PROCEEDINGS

of the CNR-GMD – Workshop, Berlin, March 9-10, 2000

Edited by

Eckart Bierdümpel, Stefan Jähnichen, Piero Maestrini and Luca Simoncini.

National Research Council (CNR) of Italy

GMD National Research Center for Information Technology

The CNR-GMD Virtual Library Project

Antonella Andreoni, Maria Bruna Baldacci, Stefania Biagioni, Carlo Carlesi, Donatella Castelli, Pasquale Pagano, Carol Peters, Serena Pisani, CNR-IEI; Reginald Ferber, Matthias Hemmje, Barbara Lutes, GMD-IPSI

Abstract

We describe the implementation of a Digital Library (DL) service for the access and dissemination of grey literature in computer science and applied mathematics as part of the CNR-GMD collaboration. The DL has been developed as a specialized sub-collection of the US Networked Computer Science Technical Reference Library (NCSTRL). The service has been developed extending the Dienst system used by NCSTRL, to meet the requirements of the European scientific community. The results achieved during the project lifetime are described and intentions for future developments are outlined.

1. Objectives

The goal of the GMD-CNR virtual library project has been to build an infrastructure for the dissemination of the technical documentation produced by these two organisations, through the circuits of the international scientific community. The project has been carried out as part of a more extensive initiative: the building of a European digital library for grey literature in computer science and applied mathematics: the ERCIM[1] Technical Reference Digital Library (ETRDL). Researchers from CNR and GMD thus collaborated with scientists from other ERCIM institutions in the development of the infrastructure.

Three main objectives have driven the development of ETRDL:

- it must encourage the exchange of ideas and dissemination of results between researchers working on similar problems around the world;
- it must respond to the needs of a specific user community, i.e. the ERCIM librarians and scientists;
- it must allow the implementation of functionality to satisfy specific local requirements of the two institutions.

There is also a secondary requirement: it should provide a testbed that can be used by scientists to experiment DL technologies.

In the development of ETRDL it was decided to follow the direction of an initiative already underway in the US: NCSTRL - the Networked Computer Science Technical Reference Library [1]. This decision was motivated by the fact that NCSTRL has extensive visibility, with a large number of European and North American participating institutions, and also because NCSTRL adopts the Dienst system [2],[3], for dissemination, searching and accessing its documents. Dienst is an open system, and can thus be extended to meet the needs of other applications.

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 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. The ERCIM digital library has thus been

[1] The European Research Consortium for Informatics and Mathematics currently consists of 14 national member institutions working in information technology and applied mathematics, and includes GMD for Germany and CNR for Italy.

implemented as part of NCSTRL collection, but with its own distinguishing characteristics and services.

2. User Requirements

From the very first discussions, between the ERCIM librarians and scientists, it was clear that what was needed was a complete digital library service covering the needs of information providers, seekers and administrators and also that each institution had its own particular requirements with respect to such a service.

The services offered were to include functionality for simple and advanced search facilities; acquisition and subject classification of documents; on-line controlled submission, updating and deletion of documents. In addition, many of the ERCIM institutions wanted to have the possibility of using the system not only in English but also in their local language. This meant implementing user interfaces in the different European languages commonly spoken by the ERCIM community, and being able to store, query and retrieve documents in these languages. The contents of the library were defined as any type of reference material, not subject to copyright, in the fields of the computer science and applied mathematics. Each document – whatever the source language – should include an abstract in English, whereas for English documents, the addition of abstracts in another language was an optional facility.

As each institution had its own consolidated practices with respect to the management of its technical documentation, the definition of the DL service had also to be flexible; a common set of core criteria would guarantee interoperability but there had to be room for flexibility at the local level. For more complete discussions of the ETRDL user requirements at the local and ERCIM levels, see [4, 5].

