The Personalized, Collaborative Digital Library Environment CYCLADES and Its Collections Management

Leonardo Candela and Umberto Straccia

I.S.T.I. – C.N.R.
Via G. Moruzzi, 1 I-56124 Pisa (PI) ITALY
{Leonardo.Candela,Umberto.Straccia}@isti.cnr.it

Abstract. Usually, a Digital Library (DL) is an information resource where users may submit queries to satisfy their daily information need. The CYCLADES system envisages a DL additionally as a personalized collaborative working and meeting space of people sharing common interests, where users (i) may organize the information space according to their own subjective view; (ii) may build communities, (iii) may become aware of each other, (iv) may exchange information and knowledge with other users, and (v) may get recommendations based on preference patterns of users. In this paper, we describe the CYCLADES system, show how users may define their own collections of records in terms of un-materialized views over the information space and how the system manages them. In particular, we show how the system automatically detects the archives where to search in, which are relevant to each user defined collection.

1 Introduction

Digital Libraries (DLs) [11] will play an important role not merely in terms of the information provided, but in terms of the *services* they provide to the information society. Informally, DLs can be defined as consisting of collections of information which have associated services delivered to user communities using a variety of technologies. The collections of information can be scientific, business or personal data, and can be represented as digital text, image, audio, video, or other media. Even though DLs have evolved rapidly over the past decade, typically, DLs still are limited to provide a search facility to the digital society at large. As DLs become more commonplace and the range of information they provide increases, users will expect more and more sophisticated services from their DLs. There is a need for DLs to move from being passive with little adaptation to their users, to being more proactive, or *personalized*, in offering and tailoring information for individual users. The requirement of a personalized search 'assistant' in the context of DLs is already known and, to date, some DLs provide related, though simplified, search functionality (see *e.g.* [3,7,8,9,10,13,14,17]). Informally, these DLs may fall in the so-called category of *alerting services*, *i.e.* services that notify a user (by sending an e-mail), with a list of references to new documents deemed as relevant. But, *searching* is just one aspect that should be addressed. Another orthogonal aspect of personalization concerns *information organization*, *i.e.* to support the users' interest in

J. Callan et al. (Eds.): SIGIR 2003 Ws Distributed IR, LNCS 2924, pp. 156–172, 2003.

being able to organize the information space they are accessing according to *their own subjective perspective* (see *e.g.* [7,9]). Additionally, very seldom[1], a DL is also considered as a *collaborative meeting place* of people sharing common interests. Indeed, a DL may be viewed as a *common working place* where users may become aware of each other, open communication channels, and exchange information and knowledge with each other or with experts. In fact, usually users and/or communities access a DL in search of some information. This means that it is quite possible that users may have overlapping interests if the information available in a DL matches their expectations, backgrounds, or motivations. Such users might well profit from each other's knowledge by sharing opinions or experiences or offering advice. Some users might enter into long-term relationships and eventually evolve into a community if only they were to become aware of each other.

CYCLADES[2] is a DL environment supporting collaboration and personalization at various level, where users and communities may search, share and organize their information space according to their own personal view. While an extensive presentation of the CYCLADES system and its algorithm for filtering and recommendation has been given elsewhere [16], in this paper we will focus on how users may tailor the information space according to their subjective view. In particular, we will address the notion of *personalized (virtual) collections*. These are user defined un-materialized views over very heterogeneous information space, consisting of the archives adhering to the *Open Archives Initiative*[3] (OAI), available within CYCLADES. The main purpose of these personalized collections is to restrict the information space during the user's search task. To this purpose CYCLADES provides techniques of automated source selection [5,12] to automatically detect the archives relevant to a view.

The outline of the paper is as follows. In the next section we will recall the main features of CYCLADES, while in Section 3 we will address the management of personalized collections in CYCLADES and report some preliminary experimental results on the automated source selection procedure, which is at the core of the personalized collection management. Section 4 concludes the paper.

2 CYCLADES: A Personalized and Collaborative DL

The objective of CYCLADES is to provide an integrated environment for users and groups of users (communities) that want to use, in a highly personalized and flexible way, 'open archives', *i.e.* electronic archives of documents compliant with the OAI standard. Informally, the OAI is an initiative between several Digital Archives in order to provide interoperability among them. In particular, the OAI defines an easy-to-implement gathering protocol over HTTP, which give *data providers* (the individual archives) the possibility to make the documents' metadata in their archives externally available. This external availability of the metadata records then makes it possible for *service providers* to build higher levels of functionality. To date, there is a wide range of archives available in terms of its content, *i.e.* the family of OAI compliant archives is multidisciplinary

[1] [7] is an exception.

[2] `http://www.ercim.org/cyclades`

[3] www.openarchives.org.

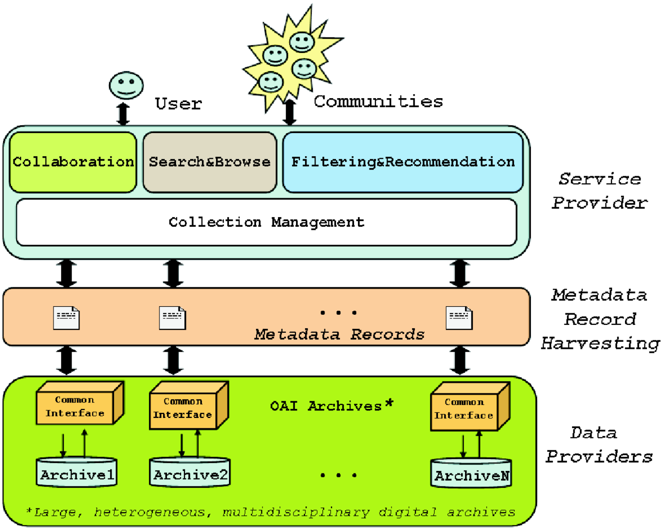
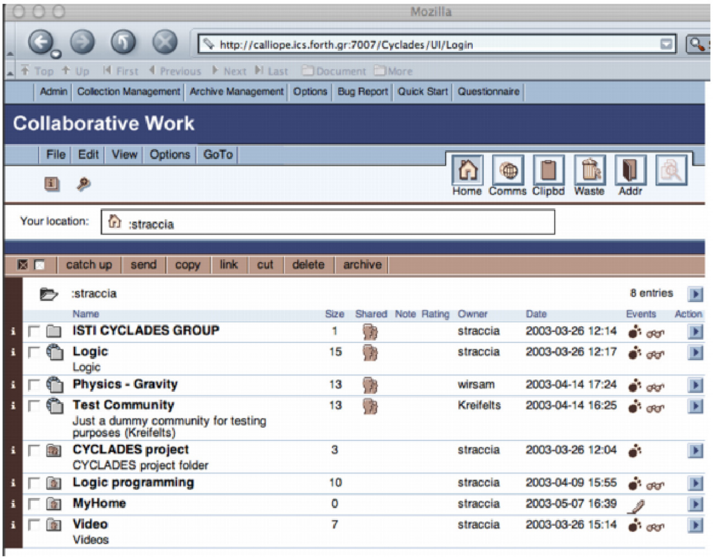


