A Digital Library for a Virtual Organization

José Luis Borbinha[(*)]
IST – Technical Superior Institute (Lisbon Technical University)
INESC – Engineering Institute for Systems and Computers
Jose.Borbinha@inesc.pt

João Ferreira[(**)]
IST – Technical Superior Institute (Lisbon Technical University)
jcaf@bruxelas.inesc.pt

Joaquim Jorge
IST – Technical Superior Institute (Lisbon Technical University)
INESC – Engineering Institute for Systems and Computers
Joaquim.Jorge@inesc.pt

José Delgado[(*)]
IST – Technical Superior Institute (Lisbon Technical University)
INESC – Engineering Institute for Systems and Computers
Jose.Delgado@inesc.pt

Abstract

In this paper we discuss the digital library metaphor and introduce the concept of *networked digital library*, here defined as a library with the additional mission of stimulating, supporting, disseminating and recording the process of creation of information. The ArquiTec project is presented as a pragmatic implementation of that concept for an example of a virtual organization: the Portuguese academic and research community.

ArquiTec is a system with a distributed architecture based on the NCSTRL technology. Local nodes manage local repositories at universities and research institutes, but all collections are freely accessible from any node. Based on that infrastructure, the Portuguese National Library will manage a central repository of official persistent digital documents.

ArquiTec further integrates a filtering service, based on user profiles, a public document annotation service, and an ontology space based in the integration of multiple statistical and formal thesauri.

1. Introduction

Vannevar Bush had a vision, 50 years ago [1]:

"Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and to coin one at random, "memex" will do so. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. (…)".

Indeed, the dream of the digital library is not new, but only now we are starting to clearly define it and having access to the required technology in order to turn it into a reality. In fact, with the power of our computers and networks, we can convert our memory and knowledge into digital bits, store it at an affordable cost and, when required, make it available to anyone, anywhere on Earth.

This challenge has been addressed in a large number of initiatives, especially after the emergence of new paradigms brought forth by the success of the Internet and the World Wide Web. However, those paradigms have not always been well understood and applied, resulting in the abuse of models and metaphors. As Mark Ackerman wrote [2]:

"With the advent of mass-market networks and information utilities, there has been a concomitant arrival of new societal metaphors. We hear such terms as "virtual communities", "digital library", "collective memory", and "information highway" bandied about by newscasters, academics, and politicians, and while these metaphors shape our understanding of the new phenomena, they are hardly examined for their efficacy or truthfulness".

This perspective raises a challenge that we decided to accept. For that we defined the concept of *networked*

[(*)] Work partially supported by the JNICT contract PRAXIS/2/2.2/TIT/1667/95
[(**)] Work supported by the JNICT fellowship PRAXIS XXI/BD/59/68/95

digital library (NDL), and put it to test in a trial, the ArquiTec project.

ArquiTec is a joint effort undertaken by INESC, the Portuguese National Library and JNICT (the Portuguese agency for research funding), to develop a networked digital library for the Portuguese academic and research community.

We intend to use ArquiTec both as a technology demonstrator and a framework to develop, test and consolidate expertise in core fields related to digital libraries, with a special emphasis on our concept of networked digital library. In that sense, ArquiTec's added value goes beyond merely providing a useful demonstration system. It will also serve as a laboratory to further identify and study the different technical, social and institutional implications raised by the NDL concept.

This paper continues in three main sections. In the next section we will discuss and present our definition of the networked digital library concept, following with the description of the ArquiTec project as a pragmatic embodiment of that concept. Finally we briefly refer some historical and related work and resume the main open issues under study or that will claim our attention in the near future.

2. Defining the Networked Digital Library

Traditionally, libraries have assumed the mission to organize, preserve and provide access to information, while trying to cope with technological evolution, from the introduction of the papyrus in ancient times to the newer challenges of digital publishing, the Internet and the World Wide Web.

Traditional libraries are centralized and organized around *books*, official and "sacred" indivisible pieces of knowledge intended to be kept "forever". In this scenario, authors decide what to write and agree with editors when to edit and publish, whilst librarians decide whether or not to buy the final product.

More recently, increasing specialization brought us thematic journals, reports and conferences, from which a new object emerged: *the paper*. The paper is a formal entity, validated by the credibility of an editor or review committee. It is not intended to be valid forever, but to be discussed during a period of time, refined and, in the end, what survives is then distilled into books. It is difficult for traditional libraries to follow the trend of an ever increasing specialization on all fields of knowledge; so libraries themselves become specialized, with a mission to serve specific communities. Since these communities are well identified, it is now possible (at least in theory) to anticipate their needs and to provide customized services, such as notify users that new journal issues have arrived, advertise of new publications, etc. In this scenario, the organization of the library changes itself. It approaches the communities by distributing competencies and functionality to specialized branches, sometimes physically located outside the central library building.

