In Proc. of IEEE Advances in Digital Libraries (**ADL 2000**)

Building a Digital Library of Newspaper Clippings: The LAURIN Project

Diego Calvanese, Tiziana Catarci, Giuseppe Santucci
Dipartimento di Informatica e Sistemistica
Università di Roma "La Sapienza"
Via Salaria 113, 00198 Roma, Italy
lastname@dis.uniroma1.it

Abstract

The field of digital libraries has been attracting a lot of research efforts during the last years. Many interesting projects have been started, dealing with the various open issues arising in the field. However, no project has specifically taken into account the problem of building a digital library of newspaper clippings. It is well known that a huge part of cultural knowledge is stored in the newspapers of yesterday. Since newspapers are not always easily accessible, special clipping archives were created in the 20th century. People interested in newspaper information benefit from these archives because the work of selecting, cutting and indexing articles is done by specialists. In order to maintain their important position in the information market, clipping archives should be able to integrate their special skills (such as professional knowledge and experience in gathering and treating newspaper information) into the new technologies of the information society. The EU-funded LAURIN *project will carry out the preliminary work necessary for an efficient and smooth shift from the "analogue" clipping archive to its "digital" successor. In order to effectively accomplish this hard task, the* LAURIN *Consortium has gathered a significant number of libraries, which are acting as final users and test sites and are continuously driving the system design and development with requirements, suggestions, testings, and criticisms. This paper presents the* LAURIN *design methodology, the main user and organizational requirements for a clipping digital library, and the overall architecture of the* LAURIN *system.*

1 Introduction

The terms "digital" and "library" have been very often coupled together in the last years. Digital libraries should represent the evolution and extension of traditional physical libraries in terms of the available information, both in formats and in extension. There are many kinds of libraries: among them those on newspapers are quite important because newspapers represent a detailed report on the cultural heritage of a country. Moreover, since newspapers are not always easily accessible, special clipping archives were created in the 20th century. In these archives specialists select, cut, and index articles, making them available to people interested in newspaper information. The wide range of subjects treated in newspapers and the great number of different newspapers made it necessary for these clipping archives to focus on particular topics. Even in the future when nearly all newspapers will have online editions with online archives attached to them, clipping archives will still be important. In order to maintain their important position in the information market, clipping archives should be able to integrate their special skills (such as professional knowledge and experience in gathering and treating newspaper information) into the new technologies of the information society, shifting from the "analogue" clipping archive to its "digital" successor.

Compared with book-oriented digital libraries, clipping libraries are more wide and unstructured, since there are not specific standards to collect and classify newspaper clippings. The subjects and content of a clipping are completely heterogeneous, as the origin of a clipping is: it could be a small article, some photos with some text as caption, or a whole article with several diagrams and photos spanning several pages. Given this diversity, it is extremely difficult to come up with a general clipping catalog system, whereas many specific clippings can be retrieved in a library systematically interested in some subjects or in a library that institutionally collects and catalogs newspapers.

Digital Libraries basically store materials in electronic format to be then accessed and possibly manipulated by different kinds of users, very often connected via the Internet. Thus, it is fundamental to develop both the infrastructure

and the user interface to effectively access the information via the Net. The key technological issues are how to search and how to display desired selections from and across large collections [9]. Effective interfaces for text-based information systems are high priority for users of these systems. The interface must support a range of functions including query formulation, presentation of retrieved information, relevance feedback and browsing [4].

LAURIN (Libraries and Archives Collecting Newspaper Clippings Unified for their Integration into Networks) is an EU-funded Project[1] involving seventeen participants from several countries, including two software companies and a large group of libraries that want to make easily available and give wide visibility to the large cultural heritage they collect and catalog daily. The high number of users/libraries involved in this project gives the opportunity to spread culture and information to a wider public by means of the Internet. LAURIN has two major goals:

1. To set up a network of eleven European libraries and archives collecting newspapers and clippings, and make the network accessible via the Internet. This will guarantee that European citizens have easy access to a general index of all collected clippings and to the digitalized clippings themselves. Users will be supported by several tools and effective search mechanisms. The LAURIN network will be the pilot for an European network, connecting a large number of libraries and archives spread over Europe.

2. To provide a generic model to be used by individual libraries for scanning, digitalizing, storing, and indexing newspaper clippings, and making them accessible via the LAURIN network. A scanned image of each article as well as the full text will be prepared for storage and retrieval.

Concerning objective 1, since many users are ill equipped to translate their search requirements into precise queries, and they often prefer to use browsing as retrieval strategy, the LAURIN interface will offer, besides traditional keyword based search methods, also the possibility of browsing the clipping collection by argument, organizing the document space in a manner that is readily understood by users. Such activity will be supported by the use of an *integrated multilingual Thesaurus*, which plays a central role in the LAURIN system. The user will see a unified search space and therefore s/he can ignore the existence of different information sources, i.e., libraries. However, s/he will also be able to select a library on demand, based on the description of its characteristics, in order to restrict her/his attention to specific topics covered by a certain library only. Requests can be formulated in any of the languages supported by the system (currently English, French, German, Italian, Norwegian, Spanish, and Swedish) and the system will provide translations for the purpose of keyword and content based search. Finally, the users will have at their disposal a rich set of help mechanisms, directly accessible from everywhere during the interaction session.

To fulfill the above requirements, the LAURIN system is organized around a *central node*, which is connected via the Internet to a set of *local nodes*, one for each participating library. The digitalized clippings and their full-text (obtained via OCR) are stored in the local nodes, together with a local, possibly personalized, copy of the Thesaurus. The central node contains indexing data about all clippings stored in the local nodes, and a centralized copy of the multilingual Thesaurus with globally validated entries. A constant flow of information from the local nodes to the central node ensures that the latter is up to date. We refer to Section 4 for a more detailed description of the system architecture.

Concerning objective 2, the integrated Thesaurus system will support librarians in indexing and handling the clippings. This will facilitate both the librarians archiving activity and the access on a local level in a previously unknown fashion.