Fig. 1. Logical view of CYCLADES functionality.

in content. Under the above definition, CYCLADES *is an OAI service provider* (see Figure 1.) and provides functionality for (i) advanced search in *large, heterogeneous, multidisciplinary digital archives*; (ii) collaboration; (iii) filtering; (iv) recommendation; and (v) the management of records grouped into *collections*. Worth mentioning, the main principle underlying CYCLADES is the *folder paradigm* (see Figure 2). That is, users and communities of users may organize the information space into their own folder hierarchy, as *e.g.* may be done with directories in operating systems, bookmark folders in Web browser and folders in e-mail programs. A folder becomes a holder of information items, which are usually semantically related and, thus, implicitly determines what the folder's topic is about. Therefore, rather than speaking about a user profile, we will deal with a *folder profile*, *i.e.* a representation of what a folder is *about*. As a consequence, the user's set of folder profiles represents the set of topics the user is interested in and, thus, the profile of a user consists of the set of profiles related to his folders.

Figure 2, shows the home (top level) folder of a user. It contains several sub-folders. Among them, there are some (shared) folders belonging to communities (created by someone) to which the user joined to, like the 'Physics-Gravity' folder (community), while others are private folders and have been created directly by the user, *e.g.* the 'Logic Programming' folder. These folders contain community or user collected OAI records relevant to some topics (*e.g.* gravity and logic programming, respectively). Figure 3 shows the content of a folder, in our case the 'Physics-Gravity' folder of the community of physicists. In it there are several other folders and metadata records. Some records have been rated (*e.g.* the 'Astronaut Protection ... ' record) and some records have notes attached (*e.g.* 'The Lunar Scout ... ' record). There is also a discussion forum. These functionality are only some of those pertaining to the collaborative support package. Note


Mozilla
http://calliope.ics.forth.gr:7007/Cyclades/UI/Login
Top Up First Previous Next Last Document More
Admin Collection Management Archive Management Options Bug Report Quick Start Questionnaire
Collaborative Work
File Edit View Options GoTo
Home Comms Clipbd Waste Addr
Your location: :straccia
catch up send copy link cut delete archive
:straccia
8 entries

Name	Size	Shared	Note	Rating	Owner	Date	Events	Action
ISTI CYCLADES GROUP	1				straccia	2003-03-26 12:14		
Logic Logic	15				straccia	2003-03-26 12:17		
Physics - Gravity	13				wirsam	2003-04-14 17:24		
Test Community Just a dummy community for testing purposes (Kreifelts)	13				Kreifelts	2003-04-14 16:25		
CYCLADES project CYCLADES project folder	3				straccia	2003-03-26 12:04		
Logic programming	10				straccia	2003-04-09 15:55		
MyHome	0				straccia	2003-05-07 16:39		
Video Videos	7				straccia	2003-03-26 15:14		

Fig. 2. User interface: a user home folder.

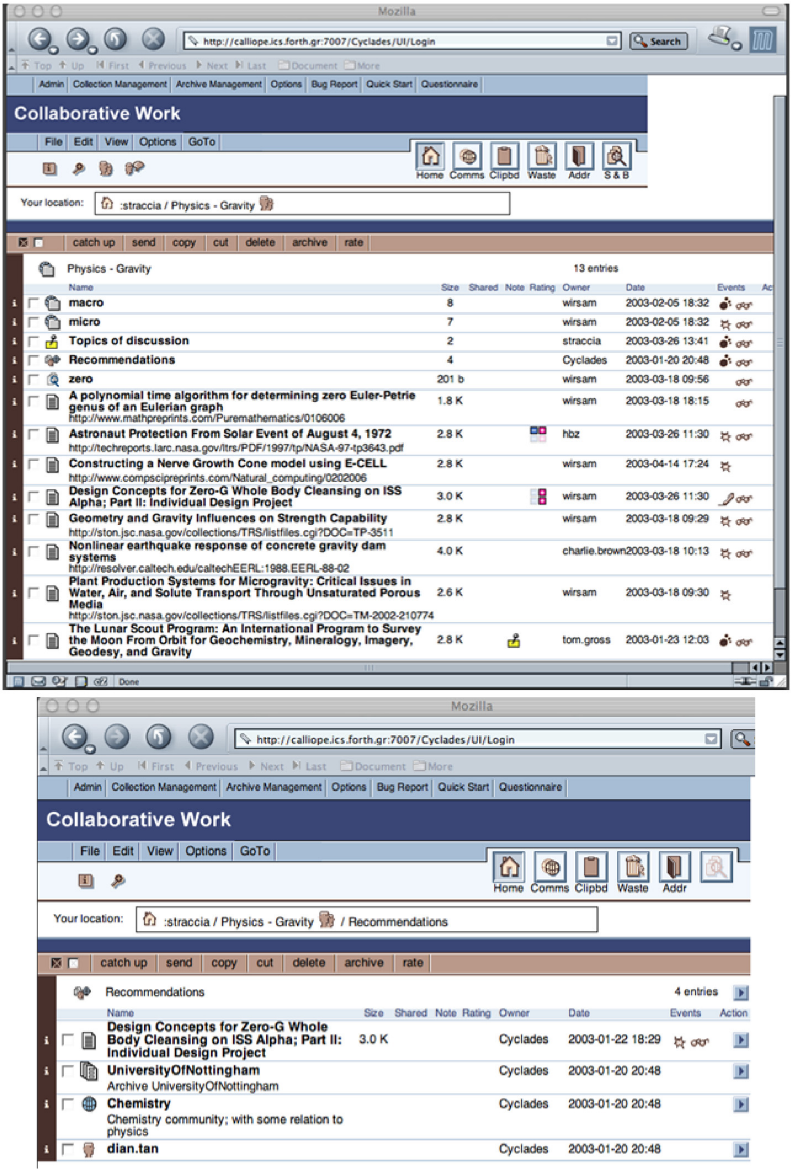
also that the CYCLADES system already provided some record, community, collection and user recommendations deemed by the system as relevant to this folder. Records retrieved after a search task may be stored in the folder by the user. This is the main way to populate folders with records gathered from CYCLADES information space (*i.e.* the OAI archives).