The scenario changes again when computers arrive. Computer networks allow communities to intensify their

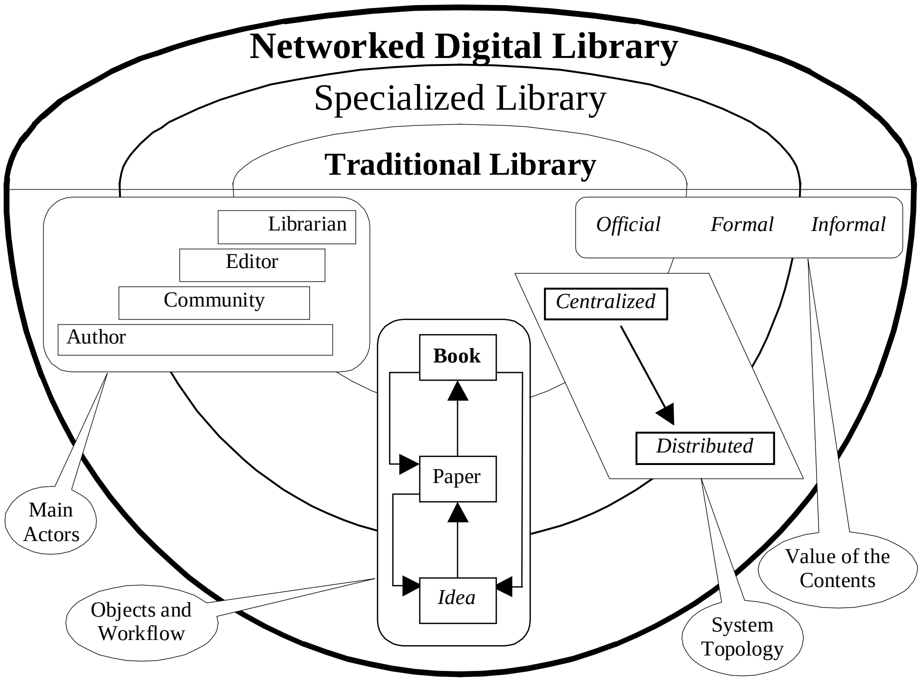

Figure 1: An evolution of the library paradigms (it is assumed that each new paradigm upgrades the former one, bringing new functions and responsibilities).

interactions. With electronic mail, desktop publishing and WWW, everyone becomes a potential publisher. Speed of interactions increases, and a new focus emerges: *the idea*. To produce fast results, ideas are presented in informal pre-prints and discussed in informal workshops. Ideas that succeed in this process result in formal papers, which are then published in journals and promoted in formal conferences. Using electronic mail and WWW, it is now easier for libraries to reach communities and to provide new services. For the same reason, it is now easier for users to interact with libraries, not only to access OPAC (Online Public Access Catalog) services but, in an extreme scenario, also to contribute with new kinds of *meta-knowledge* that can notably enrich library contents. Examples of such contributions can be the tuning and completing thesauri and document classification (allowing dynamic and collaborative classification), annotations and comments to stored documents, etc.

The perspective presented above led us to the vision of the evolution of the library paradigm as shown in Figure 1. From that vision we worked out a definition for the networked digital library that, from our point of view, comprises the most relevant concepts discussed so far:

A networked digital library is defined not only as a repository of information, with the traditional missions of preserving, organizing and providing access to its contents, but also as a system to disseminate that information and actively stimulate, support and record the process of its creation.

In the networked digital library, authors and their related communities have an important role to play, since they are the main actors in a new task: the evolution from informal contents (the idea) to formal contents (papers, books). At the same time, the distribution in the topology is enforced by the virtualization of the system components. Libraries become more independent of spatial constraints, while digital content makes it less dependent of time limitations (for example, it is no longer necessary to wait for the return of lent copies of a document, since producing new copies is nearly instantaneous and costs virtually zero).

3. ArquiTec

The ArquiTec project started in the beginning of 1997, and a limited prototype has been developed. The first phase will end with a working prototype, scheduled for public release in the first quarter of 1998, followed by a six-month trial period. After that, a second phase is expected, involving new institutional and industrial partners in order to address the open issues.

In what follows, we will detail the project, focusing our description in five steps: key concepts and requirements, system architecture and the three main conceptual entities of the system, comprising the documents, the thesaurus and the user directory.

3.1. ArquiTec’s Key Concepts and Requirements

In compliance with the NDL concept, ArquiTec is accessible over the Internet, through a WWW interface. It provides access to different kinds of technical documents (such as papers, reports, theses, dissertations, etc.), in different fields of knowledge, while special services will also be provided to the community (such as a notification service).

Following the NDL concept, the system should provide support for a three-step workflow in the production and dissemination of information and knowledge, comprising:

- *Informal documents,* such as private collections held and managed outside the library but registered and periodically checked, classified and made available trough its facilities- (usually WWW sites with “home-pages” of people, groups and projects; public repositories with the kind of publications usually known as “*grey*” literature such as position papers, pre-prints; etc.).
- *Formal documents,* such as distributed specialized collections managed by local library structures and composed of refereed and validated material (such as papers presented in conferences or published in conventional journals).
- *Official documents*, a centralized collection, fully managed by a central library structure and comprising material such as theses, dissertations, reports, electronic books, etc.

In order to follow these requirements, the system has been developed around three main entities, as shown in Figure 2: *documents*, *users* and *subjects* (the latter represented in the figure by the concept space).

Informal and formal documents exist in local repositories, managed by distributed servers. To address the problem of long term preservation, the Portuguese National Library will build and maintain an official archive with a copy of selected formal documents as part

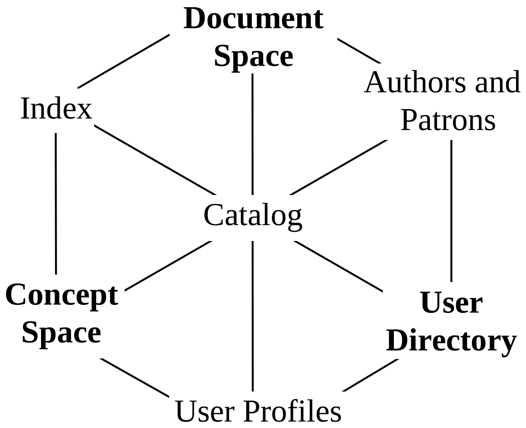

Figure 2: Main entities in ArquiTec.

of its mission as a deposit library.