In this paper we present the design methodology, the overall architecture and the main functionalities of the LAURIN framework. The paper is organized as follows: Section 2 summarizes the state of the art in the field of digital clipping archives; Section 3 introduces the LAURIN design methodology and reports the user requirements; Section 4 describes the overall system architecture; Section 5 details the systems main functionalities; finally, Section 6 summarizes the lessons learned from the project development and describes future work.

2 Related Work

While in the literature there is a huge amount of research proposals and projects dealing with digital libraries in general (see e.g, [6, 15, 12, 5, 13, 11]) only few concentrate explicitly on digital libraries of clippings. The Historical Newspaper Digital Library project [1] shares with LAURIN the objective of digitizing and doing OCR on (historical) newspapers. However, the focus is not, as in LAURIN, on providing distributed access to the collections, but rather on developing rich meta-models to capture the complex structures of newspapers.

Among the distinguishing features of digital clipping archives we have that users of these archives are typically interested in articles in the original form in which they appeared in the newspaper. As a consequence, not only the full-text of the clippings and the images possibly associated with the text must be stored in the archive, but also a picture

[1] Telematics Program, Libraries Project LB-5629/A, `http://laurin.uibk.ac.at/`

of the scanned version of the original newspaper page (actually of the rearranged columns of single articles). Since the stored picture must have a resolution which is sufficiently high for obtaining a good quality printout of the article, this requirement poses particular challenges with respect to data storage. Additionally, the information associated with clippings may differ from the usual one stored in digital libraries. For example, the author information, which is mandatory in traditional archives may be irrelevant or even missing for clippings. Moreover, particular attention must be devoted to the copyright issue, which is quite different for newspapers than for books or journals, and differs from country to country. LAURIN will specifically address the copyright problem, although we do not deal with it in the present paper.

From this point of view there is no related work to directly compare LAURIN with. Thus, a deep analysis was carried out (a complete report is in [2]) involving existing general-purpose digital libraries and newspaper archives (e.g., Library of Congress[2], NY Times Archive[3], National Archives and Record Administration (NARA)[4]) as well as related projects from both research and industry (a significant subset of the projects we reviewed may be also found in [14]).

As we said before, assuring an easy and effective access to the stored clippings is the key-point of the LAURIN project. Since it is almost always true that the only part of a system the user sees is the interface, we concentrated the above analysis mainly on the different interfaces and interaction styles offered to the user. Unfortunately, user interfaces encountered in the digital libraries, archives, and electronic journal editions that have been considered, almost look the same. They allow the user to access catalogs by submitting queries through HTML fill-in forms, and in most cases offer the choice between "simple" and "advanced" query modes. The "browsing by argument" function is almost always offered, and links to other library catalogs are often provided. The bibliographic search seems to be usually considered a quite easy task, which does not deserve very sophisticated user interfaces.

During the last years, digital library systems have not made many efforts to solve the user-interaction problems. Only recently, new projects (e.g. University of Stanford[5] and University of Michigan[6] DL Projects) are developing a more complex model of information-seeking tasks. Display of information, visualization of, and navigation through large information collections, as well as linkages to information manipulation/analysis tools can be identified as key areas for research. A significant challenge for digital libraries is to allow the users to grasp in an effective and friendly way, the huge, ever growing, and diverse information on the Internet.

The necessity for a more comprehensive understanding of user needs, objectives, and behaviour in employing digital library systems is stressed repeatedly in new projects as the basis for designing effective systems. In particular, recent proposals mainly deal with the following aspects of user interaction:

- multi-language access to digital libraries and archives;
- integration of many different services, where information search is just a subpart of a more complex task;
- easy refining of results and revisiting of search process.

For example, the expansion and refinement of queries based on lexical relationships between documents, which are automatically extracted from the document collection, is addressed in [3]. A prototype implementation of a general user interface paradigm which is capable of modelling iterative query refinement is described in [7].

Another key issue addressed by the LAURIN project is the distributed nature of the collection of clippings. A distributed query system for preexisting library catalogs and structured databases (storing bibliographic data), based on an ad-hoc query language, has been developed in the HARP project [8].

3 The LAURIN User-Centered Approach

In this section we describe the characterizing features of the LAURIN project deriving from the central role of the user. Specifically, we introduce the design methodology, the taxonomy of the system classes of users, and the user requirements with respect to (1) the user interface, (2) the available indexing mechanisms, and (3) the searching methods.

3.1 User-Centered Design Methodology

Aiming at producing a highly interactive system, the LAURIN project is being carried on by following a rigorous "user-centered" design methodology [10], so that the envisioned solutions are really driven by the user needs and requirements. Note that this kind of approach is particularly appropriate for LAURIN, given the high number of libraries involved in the project, playing the double role of end-users and test sites.

The methodology comprises the following basic activities: (1) understanding and specifying the context of

[2] http://lcweb.loc.gov/
[3] http://www.nytimes.com/
[4] http://www.nara.gov/
[5] http://www-diglib.stanford.edu/diglib/
[6] http://http2.sils.umich.edu/UMDL/

use; (2) specifying the user and organizational requirements; (3) producing design solutions; (4) evaluating design against requirements.

According to such activities, the first step of the design team was to produce a questionnaire to be filled by the LAURIN users. Generally speaking, the questionnaire is one of the key techniques adopted in user-centered projects. For LAURIN, the analysis of the users' questionnaires was a very hard task, given the many differences existing between the different libraries. For instance, all participating archives use their national languages for indexing and some use vernacular names for geographic terms. A main problem turned out to be the handling of transliterations from non-Roman alphabets, since most of the libraries have no common rules, and sometimes the handling depends on the origin of the source. Therefore, in LAURIN the analysis of the questionnaires has been integrated with direct interviews and expert analysis, in order to fully characterize the context of use as well as the user and organizational requirements.

With respect to the second step, it is worth noting that even if the introduction of LAURIN had a substantial impact on the workflow of basically all users, they were ready to do extra work and encounter initial difficulties in light of the envisioned benefits.