The architecture of the CYCLADES system is depicted in Figure 4. It should be noted that each box is a service accessible via the Web distributed over the Internet. The CYCLADES system, accessible through Web browsers, provides the user with different environments, according to the actions the user wants to perform. The functionality CYCLADES provides are developed by different services described next.

The *Collaborative Work Service* provides the folder-based environment for managing metadata records, queries, collections, external documents, received recommendations, ratings and annotations. Furthermore, it supports collaboration between CYCLADES users by way of folder sharing in communities, discussion forums and mutual awareness.

The *Search and Browse Service* supports the activity of searching records from the various collections, formulating and reusing queries associated to the folder by the user, and saving records to folders.

The *Access Service* is in charge of interfacing with the underlying metadata archives. In this project, only archives adhering to the OAI specification will be accounted for. However, the system is extensible to other kinds of archives by just modifying the Access



Name	Size	Shared	Note	Rating	Owner	Date	Events	Ac
macro	8				wirsam	2003-02-05 18:32		
micro	7				wirsam	2003-02-05 18:32		
Topics of discussion	2				straccia	2003-03-26 13:41		
Recommendations	4				Cyclades	2003-01-20 20:48		
zero	201 b				wirsam	2003-03-18 09:56		
A polynomial time algorithm for determining zero Euler-Petrie genus of an Eulerian graph http://www.mathpreprints.com/Puremathematics/0106006	1.8 K				wirsam	2003-03-18 18:15		
Astronaut Protection From Solar Event of August 4, 1972 http://techreports.larc.nasa.gov/ltrs/PDF/1997/tp/NASA-97-tp3643.pdf	2.8 K				hbz	2003-03-26 11:30		
Constructing a Nerve Growth Cone model using E-CELL http://www.compscipreprints.com/Natural_computing/0202006	2.8 K				wirsam	2003-04-14 17:24		
Design Concepts for Zero-G Whole Body Cleansing on ISS Alpha; Part II: Individual Design Project	3.0 K				wirsam	2003-03-26 11:30		
Geometry and Gravity Influences on Strength Capability http://ston.jsc.nasa.gov/collections/TRS/listfiles.cgi?DOC=TP-3511	2.8 K				wirsam	2003-03-18 09:29		
Nonlinear earthquake response of concrete gravity dam systems http://resolver.caltech.edu/caltechEERL:1988.EERL-88-02	4.0 K				charlie.brown	2003-03-18 10:13		
Plant Production Systems for Microgravity: Critical Issues in Water, Air, and Solute Transport Through Unsaturated Porous Media http://ston.jsc.nasa.gov/collections/TRS/listfiles.cgi?DOC=TM-2002-210774	2.6 K				wirsam	2003-03-18 09:30		
The Lunar Scout Program: An International Program to Survey the Moon From Orbit for Geochemistry, Mineralogy, Imagery, Geodesy, and Gravity	2.8 K				tom.gross	2003-01-23 12:03		

Name	Size	Shared	Note	Rating	Owner	Date	Events	Action
Design Concepts for Zero-G Whole Body Cleansing on ISS Alpha; Part II: Individual Design Project	3.0 K				Cyclades	2003-01-22 18:29		
UniversityOfNottingham Archive UniversityOfNottingham					Cyclades	2003-01-20 20:48		
Chemistry Chemistry community; with some relation to physics					Cyclades	2003-01-20 20:48		
dian.tan					Cyclades	2003-01-20 20:48		

Fig. 3. User interface: folder content and recommendations.

Service only. A user may also ask CYCLADES to include newly OAI compliant archives as well.

The *Collection Service* manages personalized collections (*i.e.* their definition, creation, and update) and stores them, thus allowing a dynamic partitioning of the in-

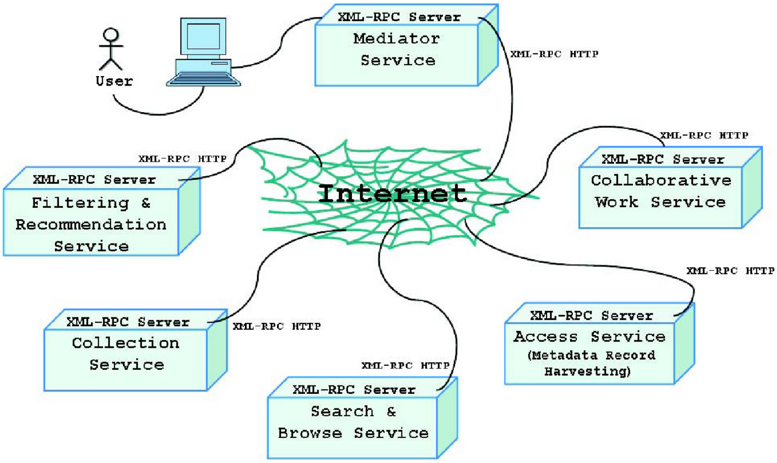


Fig. 4. Architecture.

formation space according to the users' interests, and making the individual archives transparent to the user.