3. Technical Approach

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

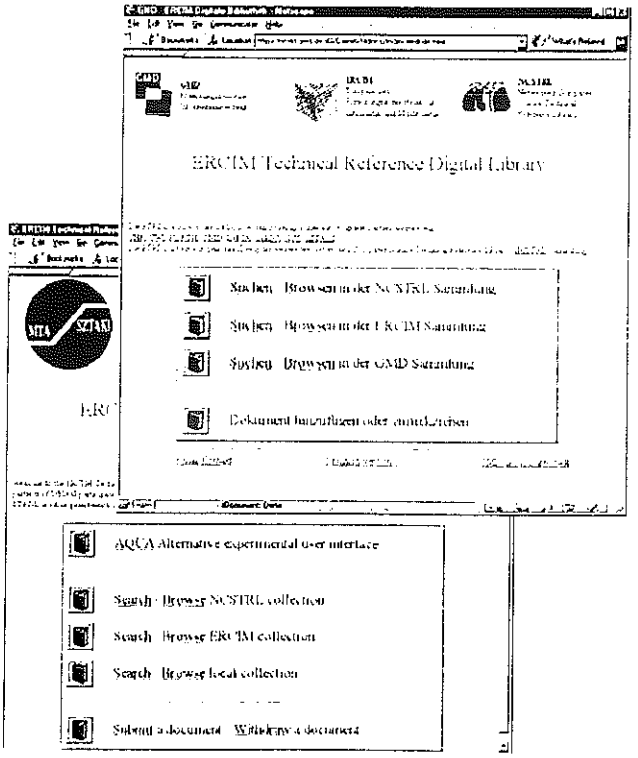
- A Repository Service that provides the mechanisms for storage of and access to the digital documents.
- An Index Service that provides the mechanisms for the discovery of DDs. It stores meta information about documents in the collection. Queries submitted to this service return a set that contains 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 "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 Dienst Standard Site (DSS) instances the functionality of the Repository Service, the User Interface Service and the Index Service for its own digital documents. A set of intercommunicating DSS and DLS constitute a *federated digital library*. The Dienst architecture also includes Regional Meta Servers and Merged Index Servers. The former provide meta-information with respect to the DSS, such as the location of the indexes and repositories in the region, the latter provide back-up index servers. They are used to improve overall connectivity by reducing the number of interconnections between servers. The existence of these servers has permitted us to create collections at distinct levels to meet the set of central (ERCIM) and local (i.e. GMD and CNR) requirements. ETRDL has been implemented as a federation of modified DSS belonging to the same region, adopting the same naming schema (ercim.xxx), and enabling the same core set of functionality. Each ERCIM institution (publishing authority) that participates in the project has instanced a modified DSS as a local server, managing one or more local collections, and also providing locally developed functionality. At the same time, each publishing authority member of ETRDL is also part of the NCSTRL federated digital library. A description of the reference implementation and the measures taken in order to specialise it to meet the requirements of ETRDL can be found in [6].

4. Accessing ERCIM and Local Collections

ETRDL has been developed within the context of the DELOS Working Group[2]. Public access to the service is provided through the DELOS Website, whereas a local home page is installed on each local server.



[2] The goal of the DELOS WG (1996-1999) has been to promote European research in DL-related areas. From 2000, the activities of DELOS will be continued and extended in a Network of Excellence (NoE), a thematic network on digital libraries funded by the European Commission.

The «views» provided by these two different Home pages respect the needs of the potential users at each site: centralised and local.

The *Centralised Home Page* appears in English, the common language. The user must select a local server to enter the system. The activation of a given server implies a choice of collection and collection services. In most cases, the choice of the local server will also offer a choice of interface language between the English and the local language.

The *Local Home Page* interface implements the commonly agreed format and set of services, but is customisable by each institution with respect to language and to any additional services to be provided locally. This can be seen in the figure which shows the SZTAKI (English switch activated) and GMD (in German) local home pages. The GMD local interface provides additional buttons to send e-mail directly to the system administrator and to access the GMD-IPSI institute (particular requirements of GMD).

In ETRDL, the functionality offered by Dienst has been expanded in order to implement the complete set of digital library services requested by the ERCIM institutions. A description of these services follows below.

5. The Browse Service

The Dienst protocol provides two ways to browse the collection: by author and by 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 for investigating the contents of the collection and thus improve the precision of their queries. We thus added a new function: browse by subject terms.

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 (free terms, CCS and MSC classification codes), and extending the UI Service, adding BrowseKeywords and ListKeywords service requests, with analogous signature and semantics to the *Browse by Authors*.

6. The Search service

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, as ETRDL is logically a part of NCSTRL, we have maintained the division introduced by NCSTRL between Simple Search and Fielded Search, even though we have modified them.

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.

With respect to the requirements of the ERCIM users, these functions were insufficient. 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.

In order to extend the services with respect to NCSTRL, the metadata set had to be extended and the Index Service altered to support the indexing of the new fields and their management. Some procedures and verbs of the standard Dienst protocol also had to be modified.