ArquiTec users can be authors, readers, or both. Users are managed in a global directory, where their identity, contacts, affiliations, etc., and a special profile are registered. Anonymous access is possible for search, browse or even retrieval, but in any case the users are always suggested to register their identity for profile management.

The third main entity in ArquiTec is the concept space, or ontology, based in the integration of multiple statistical and formal thesauri. In that sense our thesauri perform functions well beyond their usual roles as auxiliary tools for classification and search. Like the user directory, the thesauri play also an important role in the library contents, as explained below.

Documents, users and concept are interactive and dynamic entities, which means that they can change over time. For example, documents can have new releases or attachments submitted, users can become interested in new subjects, new subjects can be included in the thesaurus, new relationships can be established between existing subjects, etc. Indexes, user profiles and possible relations between documents and users (authors or just readers) associate these entities among themselves.

An *index* relates documents to thesauri entries. It is composed by three lists: a controlled list of terms (keywords and expressions) to index the documents (in fact, a non official authorities list), a controlled list of stop-words (to be ignored in the indexing) and a list of unknown words (that users and management can browse and, once understood, classify either as keywords or stop-words).

Users are identified in ArquiTec by their interests and contributions, which relate to subjects in the thesaurus (likewise for documents). In ArquiTec users are viewed not only as authors and patrons but also as important sources of information, with their profiles becoming part of the contents. Profiles serve also to provide special services to the users, such as automated notification and ranking search results.

Conceptually, a *catalog* makes it possible to explore, in an integrated perspective, the above six concepts (comprising the three main entities and the relationships between them). In that sense it becomes possible and has an equivalent meaning, for example, to search for documents or for users related to a specific subject. In an integrated perspective, it is also possible to search for both users and documents related to specific subjects, and so conceptually "sharing common interests".

A final requirement for ArquiTec is to provide persistent names for formal documents stored in the archival space. The problem of naming objects in a digital library was generically addressed in the "Kahn/Wilensky Report", from which the concept of *handle* as an URN (Uniform Resource Name) emerged [3]. A simplified version of that concept was implemented by OCLC in the *PURL* (Persistent URL) service, based on a highly reliable HTTP server [4]. A PURL is a normal URL, with a logical meaning that, when used, implies an access to the PURL server that acts as a HTTP proxy and automatically translates the logical name to the real URL of the object referred to.

A PURL service is provided in ArquiTec, which automatically generates and manages a PURL for each document archived in the central server.

3.2. System Architecture

A series of paradigmatic initiatives have been drawn on the expansion of the Internet to provide online collections of scholar and scientific documents, namely in the Computer Science area. One of the most representative efforts has been NCSTRL (Networked Computer Science Technical Reports Library), a network of servers providing three kinds of services: repository, indexing and user interface [5]. Currently NCSTRL is a worldwide

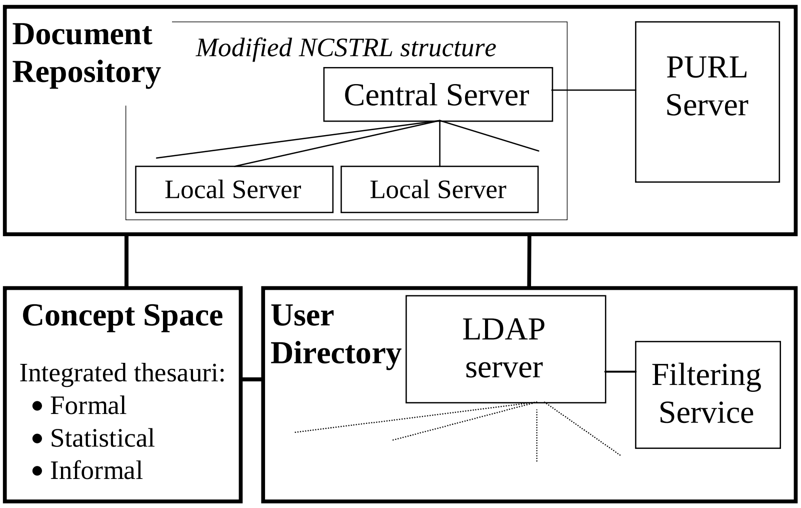

Figure 3: ArquiTec's main blocks.

service, with repositories installed in over 60 universities and research centers across the world.

We were impressed by the capabilities of the NCSTRL technology as a potential foundation for new work, especially by its open architecture model, easy installation and maintenance and its ability to handle documents in several formats. Therefore we decided to use it as the main core technology for our project.

ArquiTec is a distributed system, with each participating institution managing its own formal and informal collections (and preferably a repository server, although that is not mandatory since one local server can host more than one institutional collection). The global blocks are summarized in Figure 3, which bears a close relationship to Figure 2.

The system core is based on a modified and extended version of NCSTRL, using version 4.0 of the DIENST protocol and server [6]. This protocol, built upon HTTP, was developed by researchers at Xerox Corporation and Cornell University.

Users are managed in ArquiTec in a global X.500 directory based in a structure of LDAP (Lightweight Directory Access Protocol) servers [7], a simplified TCP/IP version of the original protocol DAP (Directory Access Protocol), defined in the ISO X.500 standard [8].

The thesauri are multilingual, and they are implemented following the ISO-5964 standard [9]. The system imports (and exports) formal thesauri defined in MCF, a simple format for meta-content representation [10].