Then, a first draft of system functionalities, accompanied by preliminary mock-ups of the user interface, has been produced and administered to the users in order to get their feedback. In LAURIN we got very interesting suggestions starting from the very first prototypes. For instance, when the librarians were asked to list requirements for the query interface in the analysis phase, they basically concentrated on query mechanisms based on keywords, and/or available information on authors, date, newspaper, etc., and/or thesaurus-based search. The various access modalities were differentiated based on the user's skill (casual, expert, etc.). However, while evaluating the first paper-based prototype, they asked for other access modalities, such as "access by collection", meaning that there is more emphasis on the single archive. Also, they required the possibility of browsing and making a "virtual tour" through the selected archive, although browsing was not considered a good access modality in the analysis phase. As a consequence of such a feedback, changes have been made to the system design, producing a new set of solutions, to be again discussed with the users, coherently with the iterative nature of the user-centered design.

3.2 User Classes

A main distinction can be operated among users of the system, classifying them in two broad categories, namely external and internal users.

External users are users who access the system, independently from the location of the nodes, to submit a query and, hopefully, get an answer. They can operate directly through an application or indirectly through the library staff. There are various categories of external users, as the questionnaires have demonstrated: casual users, student users, and expert users. A *casual user* is a user who is not very skilled in searches and is not very accustomed to libraries, so s/he seldom needs to express a precise query concerning specific clippings, but s/he may need a fast search on broad themes. A *student user* is a user who needs a deeper search in the universe of clippings for specific and particular reasons, so s/he needs an easy to use and highly selective interface. This kind of user often browses a set of clippings and searches by restricted categories, following a not-so-broad subject. S/he typically needs to incrementally refine her/his query. An *expert user* is normally a user accustomed to a library, who can spend time to learn the best access paths to information and needs very selective queries.

Internal users are part of the library staff who operate on the system to accomplish the following tasks: (a) to ask queries (in this case they embody the role of external user); (b) to input clippings; and (c) to administer the system. The main task of the internal user is clipping input, that is, scanning, OCR-ring and cataloging of clippings. This activity is performed only on local nodes (see Section 4.2). Some internal users, playing the role of system administrators, are also allowed, through a password protected interface, to deal with the inner part of the Central Node (see Section 4.1). In particular, the system provides an interface to periodically validate new Thesaurus entries, coming from the local node clipping classification.

3.3 System Requirements

In this section we describe the user requirements with respect to the interface and the system functionalities, as they emerged from the questionnaires filled by the users.

3.3.1 User Interface Requirements

The list of requirements is divided in two parts: those concerning the query expression activity and those concerning the result display activity. As for the query expression, the main requirements of the interface are as follows:

- simple use;
- incremental query refinement;
- partitioning and visual navigation in data space;
- multi access paths;
- window and icon based interaction.

The query is supposed to be formulated as follows: give me all clippings about *something*. The "something" part

must be defined in a way that produces valid results (low noise in results), which can be incrementally refined, and must be simple to define by an average user (not extremely expert on the clipping collection or "casual").

A multilevel visual interface may satisfy all above needs. Every level must be more specific than the previous one and offer a higher number of options so as to guarantee a finer selection. As for the result display, the requirements (see below) are also satisfiable by a reasonably sophisticated visual interface:

- simple use;
- data clustering and partitioning;
- visual navigation in the data space;
- multi views: different views of the same data considering diverse data characteristics;
- window and icon based interaction.

3.3.2 Indexing Requirements

These requirements come from internal users, i.e., librarians, who have to index the clippings to be stored as part of their daily activity.

The following indexing mechanisms have been devised by the librarians:

- *Prime Index*: is the basic information on clipping/article that otherwise will be lost during the clipping process (name of newspaper, page, rubric, date, ...);
- *Bibliographic Index*: basic bibliographic information on clipping/article (author, title, subtitle, text type of an article).
- *Keyword Index*: is an association of known terms from the Thesaurus with clipping/article which is automatically generated from the article full-text;
- *Content Index*: is an association of clipping/article with normalized terms from the Thesaurus resulting from a human content analysis of the clipping/article (that can use as a starting point the automatically generated terms of the keyword index);
- *Free Index*: is an association of clipping/article with subject headings that are not part of the Thesaurus resulting also from the human content analysis. The terms in the free index are candidates to become new entries in the Thesaurus.
- *Full-text Index*: is a computer based retrieving of all normalized terms in the clipping/article (including terms that are not in the Thesaurus), generated and maintained by a full-text information retrieval engine.

The above indices are used in developing different clipping classifications. The clipping main classification refers to the syntactical basic information associated with a clipping, i.e., newspaper name, language, author, newspaper sections, etc. It corresponds to the Prime Index plus the Bibliographic Index and can be a valid help to make a rough refinement on the search space. The keyword-based classification corresponds to the classical one that is available in every free text searcher. Here this kind of classification may be used in two search schemata: in the classical way, i.e., to find out all clippings containing the selected keywords in the full text (or a combination of them through AND and OR connectives); and searching in the space of the clipping prime characteristics, looking, for example, for all the clippings containing the keyword x in the title AND the keyword y in the name of the newspaper. A semantic context may be suitably adopted to limit the meaning of every keyword used in a search; moreover the scope of the search can be enlarged searching the same keyword in different languages or searching for synonyms and/or more abstract keyword. Obviously, in this context the role of a Thesaurus comprising several languages is critical.

A free index based classification can be seen as a semantic context classification arbitrarily imposed by someone to a set of clippings. The classical example is the set of directories built by a journalist. Every directory is classified under a subject and is filled with clippings related to the subject. In this case, the collection follows a completely personal classification, which the system could enforce and preserve, for historical and practical reasons.

3.3.3 Searching Requirements

A set of search fields has been devised analyzing both the user requirements and the partner library catalogs. The basic search fields include title of the article, author, various information on the newspaper, multilingual keywords, involved people, place, institutions, historical period, etc. The domain of these fields could be defined in advance (e.g., text type or language) or incrementally through the insertion of new clippings. The user is helped to choose among a set of alternatives, if the number of elements in a domain is not too high. Concerning domains that are too crowded, like, e.g., involved people, it should be possible to sort the available values in alphabetic order allowing the user to search/browse among them entering the first digits of the searched item and/or browsing a first subset of items starting with the same digits. Moreover, the system should provide the user with a Thesaurus browser, allowing for hierarchical navigation among terms. As an example, the user will be able to select the location "Rome", either using the alphabetical order of "Rome" within a subset of the geographical Thesaurus data, or following the path "Earth $\rightarrow$ Europe $\rightarrow$ Italy $\rightarrow$ Rome". Every domain must be multilingual, that is every element must be translated in the corresponding word in every language involved in the project.