7. Multilingual Information Access

The version of Dienst used by NCSTRL does not support languages other than English. However, the DL must be able to cater for different European languages used by ERCIM institutions in order to provide user-friendly access for users not familiar with English. The user interface server of the ETRDL system package has thus been made parametric with respect to language. A simple and safe mechanism is provided to permit the instantiation of the interface in the local language as well as English. Each national site is responsible for this instantiation and for the translation of the user interfaces into their own language(s). Documents included in the system can be in any of the languages supported. The bibliographic record associated with the document must specify the language of the document and include an abstract in English and an abstract in the language of the document. Users can query the system for information in languages other than English by entering free terms in the abstract field and selecting the language of interest from the associated. Separate indexes are maintained for English and for other languages. The complete Latin-1 character set (ISO_8859-1) is installed so that all diacritic characters can be viewed and searched correctly. A simple form of cross-language querying is possible using terms from the controlled languages (ACM/AMS). All documents in the ERCIM collection, in whatever language, classified using this scheme, can thus be searched. As all documents must have an abstract in English, English free term searching over documents in any language is also possible.

8. Document Submission

Document submission and withdrawal are services that are very much dependent on local factors. The system has been released with default mechanisms. It is up to each local institution to decide if and how they support them. However, common sets of mandatory fields and document formats have been determined for the bibliographic record and the document, respectively. The following comments refer to the default submission procedure. For each field, on-line helps are available to assist the compiler of the bibliographic record during submission. Subjects must be assigned to each document, either from the ACM or the AMS classification schemes, available on-line, and/or using free keywords. If compilers need assistance in assigning the correct subjects, they can contact the local librarian using the e-mail link provided. A series of automatic checks are performed by the system on the formal correctness of the completed bibliographic record and messages are sent to the compiler if problems are signaled. Once the submission form is compiled, the system displays it to the information provider and requests confirmation before it is sent to the system administrator/librarian.

9. Document Administration

We have introduced a new default administration procedure that uses a Web-based integrated work environment. The administration procedure allows the librarian to insert or delete documents from collections. Access is controlled and limited to authorised users. 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 librarian can i) insert new documents in his/her collections, ii) eliminate outdated documents; iii) limit access to bibliographic information only, once a document is published, providing the reference.

Each institution can, however, introduce its own administration procedures, according to its particular needs and customs.

10. Next steps

We have described the first implementation of the ERCIM digital library, implemented as part of the NCSTRL service, which uses a modified version of Dienst 4.1.9. A new version of Dienst (Dienst 5) is due to be released shortly. It is our intention to migrate to this version and to develop an extension to the basic ERCIM library service. Our aim is to offer our users a complete set of digital library services integrated in a homogenous work environment. We intend to augment the DL testbed functionality by adding the services for: i) advanced user interfaces; ii) multimedia data support.; iii) automatic personalized information dissemination support.

This new version of our DL system will provide the infrastructure for the service to be provided by a new project, in which this collaboration between CNR and GMD will be continued and extended. The primary goal of this project will be to provide a digital library service to support networked scholarly communities.

11. References

1. Networked Computer Science Technical Reference Library. http://www.ncstrl.org
2. Lagoze, C., Shaw, E., Davis, J.R., Krafft, D.B. (1995). 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)
3. Lagoze, C., Davis, J.R. (1995). Dienst: an Architecture for Distributed Document Libraries. In: *Communications of the ACM*, 38 (4) (April 1995), 45.
4. Baldacci, M.B., Biagioni, S., Carlesi, C., Castelli, D., Peters, C. (1998). 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, 63-72.
   http://www.ercim.org/publication/ws-proceedings/DELOS8/baldacci.html
5. Baldacci, M.B., Biagioni, S., Carlesi, C., Castelli, D., Pagano, P., Peters, C, Pisani, S. (1999). The ERCIM Technical Reference Digital Library: Meeting the Requirements of a European Community within an International Federation, *D-Lib*, Vol. 5(12).
   (http://www.dlib.org/dlib/december99/peters/12peters.html)
6. Andreoni, A. Baldacci, M.B., Biagioni, S., Carlesi, C., Castelli, D., Pagano, P., Peters, C (1999). In . S. Abiteboul A.-M. Vercoustre (eds.). *ECDL'99 Proceedings*. Lecture Notes in Computer Science Vol. 1696, Springer.