemented at that particular level. When searching on the ETRDL or the local collections, the user can switch between user interfaces in English or his/her own language. Online helps in both languages are available.

The Administrator Home Page is transparent to the general public and accessible by authorised persons only. The main functions to be provided by the administrator interfaces were decided and defined in agreement with all the partner institutions and are described above. However, no common administrator interfaces have been designed; each local institution implements them according to local requirements.

We encountered a number of problems when implementing our interfaces, mainly due to the fact that the UI Service of the Dienst protocol was not designed to support a multilingual interface. This meant a specialisation of all the UI Service modules. The aim was to simplify the operations of localisation by the ERCIM partners. The UI Service has thus been rendered parametric in function of language (see 5.1 below). This operation of specialisation has been complex as the reference implementation of Dienst architecture incorrectly conflates the functions of the user interface service with query routing.

5 Implementing a Multilingual Interface

One of the points that most distinguishes the ETRDL service from that of NCSTRL is the need to handle multiple languages. As already mentioned, the ERCIM scientific community currently consists of 14 national institutions speaking 13 different major European languages, and multilinguality is thus an issue of great relevance. While it is true that a considerable proportion of the technical documentation in the institutions is produced directly in English, provisions had to be taken to enable documents in languages other than English to be included in the collections, and to en able users that do not have a high competence in English to use the system in their own language.

As far as ERCIM is concerned, this is not only a practical but also a strategic issue, going beyond the restricted domain of computer science. The diversity of the world's languages and cultures gives rise to an enormous wealth of knowledge and ideas. It is thus essential that we study and develop computational methodologies and tools that help us to preserve and exploit this heritage. The ETRDL collection constitutes a very convenient test-bed on which study technologies for multilingual information access.

Two basic issues are involved:

1. Multiple language recognition, manipulation and display.
2. Multilingual or cross-language search and retrieval.

The first activities of ETRDL in this area have been aimed at (i) implementing user interfaces capable of handling multiple languages and (ii) providing very basic functionality for cross-language querying.

5.1 Multilingual Access

It was decided that each national site should be responsible for localisation, i.e. implementation of local site user interfaces in the national language as well as the CUI in English. At the very simplest level, this means translating the common system interfaces (including the on-line helps) into the local language. For the system home pages, at each local site we maintain a version in English and in the local language; the user can choose which to activate using the language link on the local home page. All the other interfaces of the system are generated automatically during run-time. The system code thus includes a language variable, which determines whether the procedures should invoke interfaces and system messages in English or in the local language, depending on the initial choice made by the user. Of course, localisation also implies providing the metadata field descriptors in the local language as well as in English. The group has thus been involved the activity of the Multilingual Dublin Core group [11] and the descriptors employed in each local language will conform with the decisions taken by this group.

More complex at both the interface and the system level is the question of being able to handle and visualise multiple character code sets. Each document submitted to the collection is tagged for language. Mechanisms will be provided for the local display and printing of non-Latin-1 languages (this has been implemented at ICS-FORTH but is not yet operational on-line). In the future, we must decide whether to move to Unicode.

5.2 Querying in Languages other than English

At the level of the local collections, users must be given the opportunity to formulate queries in the local language and restrict their search to documents in that language. We are thus implementing mechanisms for the indexing of documents in languages other than English. This implies managing non-English sets of characters and stop word lists. Another question to be tackled is that of handling accented characters; the Latin-1 character code set caters for most European languages and is thus able to encode and represent all the accented characters of these languages. However, European users are not always able to input all of these characters on their keyboards. This is not a problem for local language querying but can become a problem when querying over the entire ETRDL collection for authors with "foreign" names, e.g. a user querying for documents with Author = Müller, might enter Müller, Muller, Mueller. We will have to study and implement robust search and indexing mechanisms for the author field to handle such cases.

Cross-language Querying. A simple form of cross-language querying is already possible using the controlled vocabulary (ACM/AMS) terms. All documents in

the ETRDL, in whatever language, classified using this scheme, can be searched. As authors are also requested to include an abstract in English, English free term searching over documents in any language is also possible. INESC has developed an LDAP service with a multilingual repository for the ACM and AMS classification systems (currently implemented in English and Portuguese), which they intend to integrate in their version of the ETRDL system [12]. We are now investigating other strategies for cross-language querying.

6 Next Steps

In this paper, we have described the first implementation of the ERCIM digital library developed as part of the NCSTRL network, employing and adapting the Dienst infrastructure. Our reference implementation of Dienst provides a simple, monolingual free-text search service. We have extended and specialised this service by adding controlled vocabulary search facilities, multilingual interfaces, mechanisms for the guided online submission and withdrawal of documentation, and for the administration service. Our aim has been to go a step further than NCSTRL, offering our users a complete set of digital library services integrated in a homogenous work environment, or "work centre" according to the concept introduced in [13]. The difficulties we have encountered have been mainly caused by our desire to provide this specialised DL service within a much more extensive network, offering lesser functionality. The need to guarantee compatibility with NCSTRL has meant that it has not been possible to satisfy all our initial requirements and certain compromises have had to be accepted, above all the fact that the ETRDL collection cannot be viewed as such by NCSTRL users.