In figure 4 we show ArquiTec local servers main blocks. The original DIENST protocol and the NCSTRL user interface were modified, in order to support multi-lingual access. This extension is fully compatible with the original protocol, making it possible the coexistence, in the same system, of original and extended servers. It consists only in a new field, at the end of each command, with the code of the language. This field is parsed in the DIENST server and processed after that. In the absence of the language code, the default language is used (Portuguese).

We modified also the submission of documents, which can now be done remotely, and the original indexing tool, which we replaced by *Glimpse* [11]. With this tool we can index the metadata structures and, when possible, documents' full text (filters have been adapted and developed to convert other formats, such as PostScript, to plain ASCII, for indexing, and GIF, for visualization).

The ArquiTec central server has two more modules than local servers do.

After documents are selected by National Library staff, a new gather module at the central server copies the chosen documents from local servers using HTTP and then simulates a local submission. Another new component is the notification service, a "batch" service activated by relevant events in repositories (such as the submission of a new document or annotation).

The submission module was modified in the central server in order to allow managing the archive, which is seen as a collection of the central server. We also changed the way the central index is managed. In NCSTRL the central server periodically gathers local indexes from their respective servers. In ArquiTec the central server gathers metadata files from their local servers (and not the indexes) and creates the indexes locally. This is required by Glimpse, but as a positive consequence since it is possible to provide a global fault-tolerant central catalog independently of the local servers.

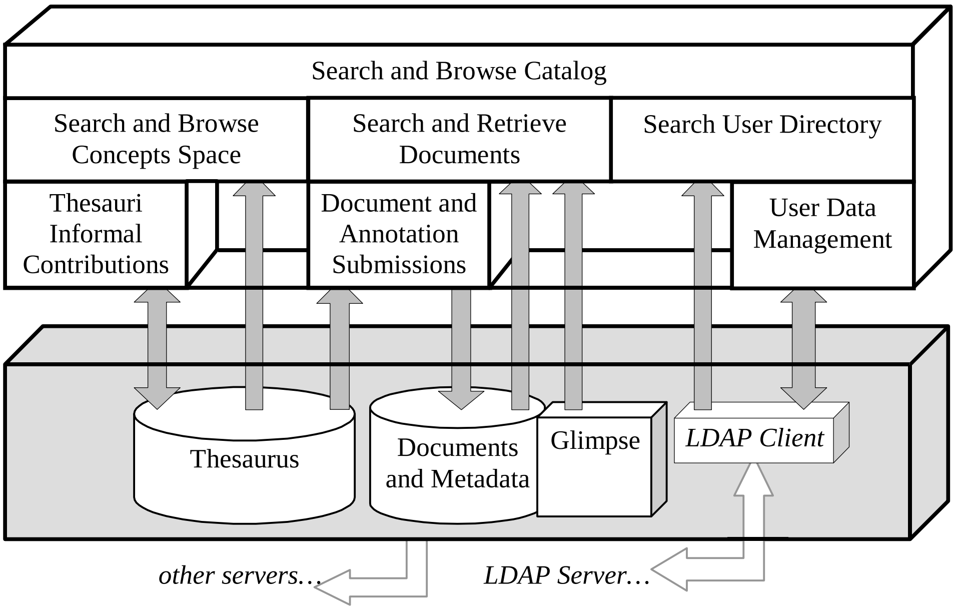
Search and Browse Catalog
Search and Browse Concepts Space
Search and Retrieve Documents
Search User Directory
Thesauri Informal Contributions
Document and Annotation Submissions
User Data Management
Thesaurus
Documents and Metadata
Glimpse
LDAP Client
other servers…
LDAP Server…

Figure 4. Structure of a local ArquiTec server.

3.3. User Directory

As it was already mentioned, users are managed in ArquiTec in a global X.500 based user directory, accessible by LDAP.

A good directory service should provide fast answers when searching information in large amounts of data, given that the information in a directory is generally read much often than it is written.

A X.500 service is based on entries. Entries are collections of attributes as defined by RFC 1779 [12]. Each entry has a type (or class), typically defined by one or more mnemonic strings, and can have one or more values. The information is supposed to be distributed and structured in a tree. At the top we have entries representing countries, below that we have entries representing national organizations. At the lowest level there are entries representing any other desired class of objects (such as people, computers, printers, etc.).

Table 1: User entry in the user directory

Generic Attributes	**Profile Attributes**
Name	**Explicit Fields:**
Institution	Relevant subjects
User ID	Non-relevant subjects
Password	**Implicit Fields:**
Password tip	Ids of documents is author
Email address	Subjects of documents is author
Telephone	Ids of submitted annotations
Fax	Ids of documents read
WWW home page	…
…	

In our system we will represent ArquiTec users with value fields such as the presented in Table 1, in a structure such as the one presented in Figure 5.

Users access ArquiTec in one of two modes: anonymous or identified. Identified users have profiles composed of explicitly provided data (their explicit interests) and data implicitly extracted from the history of their interactions with the system (such as submitted and retrieved documents, for example).

User profiles serve three main purposes:

- *Searching*: profiles are used to rank search results, for example highlighting documents that best match user's interests (but ranking will never hide or restrict the access to other documents).
- *Filtering*: profiles are used to provide an information dissemination service, supported by electronic mail, through which users can receive automatic notification of new events.
- *Collaboration: interactive services for document annotation (which can be reflected in the catalog) and for thesauri tuning are also provided.*

Concerning the user directory, we intend to distribute and replicate it in the future by several servers within the National academic network, to provide flexibility and **attributes** fault tolerance. However, the final architecture will be defined only in the second phase of the project, for which we expect a broader involvement of the National academic and research community is expect.