Not every field has to be used during the classification

process of clippings in every library. Every library can tailor the list to its best needs and preferences. Also the user may decide the completeness level of a query offered by the interface, selecting the voices s/he really needs, adopting a sort of user profile. The system itself can propose various standard sets of fields to ask queries, like simple, moderately refined, refined, advanced refined. The user should have the capability of refining previous queries and save them for later re-use. Finally, the questionnaire answers pointed out a high level of interaction when external users issue queries to the library staff, sometimes followed by browsing through the library index categories. Thus, browsing is a search mechanism which has definitely to be supported by the interface. A mixed querying/browsing mode should also be supported, where the user asks a coarse grained query and then browses the result to refine it.

Additionally, the analysis of the questionnaires has shown the existence of two main classes of user search tasks:

- *Library oriented tasks*, where the following different subtasks have been identified:
  - distributed search on all libraries (i.e., local nodes);
  - restricted search on a user defined subset of libraries;
  - user directed search on a specific local node, where either the user knows in advance the library s/he wants, or the user chooses one or more libraries based on metadata exploration.
- *Clipping oriented tasks*, with the following subtasks:
  - the user needs a specific article s/he knows in advance;
  - the user needs a set of articles based on some criteria (e.g., topic, period, newspaper, language).

4 System Architecture

The overall LAURIN architecture, depicted in Figure 1, reflects the user requirements reported in the previous section. It foresees a set of nodes connected through the Internet: one node for any participant library plus a *central node* collecting data from local nodes and providing the end user with a uniform query environment. The central node hosts a relational database in which summary data coming from the local nodes are stored (i.e., clipping title, date, newspaper, author, etc.). Local nodes are in charge of clipping scanning and indexing; moreover they store all the information about acquired clippings: summary data, full clipping text, and clipping images. LAURIN clippings are strictly related with the LAURIN *Thesaurus* that is stored in the central node and replicated in the local nodes. There is a constant flow of information from the local nodes towards the central node, updating the central database with new clippings and new thesaurus entries. Periodically, the thesaurus administrators validate the proposed thesaurus entries and the central node propagates such validations towards the local nodes. When a user formulates a query, the central node tries to obtain the result using the central data, involving the local nodes only when specific full text based queries are issued or the clipping images are requested. The central node is in charge of collecting the answers coming from local nodes and presenting the final result to the user. The central node contains a Z39.50 [16] interface as well, which allows for acting as a Z39.50 server, exporting all LAURIN summary data. Depending on local hardware/strategical issues, each local node may be directly queried by the end users through a Web interface and/or a Z39.50 interface.

We now illustrate in more detail the various components of the architecture.

4.1 Central Node

The main components of the central node are:

- *Web Server*: a standard HTML server providing remote users equipped with an HTML browser a uniform access to the LAURIN clipping archives. Users will be able to choose between a simple applet-based interaction (for very casual users) or, if the user wants to fully exploit the LAURIN system, to download a Java application providing more complex query functionalities.
- *Database Manager*: this component provides, through the Oracle DBMS, the storage and retrieval of all summary clipping data plus Thesaurus data.
- *Z39.50 Gateway*: this component provides an access to the centralized clipping data through the Z39.50 standard.
- *Query Manager*: this component analyses a query coming from the end user via the Web Server and distributes the query across the involved local nodes.
- *Local Nodes Interface*: this component handles all communication among the central node and the local nodes, including:
  - update of the central node clipping database with data coming from local nodes;
  - update of the central node candidate Thesaurus entries with data coming from local nodes;
  - propagation of validated Thesaurus entries towards local nodes;
  - distributed query processing on one or several local nodes; in this case the data flow is bi-directional:

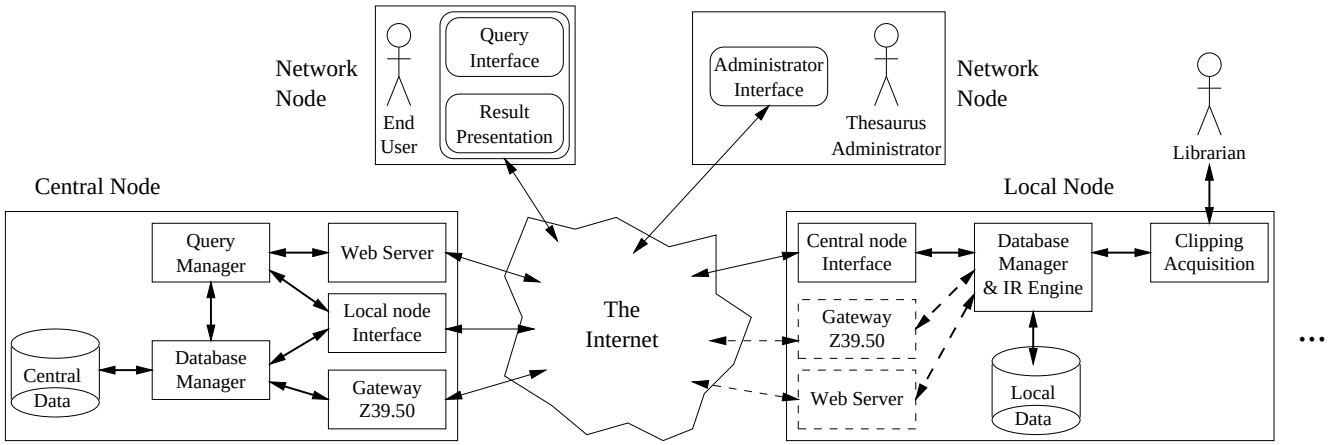

Figure 1. The overall LAURIN Architecture

the central node issues a query and the local nodes provide answers.