The *Filtering and Recommendation Service* provides filtered search, recommendations of records, collections, users, and communities deemed relevant to the user's interests.

Finally, the *Mediator Service*, the main entry point to the CYCLADES system, acts as a registry for the other services, checks if a user is entitled to use the system, and ensures that the other services are only called after proper authentication.

The Collaborative Work Service, the Search and Browse Service, the Access Service, and the Collection Service provide their own user interfaces. The Mediator Service itself provides the registration and login interface, and a system administration interface (for assigning access rights, etc.). Additionally, the Mediator Service integrates the user interfaces of the other services, and makes sure that those services and their interfaces are called only for authorized users, and only via the Mediator Service.

3 Personalized Collection Management in CYCLADES

We present some details on the management of personalized collections within CYCLADES, in particular in the Collection Service. As already addressed, the Collection Service introduces a mechanism to support users and/or communities (in the following called *agent*) to define their own information space. Usually, a collection is meant to reflect a topic of interest of an agent, *e.g.* the collection of records about information retrieval. To facilitate an agent's personalized view over the information space, an agent may organize its own defined collections into an hierarchical order. A major distinction of the Collection Service is that collections are not materialized, but are rather *personalized (virtual) collections*, *i.e.* in database terms, un-materialized views over the information

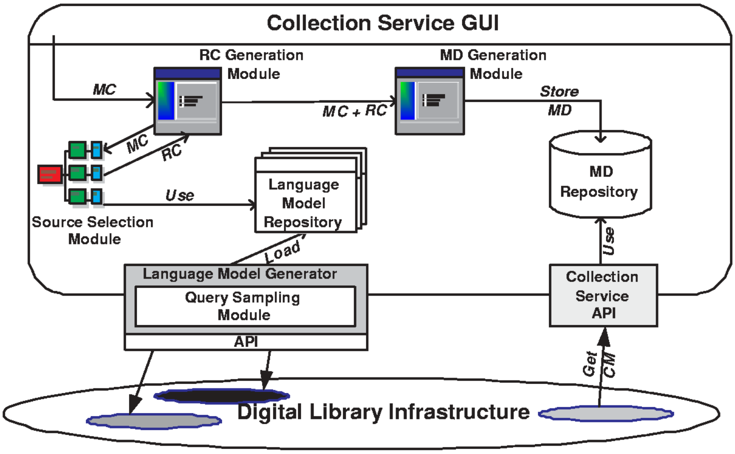
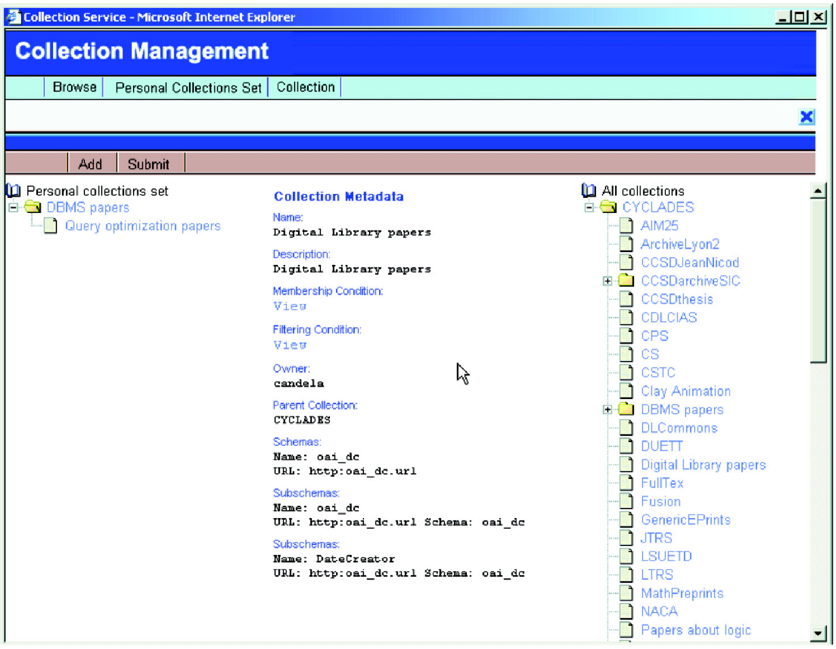


Fig. 5. The CYCLADES Collection Service Logical Architecture.

space. That is, *e.g.* a personal collection of an agent, whose aim is to collect records about ‘information retrieval’ is not a database table containing records, which are about information retrieval, but rather is a specification of the conditions that a record should satisfy in order to belong to this collection. Figure 5 shows the logical architecture of the Collection Service. The agent defines a collection using the *Collection Definition Language* (CDL). By means of the CDL, it is possible to define the *Membership Condition* (MC) of a collection that the metadata records should satisfy in order to be part of the defining collection. To reduce the information space effectively, the Collection Service uses then the MC of a collection to automatically determine the most relevant resources in which to search for records meeting the MC. This is done by relying on techniques known as *automatic source selection* (see, *e.g.* [2,15,12]). This set of determined OAI archives is then added to the MC and the resulting description is called *Retrieval Condition* (RC). The MC and the RC, together with some other collection related data forms the *Collection Metadata* (MD), which is then stored into the MD repository. While the MC is modified by an agent only, its relative RC is modified by the Collection Service whenever appropriate. So, for instance, if a new resource is added to CYCLADES then the RCs are re-computed and the new resource may become part of the MCs. This allows the Collection Service to deal with a dynamic number of OAI compliant archives, whose content may vary over time. That is, the Collection Service follows the dynamism of the underlying information space. As the Collection Service does not have the archives (which are in the Access Service), the Collection Service automatically computes an approximation of the content of each OAI compliant archive registered in CYCLADES, to support the function of automated source selection. This data is stored in the Language Model Repository. The approximation is computed by relying on the so-called *query-based sampling method* [6].

Figure 6 shows an excerpt of the Collection Service user interface. In this particular case, on the right column the set of all collections created within CYCLADES (by users, communities or automatically by CYCLADES itself) is shown. The left column shows the collections pertaining to a particular user, while the middle column shows the specification of a collection. A user may anytime define new collections, add collections to its particular collection, etc. In the following, we will address more specifically some aspects of the Collection Service. Namely, the collection definition language, the query-based sampling method, the automated source selection method and some preliminary test results.