3.4. Documents

A document can be stored in ArquiTec in multiple formats (such as ASCII, HTML, PostScript, GIF, etc.), a feature supported by DIENST. The structure for the management of the documents, supporting the already mentioned three conceptual spaces (*archive, formal* and *informal*), is shown in Figure 6.

The official space corresponds to the central archive, and any document existing there has (or had) a copy in the formal space. This central archive will be created, managed and maintained by the Portuguese National Library.

Formal spaces are intended to represent local specialized collections, managed by universities, research laboratories, departments, etc. An ArquiTec local server is itself a local digital library, depending on the central

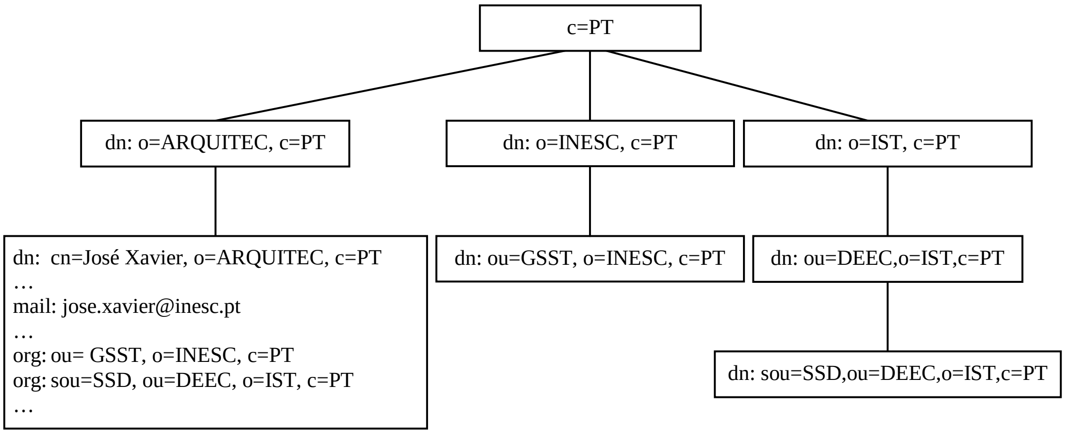
c=PT
dn: o=ARQUITEC, c=PT
dn: o=INESC, c=PT
dn: o=IST, c=PT
dn: cn=José Xavier, o=ARQUITEC, c=PT
…
mail: jose.xavier@inesc.pt
…
org: ou= GSST, o=INESC, c=PT
org: sou=SSD, ou=DEEC, o=IST, c=PT
…
dn: ou=GSST, o=INESC, c=PT
dn: ou=DEEC,o=IST,c=PT
dn: sou=SSD,ou=DEEC,o=IST,c=PT

Figure 5: The X.500 user directory in ArquiTec.

service server for updating the thesauri (functional independence of the central user directory server can be achieved by replicating it locally or at any other available site).

The support for HTTP and FTP informal spaces was done by integrating Harvest brokers in local servers. Users' "home-pages" (and associated "webs") registered in the user directory are automatically elected for classification in this space, for example. Users can register other "webs" in the informal space, just like documents, by filling a submission form were they provide the HTTP or FTP address, the "depth" of links to be followed and a normal metadata structure. Since ArquiTec doesn't control the preservation of the informal and formal spaces, PURLs are not created for documents in these spaces.

For metadata we adopted the original RFC 1807 format used by NCSTRL [13]. In fact we extended its usage, since we also decided to register annotations in that format (each annotation becomes a new metadata file). However a change to the Dublin Core [14] format is under consideration, and if agreed upon we can implement it in the short term.

As it was already mentioned, a PURL service is provided at the National Library. Each time a formal document is submitted to a local collection, a PURL for its bibliographic page is automatically created. This bibliography page is an HTML page suitable for human access. It presents the metadata associated to the document (authors, abstract, etc.) and a list of links to the available document versions and annotations (the PURLs are always related to the bibliographic pages in Portuguese, where links give access to other languages).

When a formal document is archived, the value associated to its PURL is changed to that of the archived version. However, when that document is searched via a query to its local server, its PURL is visible to the user but the versions available for retrieval are those stored locally. This is possible because bibliography page only presents links to the local versions of the document (in this sense local collections are used as caches for official documents).

Any registered user can contribute with new documents to the library and add annotations to existing documents. Annotations are normal metadata structures automatically stored and indexed by the digital library, which become conceptually attached to their corresponding documents.

Submitting an annotation may originate automatic notifications, composed by electronic mail messages sent to specific users such as the document authors and users that had retrieved it. A similar situation occurs when a new document is submitted. With this service we intend to comply with the NDL concept, in order to stimulate the process of creation of information. In fact, and as Jean-Claude Guedon wrote it:

"*In all cases, print serves diffusion rather than communication, because it does not lend itself in reality to two-way dialogs. Readers' feedback sometimes appears, but in the best of cases it remains a minor part of the printing enterprise.*

By contrast, electronic publishing, even when designed to look as much as possible like traditional print, lends itself naturally to dialog and feedback" [15].

As it was mentioned, the Portuguese National Library will maintain a central archive with a copy of formal or refereed documents (and related annotations). Currently, the documents to archive are copied from local collections only after explicit selection by the library staff. The reason for this procedure is not technical, but organizational, since it is difficult to foresee how often and in what manner different communities will use ArquiTec. This is also a completely new reality for the National Library which, despite their experience with printed material, have yet to accumulate similar know-how in order to establish equivalent rules for on-line publications. In the future, it is expected that the central server will automatically gather from the local sites specific kinds of well identified genres of documents, such as theses, dissertations and registered on-line journals.