4.2 Local Node(s)

The main components of a local node are the following:

- *Central Node Interface*: this component handles all communication with the central node including:
  - sending new clippings to the central node;
  - sending new candidate Thesaurus entries to the central node;
  - accepting from the central node validated Thesaurus entries;
  - accepting a central node query based on the full text clipping content;
  - answering a central node query (i.e., sending the list of clipping IDs satisfying the query).
- *Web Server* (optional): a standard HTML server providing remote users using an HTML browser a (customized) access to the local node clipping archive.
- *Z39.50 Gateway* (optional): this component provides an access to the local clipping data through the Z39.50 standard.
- *Database Manager & Information Retrieval Engine*: this component, based on the Oracle DBMS and Oracle ConText Option Cartridge, allows for storing all clipping data (text, images, and summary data) plus the Thesaurus data. This module is able to answer in an efficient way text queries on clipping content.
- *Clipping Acquisition*: this module allows local librarians for scanning and indexing new clippings. The acquired data are stored locally and the central node interface provides updates to the central node.

Note that Figure 1 describes these components in a logical way: as an example, the Web Server and the Database Manager may be physically hosted by two different computers.

4.3 Data Workflow

We can single out four main workflows across the LAURIN architecture:

Clipping Acquisition: This is mainly a local node activity that involves the physical scanning of the clipping, the optical character recognition, the indexing of the clipping (according to the indexing criteria described in Section 3.3.2), and the clipping data memorization. The result of such an activity is a new set of clippings and a new set of candidate Thesaurus entries. This information is propagated towards the central node that updates its central database;

Clipping Database Maintenance: This activity involves a local node and the central node. With a timing that depends on the local node constraints (rate of clipping acquisition, urgency, speed of connection, etc.) the summary data of the new scanned clippings plus their connections with Thesaurus concepts is automatically sent through the network to the central node that updates its clipping database, keeping track of the clipping owner and making this new information available to the end user. A simple but effective general agreement about clipping IDs allows for avoiding collision of clipping keys. It is worth noting that this activity is an incremental one: it implies only adding clippings to the central node database that stores the related information as it is provided by the local nodes.

Thesaurus Database Maintenance: It may happen that, while indexing a clipping using the thesaurus concepts, a librarian is not able to find a thesaurus entry satisfying her/his needs. In this case the clipping acquisition module allows for associating the clipping with a *new* Thesaurus concept (*candidate concept*) that is sent to the central node together

with the clipping data. That implies that the central node handles a set of canonical Thesaurus entries plus, for each local node, a set of candidate concepts. Candidate concepts are available for query formulation as soon as they reach the central node but are *not* replicated on local nodes. Periodically, the Thesaurus administrators inspect and validates such concepts (see Section 5.1). Once a concept has been validated it is sent across the network to all local Thesauri. A full handshake protocol is adopted in this phase to avoid inconsistent clipping classification.

Distributed Query Processing: Once the user has formulated a query, the central node query manager answers it as follows. First of all, it analyses the query, splitting it into two parts, one related to the central node database and one related to the local nodes (i.e., the part of the query that refers to the full text of clippings). To solve the former it starts a query against the central database; to compute the latter it selects the local nodes that may possibly contribute to the query (e.g., to look for an Italian clipping at the Uppsala node in Sweden makes no sense) and then sends the query to the selected local nodes. Once each local node has returned a list of clipping IDs the query manager merges such lists with the answer it got by the central database, presenting the final result to the end-user.

5 System Main Functionalities

In this section we describe the main system functionalities of the LAURIN system, referring to the user types and their corresponding tasks as highlighted from the user requirement analysis. Note that, in order to better illustrate the activities related with the Thesaurus maintenance, inside the class of internal users, two subclasses of "Thesaurus Administrators" (either central or local) have been considered explicitly. We also briefly present the schema of the internal LAURIN database, supporting all users activities.

5.1 User Activities

External users interact with the central node and the system provides (on demand) a description of the LAURIN consortium and of the involved local nodes, allowing a direct connection to local nodes hosting a Web query interface. If a user wants to ask a query across two or more local nodes (all nodes as an extreme case) s/he interacts only with the central node that acts as a broker with respect to the local nodes. Also, an external user connected to the central node can browse the central Thesaurus to search for associated clippings[7]. Using several kinds of interfaces the user is allowed to formulate a multilingual query in which the Thesaurus plays three different roles:

1. it is a guide to understand the *classification* of the clippings stored in the LAURIN distributed database;

2. if the user has requested a multilingual search it translates the involved terms;

3. if requested by the user, it can be used to modify the scope of a query (e.g., finding not only the clippings containing the word X but also the clippings containing a synonym of X or a more specific term for X).

Summarizing, when keywords are used in a query, the Thesaurus is accessed to expand the set of keywords according to the user specified criteria (more general terms, related terms, terms in different languages, etc.). For keyword expansion the Thesaurus of the node to which the user is connected (either central or local) is used. The identifiers of clippings associated with the expanded set of keywords can then be retrieved and presented to the user.

When the query has been processed, the user can interactively refine the result. When s/he has reached her/his goal, s/he can ask the system for a summary of the results, containing all the necessary information needed to get the clippings (involved nodes, cost, etc.).

Internal users perform their activities related with indexing and storing clippings only on local nodes through an ad-hoc interface to a sophisticated OCR system. Whenever they want to make a query they are considered as expert external users and can use the corresponding interface.

Local node Thesaurus Administrators are special internal users, whose main goal is to administrate the local node Thesaurus. They typically update the local node Thesaurus with information associated with new clippings: The Thesaurus is queried and/or browsed to find relevant entries that can be associated with a clipping. Whenever an entry that is already in the Thesaurus needs to be associated with a clipping the association is stored in the local node database and transmitted to the central node together with the clipping data. Whenever an entry that is not in the Thesaurus needs to be associated with a clipping, a *candidate entry* is generated, inserted in the local Thesaurus, and associated with the clipping. At least a preferred name in the local language and an English name must be associated with the candidate entry. The system assigns a global unique identifier to the candidate entry. The candidate entries are transmitted to the central node for validation and also kept in the local node (together with their association with clippings) until validation is performed.