As stated in the paper, our reference implementation has been Dienst, version 4.1.9. However, a new (and final) version of Dienst has now been developed at Cornell. This version provides functionality to order the results (including ranking). NCSTRL has adopted this new version. If we want to maintain the same level of compatibility with NCSTRL, we must produce a new version of ETRDL which incorporates the new functionality.

In any case, our intention in the future is to continue in the direction of a dedicated "work centre", implementing a series of more sophisticated user-oriented services. These include tools for semi-automatic document classification, procedures for free-text non-English and cross language querying, mechanisms for indexing and querying mathematical formulae, systems for document filtering and user profile modelling, procedures for the automatic classification of existing collections, gateways to other online digital libraries and catalogues for related areas of interest.

Some of these procedures are already in an advanced state of development, e.g. the semi-automatic classification tool and the user profiling, others are still at the level of "wish-list". The existing ETRDL collection functions as a test-bed for the study and development of these advanced services.

References

1. Biagioni, S., Borbinha, J., Ferber, R., Hansen, P., Kapidakis, S., Kovacs, L., Roos, F., Vercoustre, A.M.: The ERCIM Technical Reference Digital Library. In: Nikolaou, C., Stephanidis, C. (eds.): Research and Advanced Technology for Digital Libraries. Lecture Notes in Computer Science, Vol. 1513. Springer-Verlag, Berlin Heidelberg New York (1998) 905-906. (http://www.iei.pi.cnr.it/DELOS/EDL/ETRDL98.html).
2. ETRDL Demo Decription: Handout distributed at ECDL'98, Crete, Greece, September 1998 (http://www.iei.pi.cnr.it/DELOS/EDL/JPEG/etrdl0998.html).
3. Belkin, N.: Understanding and Supporting Multiple Information Seeking Behaviors in a Single Interface Framework. In: Proceedings of Eighth DELOS Workshop: User Interfaces in Digital Libraries. DELOS Working Group Report No.99/W001, (1998) 11-18.
4. Networked Computer Science Technical Reference Library. http://www.ncstrl.org
5. Leiner, B.M.: The NCSTRL Approach to Open Architecture for the Confederated Digital Library. In: D-Lib Magazine (December 1998) (http://www.dlib.org/dlib/december98/leiner/12leiner.html)
6. Lagoze, C., Shaw, E., Davis, J.R., Krafft, D.B.: Dienst: Implementation Reference Manual. Cornell Computer Science Technical Report TR95-1514 (http://cstr/cs/cornell.edu:80/Dienst/UI/2.0/Describe/ncstrl.cornell/TR95-1514)
7. Lagoze, C., Davis, J.R.: Dienst: an Architecture for Distributed Document Libraries. In: Communications of the ACM, 38 (4) (April 1995), 45.
8. Lagoze, C., Fielding, D.: Defining Collections in Distributed Digital Libraries. In: D-Lib Magazine, (November 1998). (http://www.dlib.org/dlib/november98/lagoze/11lagoze.html)
9. Dublin Core Metadata Element Set: Resource Page. http://purl.org/metadata/dublincore.
10. Baldacci, M.B., Biagioni, S., Carlesi, C., Castelli, D., Peters, C.: Implementing the Common User Interface for a Digital Library: The ETRDL experience. In: Proceedings of Eighth DELOS Workshop: User Interfaces in Digital Libraries. DELOS Working Group Report No.99/W001, (1998) 63-72. http://www.ercim.org/publication/ws-proceedings/DELOS8/baldacci.html
11. Multilingual Dublin Core: http://www.cs.ait.ac.th/ tbaker/dc-multilingual.html
12. Freire, N.: Integration of Multilingual Classification Systems with the Dienst digital library system. In: Proceedings of Eighth DELOS Workshop: User Interfaces in Digital Libraries. DELOS Working Group Report No.99/W001, (1998) 27-31. http://www.ercim.org/publication/ws-proceedings/DELOS8/freire.html
13. Cousins, S.B., Paepcke, A., Winograd T., Bier, E.A., Pier, K.: The Digital Library Integrated Task Environment (DLITE). In: Proceedings of the Fourth Annual Conference on the Theory and Practice of Digital Libraries, (1997).

Acknowledgements

The development of ETRDL is the result of a collaborative activity, the implementation was the responsibility of IEI-CNR; the authors would like to gratefully acknowledge the assistance of the other ERCIM participants, both in the initial formulation of the specifications, and in the feedback received as a result of testing the first prototype. They would also like to thank the developers of the Dienst system and, in particular, Carl Lagoze and David Fielding for their generous assistance and advice.