3.5. Concept Space

A thesaurus is traditionally used in a library in pre-coordination tasks, such as to help in the classification of documents, or in post-coordination tasks, such as to help in the search.

In a pre-coordination scenario the thesaurus helps the librarian to choose the correct terms to classify documents, avoiding ambiguities. Usually documents are classified using accepted and well-defined terms, maintained in an index or authorities list. When a librarian intends to classify a document with a term that does not exist in that index, it is necessary to consult the thesaurus and try to find the correct equivalence in the authorities list. In a search task, the user must check that list before (and also the thesaurus if necessary) in order to ensure that terms they want to use exist. These pre-coordination practices are very common in central libraries. Although these require considerable manpower to classify documents, they ensure good consistency (the maintenance of thesauri and authorities lists require very formal and rigid procedures, with very well defined updates).

A post-coordination scenario affords more freedom to the classification task. It becomes possible to use "more correct" terms, independently of earlier decisions taken when classifying other related documents. When searching, users expect the system thesaurus to provide assistance (whether "manual" or "automatic") to find equivalent or related terms. The post-coordination practice needs much less effort to classify documents, at the cost of increased efforts in thesaurus maintenance and considerable resources dedicated to the search task (usually possible only with the help of computers).

Most recently, the increasing scholarly and scientific activity has resulted in a growth of publications rich in new and interdisciplinary perspectives, raising serious problems for traditional libraries where collections were usually classified with the help of formal thesauri and catalogs defined by static structures.

In order to deal with this dynamic classification problem, ArquiTec incorporates an ontology space, composed by user contributions and the integration of multiple thesauri.

We supplement formal thesauri with *statistical thesauri*, based in relationships automatically extracted from ArquiTec´s document repositories and user directory. For example, we assume that two subjects are related if there is at least one document explicitly or implicitly classified with both the subjects, or even if there is a user interested in both subjects. Such relationships are maintained in internal thesauri and can be presented as links to the document or to the user profile.

User contributions are also an important feature of our ontology, giving users the chance to contribute with new meta-knowledge to enrich the system. In that sense, users can suggest new keywords for documents or new relationships to thesauri, for example.

In summary, two sources support our ontology, reflecting complimentary perspectives:

- *Formal sources*: it is possible to import external formal thesauri, developed independently of ArquiTec and currently used as guidelines to users for classification tasks, in a pre-coordination scenario.
- *Informal sources*: statistical thesauri are created with extra relationships automatically extracted from document repositories, user directory and document annotations (the latter can contribute to document classification).

The thesauri follow the ISO-5964 standard (for multilingual thesauri), and it is possible to import them to the system (or export them) in a portable format defined in MCF.

4. Related Work and Open Issues

Several paradigmatic approaches have been developed to provide access to technical and scientific documents in digital formats.

For a short historical perspective, we could start by referring to the CORE project, started in 1991 aimed at building a database of scanned journals published by the American Chemical Society [16]. By the end of 1994 the project had a database of more than 400,000 pages of full text and graphics (in magnetic tapes and CD-ROM). The text was converted to ASCII and marked-up with SGML (Standard General Markup Language), and dedicated X-Windows interfaces were developed to access the database. While reported user acceptance of project results was high, "*the task of building and maintaining electronic journal databases remains formidable*."

Another contemporary initiative was the TULIP project sponsored by Elsevier Science. It required the efforts of nine American universities from March 1991 to the end of 1995 [17] to research and test systems for networked delivery and use of scanned journals. Elsevier contributed the scanned page images, OCR generated text and bibliographic data from 43 engineering and materials science journals. When the project started, the Internet was already a reality, but the Web was still in an embryonic state. Due to that, the delivery technology was based on alphanumeric clients for mainframe terminals and dedicated graphical clients for X-Windows, MS-Windows and Apple Macintosh. But soon the maintenance costs were evident, and the project shifted to WWW technology when its advantages and maturity became recognized. In its final conclusions, the project pointed out that the transition from conventional to digital libraries (defined here as libraries with full digital contents), will take much longer and cost more than commonly thought, mainly due to maintenance costs, network bandwidth and storage limitations. However, and as pointed out by the CORE project, we think that this conclusion can not be dissociated from the approach taken: to scan the original material. For example, it was estimated in TULIP that a typical journal issue, with 20 articles and 200 pages, requires approximately 17 Mbytes of storage, with 16 Mbytes for the scanned TIFF pages. By comparison, the ASCII information resulting from the OCR requires only 800 Kbytes, plus 200 Kbytes for indexing and bibliographic information in SGML format (for ArquiTec we assume a different scenario, since nowadays documents are commonly produced in digital format using word processors).

Recently more pragmatic approaches were taken in a series of projects in the Computer Science Reports area, using this perspective. Some of the most representative were UCSTRI (Unified Computer Science Technical Report Index [18]), NTRS (NASA Technical Report Server [19]), WATERS (Wide Area Technical Report Service [20]) and CSTR (Computer Science Technical Reports [21]). A common goal of those projects has been easy installation and maintenance of server sites and support for heterogeneous collections. The idea has been not only to provide scanned versions of printed documents, but also to take advantage of the fact that most printed documents are already produced in digital format (such as MS-Word, PDF, HTML, etc.).

In April 1995, WATERS and CSTR projects joined efforts and conceived NCSTRL, a system that had an important influence in ArquiTec, as already described. Other recent similar initiatives that we will follow with a special interest are projects MEDOC (http://medoc.informatik.tu-muenchen.de) and NDLTD (Networked Digital Library of Theses and Dissertations [22]). While MEDOC concentrates in Computer Science contents, NDLTD, like ArquiTec, aims to extend the content base to provide a digital library of theses and dissertations in a wide range of technical and scientific fields. NDLTD started as a national initiative, limited to the United States, but it is now open to cooperation with other countries and regions, with the aim to become a worldwide system. In that sense we have been developing contacts in order to integrate and have ArquiTec accessible through NDLTD.