Central node Thesaurus Administrators have two main tasks, which they accomplish through a password protected Web interface[8]:

[7] An external user connected to a local node browses directly the local Thesaurus.

[8] Hence a Central node Thesaurus Administrator may physically reside in a different location than the central node itself, e.g., in one of the local node libraries.

- *Build, refine, and modify the Thesaurus*. This is an off-line activity, that alters the Thesaurus content, independently from the activity of local nodes (e.g., correcting errors, adding new terms for existing concepts, etc.). The updates resulting from such an activity are propagated towards local nodes.

- *Validate candidate entries*. This is part of the routine LAURIN job, and implies the analysis of the candidate entries coming from local nodes. For each candidate entry one of the following actions is performed: (1) the candidate entry is recognized as one already present in the Thesaurus and merged with it (choosing one preferred name). Then, the identifier of the candidate entry is removed and the clippings associated with the candidate entry are associated with the entry in the Thesaurus. The update is transmitted to the local node that had generated the candidate entry. (2) The candidate entry becomes a new entry in the Thesaurus. It is inserted in the global Thesaurus and relationships between the new entry and other entries in the Thesaurus are established. The update is transmitted to the local node that had generated the candidate entry plus all the other nodes.

  Note that the complete deletion of a candidate entry is not a viable option during validation, since the entry has already been associated with clippings in the local node that generated it. Moreover, it is assumed that the librarian, who is an expert in his field, had valid reasons for inserting the term in the Thesaurus, and therefore wants the term to be available.

5.2 LAURIN Distributed Database

Figure 2 depicts the overall schema of the LAURIN database, which is an integrated view of five information sources, namely the *clipping archive*, the *thesaurus*, the *periodical archive*, the *author archive*, and the *administrative archive*. In Figure 2 different grey-scales are used for illustrating concepts belonging to the various sources.

5.2.1 Clipping Archive

The main purpose of the clipping archive is to store all information and fields related to an article, namely the so called "structured" part of an article constituted by the attributes/metadata associated with it; the multimedia and textual objects associated with the clipping; the free-text part, managed by the retrieval engine; the multilingual implementation of the information about text type and object type. The clipping archive is distributed over the central node and all local nodes as follows: the central node contains a copy of all article data except the associated objects (i.e., clipping image and clipping text); each local node stores all the data about its own articles including the associated multimedia and textual objects; the association between Thesaurus entries and clippings is replicated in the central node, and each local node maintains the associations for the clippings stored in that node.

Access and updates to the clipping archive are performed according to the workflow described in the previous section. The archive is accessed by four typologies of users: (1) external users who query the LAURIN system through the central node interaction; (2) external users who query the LAURIN system directly from the local node (librarians who make queries for supervision activity belong to this category as well); (3) librarians at a local node who perform the indexing activity; (4) librarians at a local node who perform management activity such as adding/editing/deleting clippings.

5.2.2 Thesaurus

The database containing the Thesaurus is distributed over the central node and all local nodes as follows: the central node contains all validated thesaurus entries and the entries that are candidates for validation; each local node contains a copy of all validated thesaurus entries from the central node plus the local candidates for validation; the association between thesaurus entries and clippings is stored in the central node, and each local node replicates the associations for the clippings stored in that node.

5.2.3 Periodical Archive

As shown in Figure 2, the clipping archive is related to the periodical archive as well. The main issue of this database portion is to store all information and fields related to a periodical, namely the historical tracking of the newspaper/magazine, the place of publication (city and country), editors, frequency of publication, title, political attitude, supplements and merges of newspapers/magazines. All the above information exhibit a multilingual implementation.

5.2.4 Author Archive

Although authors could represent specific thesarus entries, for efficiency reasons they are stored separately. Instead, the country an author belongs to is a thesaurus entry.

5.2.5 Administrative Archive

The database containing the administrative information is maintained mainly for monitoring purposes. For instance, it can store statistical data about the activity of the users, such as how many articles have been downloaded, the number of connections, etc. In addition, it will be possible to integrate