Personal collections set
- DBMS papers
  - Query optimization papers

Collection Metadata

Name:
Digital Library papers

Description:
Digital Library papers

Membership Condition:
View

Filtering Condition:
View

Owner:
candela

Parent Collection:
CYCLADES

Schemas:
Name: oai_dc
URL: http:oai_dc.url

Subschemas:
Name: oai_dc
URL: http:oai_dc.url Schema: oai_dc

Subschemas:
Name: DateCreator
URL: http:oai_dc.url Schema: oai_dc

All collections
- CYCLADES
  - AIM25
  - ArchiveLyon2
  - CCSDJeanNicod
  - CCSDarchiveSIC
  - CCSDthesis
  - CDLCIAS
  - CPS
  - CS
  - CSTC
  - Clay Animation
  - DBMS papers
  - DLCommons
  - DUETT
  - Digital Library papers
  - FullTex
  - Fusion
  - GenericEPrints
  - JTRS
  - LSUETD
  - LTRS
  - MathPreprints
  - NACA
  - Papers about logic

Fig. 6. CS Graphical User Interface for select the Personal Collections Set.

Collection definition language. The Collection Service allows users to specify their own information needs via a declarative collection definition language. Each definition specifies a list of conditions that a record has to satisfy in order to belong to the set. The definitions are soft in the sense that each record may satisfy them to some degree in the unit interval $[0, 1]$. Therefore, a collection may be seen as a *fuzzy set* [19] of records. The language is simple, but expressive and quite usable. Below, we present the CDL syntax in Backus Naur Form (BNF).

```
query       ::= condition* [, (archiveList)]
condition   ::= ([weight,] field, predicate, value)
weight      ::= + | - | 1..1000
field       ::= [schemaName":"]attributeName
predicate   ::= cw | < | <= | >= | > | = | ! =
archiveList ::= archiveName | archiveName, archiveList
```

Roughly, this is an ALTAVISTA-style language where a query is a set of conditions, which are either optional, *mandatory* (+) or *prohibitive* (-). In addition, it allows for weighting of optional conditions. With respect to the structure of metadata records it assumes that they have a one-level structure[4]. It allows the use of a name space (*e.g.* `schemaName`). The set of predicates supported is composed by the classical comparison operators ($<$, $<=$, $>=$, $>$, $=$ and $!=$) plus the `cw` operator used to specify a condition on the content (aboutness) of a text field, *e.g.* `(description, cw, library)` stands for "the value of the attribute `description` is relevant to the term `library`".

Automated source selection in the Collection Service. In the CYCLADES system, a query is issued from the Search and Browse Service. As specified early, from the design choice of CYCLADES a global index for all OAI compliant archives does not exists. Therefore, query evaluation is performed by dispatching a query to all OAI compliant archives. A feature of the Search and Browse Service is that the query may be accompanied with the specification of a collection[5]. In this latter case, the query has to be understood as a refinement of the information space and records are searched only in OAI archives relevant to the collection specification. To this end, the RC, automatically built by the Collection Service from the membership condition, is coupled with the query and the result of this combination is sent to the Access Service to compute the result. The RC consists of the MC plus a set of automatically determined archives most relevant to the conditions specified in the MC only. The language chosen for specifying the MC is similar to the query language supported by the Search and Browse Service. Therefore, no significant query reformulations are necessary to build the RC, except the addition of the determined relevant archives.

The computation of the RC from the MC needs two steps:

- the computation of an approximation of the content of each OAI compliant archive registered in CYCLADES;
- the selection of those archives deemed as most relevant to the collection definition, relying on the approximations of the archives' content.

The first step is done periodically and each time a new archive is added to CYCLADES, while the second step is done immediately after a personalized (virtual) collection has been defined by an agent. RCs are updated periodically as well. In the following, we will address these two steps further. In order to select the relevant information sources for a query the Collection Service has to have a sort of knowledge about their content. For each archive, we build a simple language model of it (a list of terms with their term weight information). We rely on the so-called *query-based sampling method* [6], which has been proposed for automatically acquiring statistical information about the content of an information source. A major feature is that it requires only that an information source provides a query facility and access to the documents, that are in the result of a query. Informally, the method is an iteration of the following steps 2 and 3: (1) issue a random query to the Access Service (as start-up); (2) select the top-k documents;

[4] The assumption about the one-level metadata record structure can be removed using the attribute name path instead of the attribute name.

[5] More precisely, a list of collections is allowed.

Table 1. Sampling Algorithm.

```
1: query = generateInitialTrainingQuery();
2: resultSet = run(query);
3: if(|resultSet| < $L_{tr}$){
4:     go to 1;
5: }else{
6:     updateResourceDescription(resultSet);
7:     if(NOT stoppingCriteria()){
8:         query = generateTrainingQuery();
9:         resultSet = run(query);
10:        go to 6;
11:    }
12: }
```

and (3) build m new queries from n randomly selected terms within the top-k ranked documents. The iteration continues until a stop criterion is satisfied. While [6] worked with plain text, in our context we have Dublin Core metadata records. Table 1 shows the detailed sampling algorithm, customized to the case where text databases have multiple text attributes (*e.g.* bibliographic records, similar to [18]).