Other important and interesting initiatives have addressed the "digital library" problem from a broader perspective.

Telematics for Libraries is a part of a vast Telematics Program launched in 1991 by the European Community (http://www.echo.lu). The first phase ran from 1991 to 1994, and launched about 80 actions and projects. The majority of these initiatives focused in institutional library services and inter-library networking (an area where Europe was relatively lagging behind, mainly due to its diversified political and cultural structure), but other topics have also been addressed, such as electronic publishing, delivery of electronic publications, copyright, etc. A second phase of the program started in 1995, and a final third phase will start in the end of 1997. INESC and the Portuguese National Library are associated, with seven other European national libraries, in NEDLIB, an European consortium that will run a project in the third phase of Telematics to address the problems of gathering, installing and preserving digital publications by national libraries. The second phase of ArquiTec will run in parallel with this project, and we expect a strong interaction.

Another interesting initiative is the Electronic Library Programme (eLib), a national UK program launched in the end of 1994 [23]. This program follows the recommendations of the Follett Report, the result of a joint commission of several higher education and library funding institutions in the UK. That commission had the mission to investigate "*how to deal with the pressures on library resources caused by the rapid expansion of student members and the world-wide explosion in academic knowledge and information*." As a result, eLib has supported a large number of pragmatic small projects, with more than 60 already approved (http://www.ukoln.ac.uk/elib/flyers). The program has a broad range, covering areas also addressed by ArquiTec. Until now we had a few informal contacts with projects of this initiative, but we expect to have the chance to provide them more attention in the future.

Especially interesting for us are the projects in the areas of Pre-Prints and Access to Network Resources, with similar approaches to our concept of informal collection (such as the projects ROADS and SOSIG), as also the projects in the area of Electronic Journals.

Finally we must refer the Digital Library Initiative (DLI), a four-year program started in the end of 1994 in the United States and managed jointly by NSF (National Science Foundation), ARPA (Advanced Research Projects Agency), and NASA (National Aeronautics and Space Administration) [24]. The global program focus in ways to advance the means to collect, store and organize information in digital form, and to make it available for searching, retrieval and processing via communication networks with emphasis on the Internet. Six research projects have been funded, each one lead by one university that will develop a testbed mainly for research and prototyping purposes.

While these projects cover a broad spectrum, the most interesting for ArquiTec are the projects ongoing at Berkeley (automated indexing and intelligent retrieval) [25], Michigan (large-scale multimedia libraries, software agents and ontologies) [26], Stanford (heterogeneous collections) [27], and Illinois (semantic retrieval and sociological evaluation) [28]. Other projects are addressing digital video and natural language (CMU) [29] and spatially indexed and graphical information (Santa Barbara) [30].

Concerning the main research issues that specifically require our attention in ArquiTec, we have:

- *URNs: a support for a more complete URN service than the PURL servers is under consideration.*
- *Copyright: the problem of definition and management of the copyright of the used material is an open issue, very important if we intend to include other kinds of material in our system.*
- *Security: requirements for secure authentication and certification authorities, important for example for managing documents with access restrictions (an issue related with the URN and copyright management problems).*
- *Natural language classification and search: with a special focus on the Portuguese language.*
- *Long term preservation: to ensure that official repositories will survive technology changes, such as new storage systems, document formats, etc. (this is of special concern to the National Library).*

We have been carrying out work concerned also with the integration of other spaces, accessible through new interfaces to local servers. Examples are interfaces for Z39.50, useful for the integration of OPAC systems such as the catalogs of conventional libraries, and HARVEST brokers, a work already started and useful for the support of informal publications and other similar material such as "home-pages", archived mailing lists, etc. This work is intend to expand the range of materials and services available through ArquiTec, in order to better fulfil its role as a Networked Digital Library.

References

[1] Bush, V. (1945). As We May Think. The Atlantic Monthly, July 1945. Available on-line in 13 May 1887 at http://www2.theatlantic.com/Atlantic/atlweb/flashbks/computer/Bushf.htm

[2] Ackerman, M. (1994). Metaphors along the Information Highway. Proceedings of the Symposium on Directions and

Impacts of Advanced Computing (DIAC'94), Cambridge, MA, April 1994. Available on-line in 13 May 1997 at http://www.ics.uci.edu/ackerman/docs/diac94/diac.final.html

[3] Kahn, R.; Wilensky, R. (1995). A Framework for a Distributed Digital Object Services. CS-TR Report, May 1995. Available on-line in 13 May 1997 at http://WWW.CNRI.Reston.VA.US/home/cstr/arch/k-w.html.

[4] Weibel, S.; Jul, E. (1995). PURLs to improve access to Internet. OCLC Newsletter, November/December 1995, 19. Updated version Available on-line in 13 May 1997 at http://purl.oclc.org/OCLC/PURL/SUMMARY.

[5] Davis, J. R. (1995). Creating a Networked Computer Science Technical Report Library. D-Lib Magazine, September 1995. Available on-line in 13 May 1997 at http://www.dlib.org/dlib/september95/09davis.html

[6] Davis, J. R.; Lagoze, C. (1994). A protocol and server for a distributed digital technical report library. Technical Report TR94-1418, Computer Science Department, Cornell University, 1994.