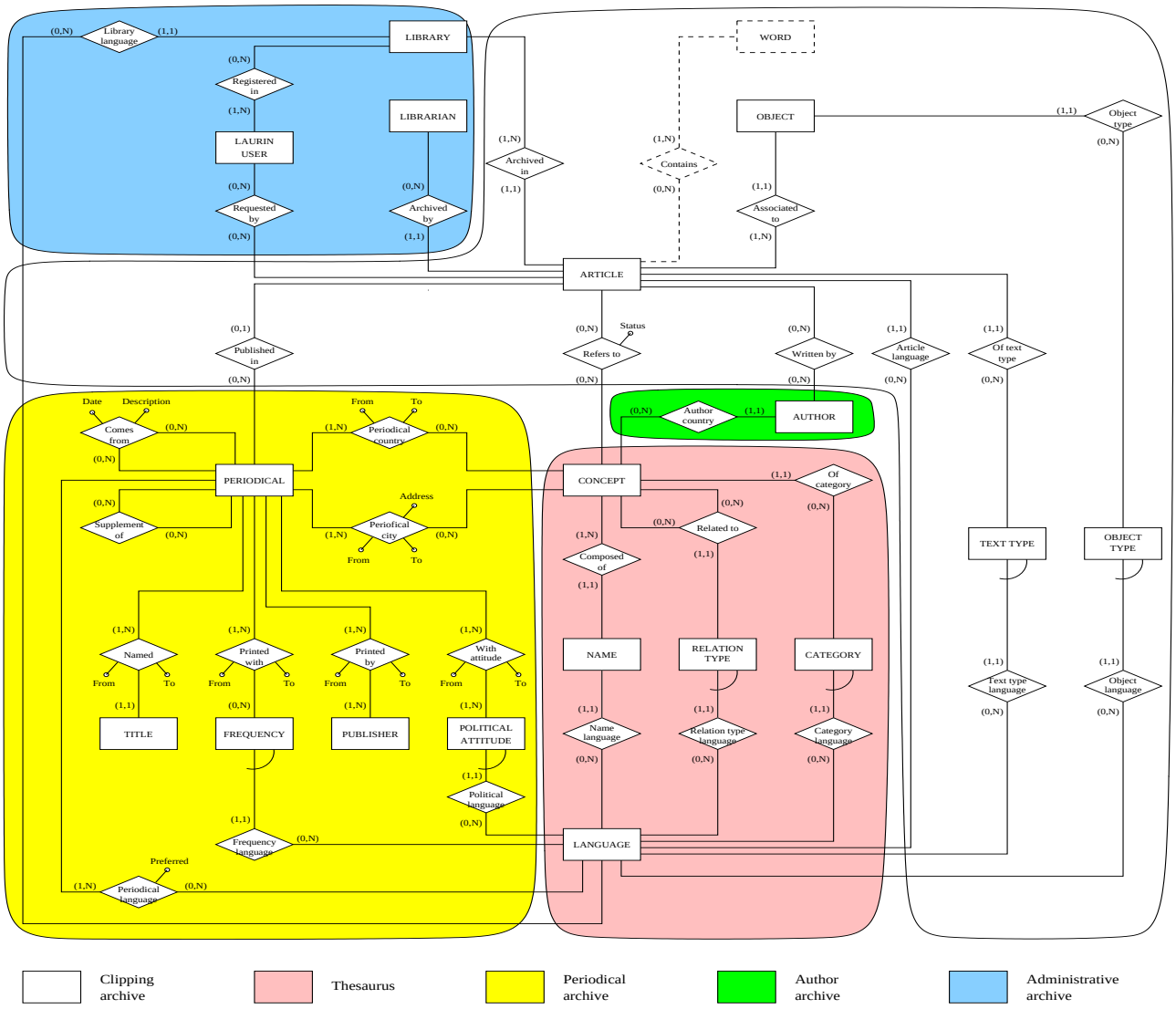

Figure 2. The LAURIN distributed database

specific data in order to calculate the cost of a transaction, useful for the payment functionality.

Note that the aspects related to copyright and payment are still under analysis and, even if several alternatives and possibilities have been discussed, for the moment the Consortium decided to start with a system that fits the requirements of all partners: a centrally administered payment-in-advance registration system that covers all LAURIN partners, with one global subscription or with several "library oriented" registrations. Referring to Figure 2, this is captured by the entity `LAURIN_USER` and by the relationships `Registered_in` and `Requested_by`. This approach allows for handling a simple access control policy by assigning each user a password. In this way the user access can be restricted to one or several libraries.

6 System Implementation

A first prototype of the LAURIN system has been implemented under the Windows NT operating system, using Jbuilder 2 with Java 1.1.7 and Oracle 8. All the modules foreseen for the Local Nodes have been implemented and installed at each participating library site. The Local Node system includes a first version of the Thesaurus as well, which contains a large set of geographical data extracted from the TGN thesaurus, and a set of names of famous people (artist, writers, etc.), together with a predefined set of relationships among concepts. In this very moment librarians are starting to clip and index articles, filling the thesaurus with new concepts.

A first version of the Central Node has been imple-

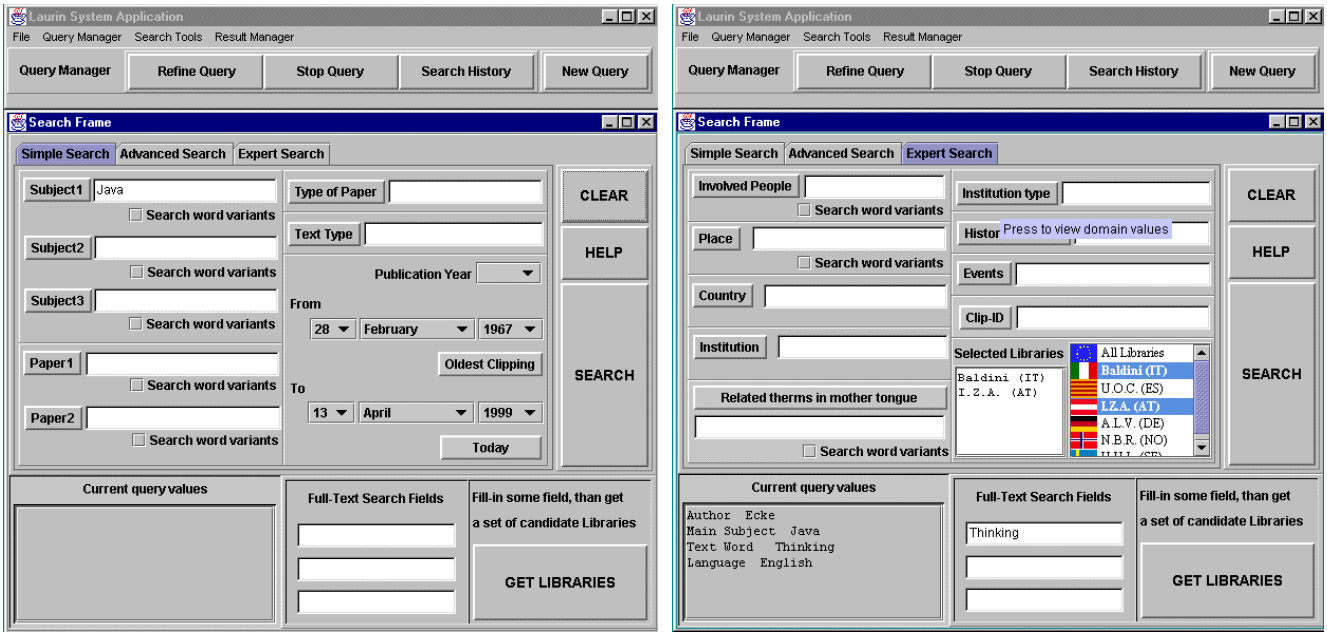

Figure 3. The Simple and Expert Search Interfaces

mented as well, providing the main query functionalities. In particular, the system allows for querying the whole LAURIN system through a form-based interface capturing the user requirements described in Section 3.3.3. The interface is based on a set of panes allowing for expressing queries characterized by increasing complexity. In Figure 3 the Simple and Expert Search panes are shown. The Simple Search pane allows for retrieving clippings through subjects, newspaper names, date, paper and clipping types. Moreover, the user is able to restrict the search space by launching the query on a subset of the LAURIN libraries. While interacting with a query pane, the user can change the interaction modality to ask more complex queries, such as looking for clippings written in a certain language and about people and places. A continuous feedback about the query is provided, allowing the user to have a full control on the whole process.