This algorithm uses a set of functions:

generateInitialTrainingQuery() it generates the *start* training query. In order to generate a query we need: (a) a set of attributes among which we randomly choose the ones to be used to build the initial condition; and (b) a set of terms among which we randomly choose the ones to fill-in the attributes values. For each selected attribute we randomly select 1 to max_t distinct terms and for each (attribute, value) pair we choose an operator to relate attribute and term into the condition. In the CYCLADES Collection Service prototype we have adopted the following decision: (a) concerning the terms, we use the set of terms that characterize the second and the third level of Dewey Decimal Classification [1], (b) concerning the attributes, we have used Dublin Core[6] fields, (c) $max_t = 4$ and (d) the operator to use is always the `cw` operator.

updateResourceDescritpion() it updates the set of records that represents the resource description. Note that a query must return at least L_{tr} records before the records collected (the top L_{tr}) can be added to the resource description record set. This minimum result size is required because query returning small results do not capture source content well. In our prototype we have used $L_{tr} = 4$ as proposed in [18], this is just another configuration aspect.

stoppingCriteria() it evaluates if the stopping criteria was reached. Callan and Connell [6] stop the sampling after examining 500 documents, a stopping criteria chosen empirically observing that increasing the number of documents examined does not improve significantly the language model.

generateTrainingQuery() it generates the *next* training query. Training queries are generated as follow:

[6] `http://dublincore.org/`

1. randomly select a record R from the resource description record set;
2. randomly select a set of attributes of R to be used in the training query;
3. for each attribute to be included in the training query, construct a predicate on it by randomly select 1 to max_t distinct terms (stopwords are discarded) from the corresponding attribute value, by using the `cw` operator.

At the end of this process a sample of records of the information source is acquired. This set is called resource description and the Collection Service uses it to build the language model of the archive.

We conducted some preliminary experiments to evaluate the quality of the computed archive approximations. We have considered two bibliographic information sources. A very small and homogeneous information source (*Archive 1*, 1616 records, 13576 unique terms after stopwords removal, papers about computer science published by the same authority) and a more large and heterogeneous information source (*Archive 2*, 16721 records, 79047 unique terms after stopwords removing, papers published by different authorities). Note that OAI compliant archives are characterized to be very small in terms of numbers of records ($\leq$ 2000 records) except some few exceptions, like arXiv[7] ($\approx$ 270000 records).

The experimental method was based on comparing the learned resource description of an information source with the real resource description for that information source. Resource descriptions can be represented using two information, a vocabulary V of the set of terms appearing in the information source records and a frequency information for each vocabulary term: the number of documents containing the terms, called *document frequency* (*df*).

In accordance with [6], we have used two metrics to evaluate the quality of resource descriptions acquired by sampling: the *ctf ratio* (CTF) to measure the correspondence between the learned vocabulary (V') and the real vocabulary (V) and the *Spearman Rank Correlation Coefficient* (SRCC) to measure the correspondence between the learned and the real document frequency information. This metrics are calculated using Equation (1) and (2) where ctf_i is the number of times term i occurs in the resource description of information source i, d_i is the rank difference of a common term $t_i \in V' \cap V$. The two term rankings are based on the learned and the actual document frequency df_i. n is the total number of common terms.

$$\mathrm{CTF} = \frac{\sum\limits_{i \in V'} ctf_i}{\sum\limits_{i \in V} ctf_i} \tag{1}$$

$$\mathrm{SRCC} = 1 - \frac{6}{n^3 - n} \sum_{t_i \in V' \cap V} {d_i}^2 \tag{2}$$

Five trials were conducted for each information source and for each trial a resource description of 500 records has been acquired to illustrate the behavior of the measures

[7] http://arxiv.org

above. The results reported here are the average of the results of each trial. Figures 7–10 show respectively the CTF and the SRCC metrics calculated for some Dublin Core attribute. On the x-axis, we varied the number of acquired records.

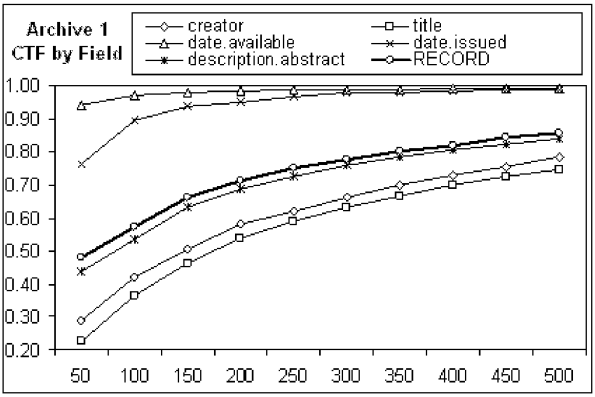


Fig. 7. Archive 1: CTF.

By observing the CTF graphics we can note that the language model acquired for the first archive is better then the one acquired for the second. Moreover we can note that the language model acquired for a field has different characteristics than the one acquired for other fields. The reasons for this behaviour are twofold: *Archive 2* contains more records and is more heterogeneous than *Archive 1* and some attributes, *e.g.* `creator`, are more heterogeneous than others, *e.g.* `date`. The more heterogeneous the values of an attribute are, the more difficult it is to approximate it.

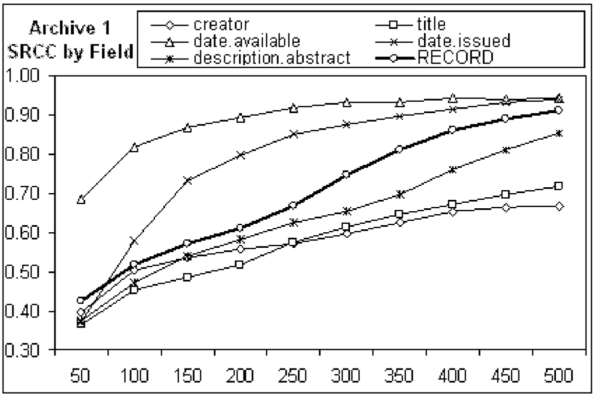


Fig. 8. Archive 1: SRCC.

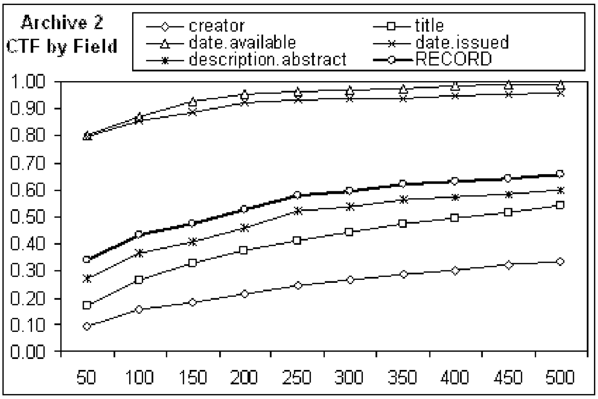


Fig. 9. Archive 2: CTF.