[7] Yeong, W.; Howes, T.; Kille, S. (1995). RFC 1777: Lightweight Directory Access Protocol. IETF Network Working Group, March 1995. Available on-line in 13 May 1997 at http://ds.internic.net/rfc/rfc1777.txt

[8] CCITT (1988). X.500 The Directory: Overview of Concepts, Models and Service. CCITT Recommendation X.500, 1988.

[9] International Organization for Standardization (1985). ISO-5964: Documentation Guidelines form the establishment and development of multilingual thesaurus. ISO, 1985.

[10] Gutha, R. V. (1997). Meta-Content Framework. Apple Computer White Paper Available on-line in 13 May 1997 at http://mcf.research.apple.com/hs/mcf.html

[11] Manber, U.; Wu, S. (1993). GLIMPSE: A Tool to Search Through Entire File System. University of Arizona Technical Report TR 93-34.

[12] Kille, S. (1995). RFC 1779: A string Representation of Distinguished Names. IETF Network Working Group, March 1995. Available on-line in 13 May 1997 at http://ds.internic.net/rfc/rfc1807/rfc1779.txt

[13] Lasher, R.; Cohen D. (1995). RFC 1807: Format for Bibliographic Records. June 1995. Available on-line in 13 May 1997 at http://ds.internic.net/rfc/rfc1807.txt

[14] Weibel, S. (1995). Metadata: The Foundations of Resource Description. D-Lib Magazine, July 1995. Available on-line in 13 May 1997 at http://www.dlib.org/dlib/July95/07weibel.html

[15] Guedon, J. (1994). Why are Electronic Publications Difficult to Classify? : The Orthogonality of Print and Digital Media. Directory of Electronic Journals, Newsletters and Academic Discussion Lists, Association of Research Libraries, May 1994. Available on-line in 13 May 1997 at gopher://arl.cni.org/00/scomm/edir/guedon.94

[16] Entlich, R.; Garson, L.; Lesk, M.; Normore, L.; Olsen, J.; Weibel, S. (1995). Making a Digital Library: The Chemistry Online Retrieval Experiment. Communications of the ACM, April 1995, Vol. 38, Number 4, 54.99

[17] Elsevier Science (1996). TULIP Final Report. Elsevier Science Edition. Available at http://www.elsevier.nl/locate/tulip

[18] VanHeyningen, M. (1994). The Unified Computer Science Technical Report Index: Lessons in Indexing Diverse Resources. Second International World Wide Web Conference, 1994, 535-543.

[19] Nelson, M. L.; Gottlich, G. L.; Bianco, D. J.; Paulson, S. P.; Binkley, R. L.; Kellog, Y. D.; Beaumont, C. J.; Schmunk, R. B.; Kurtz, M. J.; Accomazzi, A.; Syed, O. (1994). The NASA Technical Report Server. Internet Research: Electronic Network Applications and Policy, Vol. 5, No 2, 25-36.

[20] French, J. C.; Fox, E. A.; Maly, K. (1995). Wide Area Technical Report Service: Technical Reports Online. Communications of the ACM, April 1995, Vol. 38, Number 4, 45.

[21] Anderson, G.; Lasher, R.; Reich, V. (1996). The Computer Science Technical Report (CS-TR) Project: A Pioneering Digital Library Project Viewed from a Library Perspective. The Public-Access Computer Systems Review 7, No 2, 1996. Available on-line in 13 may 1997 at http://info.lib.uh.edu/pr/v7/n2/ande7n2.html

[22] Fox, E. A.; Eaton, J. L.; McMillan, G.; Kipp, N. A.; Weiss, L.; Arce, E.; Guyer, S. (1996). National Digital Library of Theses and Dissertations. D-Lib Magazine, September 1996. Available at http://www.dlib.org/dlib/september96/theses/09fox.html

[23] Rusbridge, C. (1995). The UK Electronic Libraries Programme. D-Lib Magazine, December 1995. Available at http://www.dlib.org/dlib/december95/briefings/12uk.html

[24] Schatz, B.; Chen, H. (1996). Building Large-Scale Digital Libraries. IEEE Computer, May 1996, Vol. 28, Number 5, 22-26.

[25] Wilensky, R. (1996). Toward Work-Centered Digital Information Services. IEEE Computer, May 1996, Vol. 28, Number 5, 37-44.

[26] Atkins, D. E.; Birmingham, W. P.; Durfee, E. H.; Glover, E. J.; Mullen, T.; Rundensteiner, E. A.; Soloway, E.; Vidal, J. M.; Wallace, R.; Wellman, M. P. (1996). Toward Inquiry-Based Education Trough Interacting Software Agents. IEEE Computer, May 1996, Vol. 28, Number 5, 69-76.

[27] Paepcke, A.; Cousins, S. B.; Garcia-Molina, H.; Hassan, S. W.; Ketchpel, S. P.; Röscheisen, M.; Winograd, T. (1996). Using Distributed Objects for Digital Library Interoperability. IEEE Computer, May 1996, Vol. 28, Number 5, 61-68.

[28] Schatz, B.; Mischo, W. H.; Cole, T. W.; Hardin, J. B.; Bishop, A. P.; Chen, H. (1996). Federating Diverse Collections of Scientific Literature. IEEE Computer, May 1996, Vol. 28, Number 5, 28-36.

[29] Wractlar, H. D.; Kanade, T.; Smith, M. A.; Stevens, S. M. (1996). Intelligent Access to Digital Video: Informedia Project. IEEE Computer, May 1996, Vol. 28, Number 5, 46-52.

[30] Smith, T. R. (1996). A Digital Library for Geographical Referenced Materials. IEEE Computer, May 1996, Vol. 28, Number 5, 54-60.