The communication protocols among the Central Node and the Local Nodes (i.e, clipping and Thesaurus database maintenance, see Section 4.3) have been implemented and their testing will start after the Local Nodes have been filled with a significant set of clippings and Thesaurus entries.

7 Lessons Learned and Future Work

In this paper we presented the key points and the overall architecture of the LAURIN project. LAURIN is the first project which specifically deals with clipping databases, with the final goal of building a large European network of libraries hosting clippings, which will allow users to easily find the articles (physically distributed over Europe) better matching their interests through a centralized Internet access.

One of the key aspects of the project is the presence in the Consortium of several libraries, which act as end-users. This has permitted to develop a truly user-centered design methodology and to test all design choices against real user requirements. We have discovered that the users of the LAURIN system are very conscious of their needs and have clear expectations (in particular, the librarians, with whom most of the tests were done). They have been exposed to several mock-ups of the interface and workflow simulations, and very often have reacted with constructive criticisms, which has led to several improvements of the system. For instance, the librarians rejected too complex interfaces, even if these offered more functionalities, since such extra functionalities were considered marginal with respect to the easy accomplishment of the overall tasks. In contrast, the various designers of the team first tried to make interfaces able to satisfy all functional requirements collected during the analysis phase. It is worth noting that users were ready to give up their own requirements in favor of a simpler interaction. This is further proof of the fact that majority of users prefer to have a simple way to get basic things than a complex way to get almost everything she or he would like in principle. Thus, in practice, over-featured interfaces do not work well. Cultural diversity (e.g., between Moi-Rana library and Catalonia Library) has not represented a real problem, instead it contributed to make LAURIN a richer project. Nevertheless, each country has strongly pushed to

preserve its identity. For instance, the Thesaurus interface had to be developed in such a way that each participating library could use its native language. A unified English-based interface was not accepted by the librarians.

From the technical point of view, first of all we discovered that the native Java RMI (Remote Method Invocation), in spite of its elegance, was not suitable to handle several connections across a (usually) not reliable network. Therefore we decided to use the socket approach that gave us full control on each single connection (retry, timeout, etc.). As a second consideration, our application is strongly based on the use of threads, which are extremely well supported by Java. However, to slightly mitigate our enthusiasm about Java threads, we have to note that a high usage of threads makes the debug activity really hard. Moreover we discovered that a modal dialog (e.g., one in which the system is waiting for an OK) stops all the running threads. However, we did not verify if this is version dependent, and if the same happens in more recent versions of Java.

It is worth noting that the most hard problem still to be solved within the LAURIN project is not technical but legal, namely the issue of copyright.

References

[1] R. B. Allen and J. Schalow. Metadata and data structures for the historical newspaper digital library. In *Proc. of the 8th Int. Conf. on Information and Knowledge Management (CIKM'99)*, pages 147–153, 1999.

[2] D. Calvanese, T. Catarci, V. Curci, E. Melis, A. Rastellini, and G. Santucci. The overall laurin architecture. Technical Report Deliverable Nr. D3.10.2, Laurin Project, Dipartimento di Informatica e Sistemistica and CM Sistemi S.p.A., 1999.

[3] J. W. Cooper and R. J. Byrd. Lexical navigation: Visually prompted query expansion and refinement. In *Proc. of the 2nd ACM Int. Conf. on Digital Libraries (DL'97)*, pages 237–246, 1997.

[4] W. B. Croft. What do people want from information retrieval? (The top 10 research issues for companies that use and sell IR systems). *D-Lib Magazine*, Nov. 1995.

[5] J. Frew, M. Freeston, R. B. Kemp, J. Simpson, T. Smith, A. Wells, and Q. Zheng. The Alexandria digital library testbed. *D-Lib Magazine*, July 1996.

[6] T. S. D. L. Group. The stanford digital library project. *Communications of the ACM*, 38:59–60, 1995.

[7] L. Kovács, A. Micsik, and B. Pataki. AQUA: Query visualization for the NCSTRL digital library. In *Proc. of the 4th ACM Conf. on Digital Libraries (DL'99)*, pages 230–231, 1999.

[8] E.-P. Lim and Y. Lu. HARP: A distributed query system for legacy public libraries and structured databases. *ACM Trans. on Information Systems*, 17(3):291–319, 1999.

[9] C. Lynch and H. Garcia-Molina. Interoperability, scaling, and the digital libraries research agenda: A report. In *Proc. of IITA Digital Libraries Workshop*, Aug. 1995.

[10] D. Norman and S. Draper. *User Centered System Design*. LEA, Hillsdale, N.J., 1986.

[11] J. Ober. The california digital library. *D-Lib Magazine*, Mar. 1999.

[12] V. Ogle and R. Wilensky. Testbed development for the Berkeley digital library project. *D-Lib Magazine*, July 1996.

[13] A. Peterson Bishop. Measuring access, use, and success in digital libraries. *The Journal of Electronic Publishing*, 4(2), 1998.

[14] B. Schatz and H. Chen, editors. *Digital Libraries: Technological Advances and Social Impacts*. Feb. 1999. Special Issue of IEEE Computer.

[15] N. A. Van House, M. H. Butler, V. Ogle, and L. Schiff. User-centered iterative design for digital libraries: The Cypress experience. *D-Lib Magazine*, Feb. 1996.

[16] Z39.50 Maintenance Agency. *Information Retrieval (Z39.50): Application Service Definition and Protocol Specification (ANSI/NISO Z39.50-1995)*, july 1995.