By observing the SRCC graphics we can note that the *quality* of the language model acquired via sampling is high, considering the record as a plain text we can found values greater that 80% (see RECORD line).

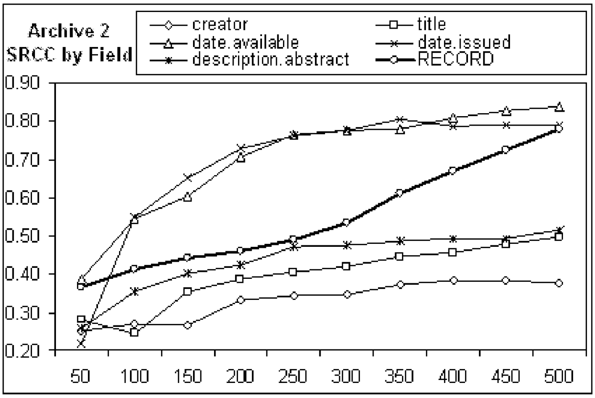


Fig. 10. Archive 2: SRCC.

Let us now deal with the automated source selection problem. Source selection is the problem of selecting from a large set of accessible information sources the ones relevant to a given query. In our case the query is the MC, *i.e.* the collection characterization criteria, while the selected information sources are used in the generation of the RC.

Our approach to automated source selection extends the CORI scheme [4] to the case of OAI-compliant archives, *i.e.* where records in the Dublin Core format are provided. Informally, given a membership condition MC, to each archive IS_i we associate a goodness value $G(MC, IS_i)$ and then select the top-k ranked archives. Formally, let MC be a set of conditions $c_j = (w_i, a_i, o_i, v_i)$ where:

- w_i is the *weight* of this condition, where $w_k \in [1..1000] \cup \{+, -\}$. $+$ means that the condition must be fulfilled, $-$ means that the condition must not be fulfilled (the boolean NOT);
- a_i is the attribute of the Dublin Core record involved in the condition;
- o_i is the operator, *e.g.* <=, =, `cw`, etc.;
- v_i is the term.

For instance, `(+,author,cw,''Straccia'')(+,subject,cw,''Cyclades'')` denotes the (fuzzy) set of records having author "Straccia" and "subject" is related to "Cyclades".

The $Goodness$ score $G(MC, IS_i)$ for an information source IS_i and membership condition MC is defined as follow:

$$G(MC, IS_i) = \begin{cases} 0 & \text{if } \exists k \in [1..|MC|], s.t., w_k \in \{+, -\} \wedge p(c_k|IS_i) = 0 \\ \dfrac{\sum_{k=1}^{|MC|} p(c_k|IS_i)}{|MC|} & \text{otherwise} \end{cases}$$

where the "belief" $p(c_k|IS_i)$ in IS_i, for condition c_k is defined as

$$p(c_k|IS_i) = \begin{cases} T_{i,k} \cdot I_k \cdot w_k & \text{if } w_k \in [1..1000] \\ T_{i,k} \cdot I_k & \text{if } w_k = \text{“}+\text{” or } w_k = \text{“}-\text{”} \end{cases}$$

$$T_{i,k} = \frac{df_{i,k}}{df_{i,k} + 50 + 150 \cdot \dfrac{cw_{i,k}}{\overline{cw_k}}}$$

$$I_k = \frac{\log\left(\frac{|S|+0.5}{cf_k}\right)}{\log\left(|S| + 1.0\right)}$$

where:

$df_{i,k}$ is the number of records in the approximation of IS_i satisfying c_k;
$cw_{i,k}$ is the number of terms in attribute a_k of records in the approximation of IS_i;
$\overline{cw_k}$ is the mean value of $cw._{,k}$ over the approximation of IS_i;
cf_k is the number of approximated information sources that satisfy c_k;
$|S|$ is the number of the information sources.

We carried out some preliminary experiments to evaluate the efficiency and effectiveness of the above source selection method. Indeed, we generated randomly 200 collections using Dublin Core fields. The collections generated are of two kinds: 100 collections (T1) are generated using a combination of conditions on `description` and `title` fields, 100 collections (T2) are generated using a combination of conditions on all fields of the Dublin Core schema. Given a collection definition MC_i and the relative RC_i obtained after source selection, $Precision_i$ is defined as the quantity

$$Precision_i = \frac{|ret(RC_i) \cap ret(MC_i)|}{|ret(RC_i)|}$$

and $Recall_i$ is defined as the quantity

$$Recall_i = \frac{|ret(RC_i) \cap ret(MC_i)|}{|ret(MC_i)|} \ .$$

$ret(MC_i)$ is the set of records retrieved by submitting the MC_i query to CYCLADES consisting of 50 OAI archives (taking the top-100 records). $ret(MC_i)$ is considered as the set of records effectively to be retrieved. $ret(RC_i)$ is the set of records retrieved by submitting RC_i as query (restricting the set of archives in which to search for records). Precision and recall measure how effective the source selection method is with respect to the original query MC_i, which considered all information sources stored in CYCLADES. For each pair (MC_i, RC_i), precision measures the conditional probability $\mathcal{P}(ret(MC_i)|ret(RC_i))$, while recall measures $\mathcal{P}(ret(RC_i)|ret(MC_i))$. For each pair (MC_i, RC_i) we have a precision/recall value $(Precision_i, Recall_i)$. These pairs of values have been partitioned into precision/recall levels and are summarized in Table 2. In it, each row/column pair (r, p), where r and p are intervals denoting respectively recall level and precision level, dictates the percentage of test pairs (MC_i, RC_i) such that $Recall_i \in r$ and $Precision_i \in p$. Furthermore, the right most column and the bottom row report the total amount w.r.t. a row and a column, respectively. For instance, from Table 2 we have that 27.5% of the test pairs (MC_i, RC_i) have recall and precision level in $[0.91, 1]$, while 96.66% of the tests have precision level in $[0.91, 1]$.

Table 2. Source selection: precision and recall.

		Precision										
		0.00 – 0.10	0.11 – 0.20	0.21 – 0.30	0.31 – 0.40	0.41 – 0.50	0.51 – 0.60	0.61 – 0.70	0.71 – 0.80	0.81 – 0.90	0.91 – 1.00	
Recall	0.00 – 0.10	0.33%	0	0	0	0	0	0	0	0	8%	8.33%
	0.11 – 0.20	0	0	0	0	0	0.16%	0	0	0	5.83%	6%
	0.21 – 0.30	0	0	0	0	0	0	0	0	0	5.83%	5.83%
	0.31 – 0.40	0	0	0	0	0	0	0	0	0	7.5%	7.5%
	0.41 – 0.50	0	0	0	0	0	0.16%	0	0	0.16%	12.16%	12.5%
	0.51 – 0.60	0	0	0	0	0	0.16%	0	0	0	2.5%	2.66%
	0.61 – 0.70	0	0	0	0	0	0	0.16%	0	0	8.66%	8.83%
	0.71 – 0.80	0	0	0	0	0	0	0	0.5%	0.33%	8.83%	9.66%
	0.81 – 0.90	0	0	0	0	0	0	0	0	1.33%	9.83%	11.16%
	0.91 – 1.00	0	0	0	0	0	0	0	0	0	27.5%	27.5%
		0.33%	0	0	0	0	0.5%	0.16%	0.5%	1.83%	96.66%	

The RC effectively improves the performance. Table 3 shows a measure of this improvement. In particular, it compares the average query response times obtained by retrieving the set of documents matching the MC_i with those obtained using the RC_i.

In summary, Table 2 and Table 3 show that there is an high improvement in response time with little loss in the set of records retrieved after automatic source selection.

Table 3. Source selection: average response time.

	T1	T2	Average
MC	162874 ms	186909 ms	174892 ms
RC	48469 ms	52253 ms	50361 ms
Improvement in ms	114405 ms	134655 ms	124530 ms
Improvement in %	70.24%	72.04%	71.20%

4 Conclusions

Since the Web and the information contained in it, is growing rapidly, every day a huge amount of "new" information is electronically published and new Digital Libraries are available to satisfy the user information needs. In this paper, we described a Digital Library that is not only an information resource where users may submit queries to get what they are searching for, but also a personalized, collaborative working and meeting space in which the user functionality may be organized into four categories: users may (i) search for information; (ii) organize the information space (according to the folder and personalized collections paradigm); (iii) collaborate with other users sharing similar interests; and (iv) get recommendations. Particular attention has been paid to the notion of personalized collections, which are user defined un-materialized views of the information space, and how the system automatically determines the most relevant information sources related to the views.

Acknowledgements. This work is funded by the European Community in the context of the CYCLADES project IST-2000-25456, under the Information Societies Technology programme.

References

1. Dewey Decimal Classification. `http://www.oclc.org/dewey`.
2. Christoph Baumgarten. A probabilistic solution to the selection and fusion problem in distribute information retrieval. In *Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 246–256, Berkeley, CA USA, 1999.
3. K. Bollacker, S. Lawrence, and C. L. Giles. A system for automatic personalized tracking of scientific literature on the web. In *The 4th ACM Conference on Digital Libraries*, pages 105–113, New York, 1999. ACM Press.
4. J. P. Callan, Z. Lu, and W. Bruce Croft. Searching distributed collections with inference networks. In E. A. Fox, P. Ingwersen, and R. Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–28, Seattle, Washington, 1995. ACM Press.

5. Jamie Callan. Distributed information retrieval. In W.B. Croft, editor, *Advances in Information Retrieval*, pages 127–150. Kluwer Academic Publishers, Hingham,MA, USA, 2000.
6. Jamie Callan and Margaret Connell. Query-based sampling of text databases. *ACM Transactions on Information Systems*, 19(2):97–130, 2001.
7. M. Di Giacomo, D. Mahoney, J. Bollen, A. Monroy-Hernandez, and C.M. Ruiz Meraz. Mylibrary, a personalization service for digital library environments, 2001.
8. D. Faensen, L. Faulstich, H. Schweppe, A. Hinze, and A. Steidinger. Hermes: a notification service for digital libraries. In *ACM/IEEE Joint Conference on Digital Libraries*, pages 373–380, 2001.
9. L. Fernandez, J. A. Sanchez, and A. Garcia. Mibiblio: personal spaces in a digital library universe. In *ACM DL*, pages 232–233, 2000.
10. P.W. Foltz and S.T. Dumais. Personalized information delivery: an analysis of information filtering methods. *Communications of ACM*, 35(12):51–60, 1992.
11. Edward A. Fox and Gary Marchionini. Digital libraries: Introduction. *Communications of the ACM*, 44(5):30–32, 2001.
12. Norbert Fuhr. A decision-theoretic approach to database selection in networked IR. *ACM Transactions on Information Systems*, 3(17):229–249, 1999.
13. *Information Filtering Resources:* `http://www.enee.umd.edu/medlab/filter`, WWW.
14. A. Moukas. *Amalthaea*: Information discovery and filtering using a multiagent evolving ecosystem. In *Proceedings Practical Applications of Agents and Multiagent Technology*, London, GB, 1996.
15. Allison L. Powell, James C. French, and Jamie Callan. The impact of database selection on distributed searching. In *Proc. of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–239, Athens, Greece, 2000.
16. M. Elena Renda and Umberto Straccia. A personalized collaborative digital library environment. In *5th International Conference on Asian Digital Libraries (ICADL-02)*, number 2555 in Lecture Notes in Computer Science, pages 262–274, Singapore, Republic of Singapore, 2002. Springer-Verlag.
17. L.M. Rocha. Talkmine and the adaptive recommendation project. In *ACM DL*, pages 242–243, 1999.
18. Jian Xu, Yinyan Cao, Ee-Peng Lim, and Wee-Keong Ng. Database selection techniques for routing bibliographic queries. In *Proceedings of the third ACM conference on Digital Libraries*, pages 264–274. ACM Press, 1998.
19. L. A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.