A Federated Digital Library for Universities' Grey-Literature:

Practical Issues and a "Transient" Analysis

Christos Nikolau[1], George F. Georgakopoulos[2]
Computer Science Dep., University of Crete, GREECE

Hariklia Tsalapata[3]
FORTH-ICS, Crete, GREECE

1. Introduction

A ***library*** is an organization consisting of persons, procedures, and tools co-operating towards the following goals: (1) the collection and safe storage of material, (2) the organization of the collected material so that it can be accurately and efficiently located and retrieved, and (3) user-friendly material access mechanisms.

Digital libraries store and manage material in electronic format. Thus, all material processing (collection, indexing, retrieval, and distribution) is performed through electronic and computational tools (software/hardware). Networking capabilities add a crucial dimension to digital libraries: material can be submitted and accessed from remote locations and at any time. Furthermore, digital libraries of related material can be integrated into a single federated system, thus enabling users to locate and retrieve digital and possibly heterogeneous material stored in various locations through a single point of access and in a uniform way.

In this report we present the design and implementation of a digital library of teaching material used in university classes. Typically, universities supply or suggest textbooks as the main reference material. However, in many cases they rely on what we like to call **gray literature**. By this we mean:

1) Class notes (which in many situations may be hand written).

2) Exercises, exercise solutions, and examples/counterexamples.

3) Examination questions.

[1] nikolau@csd.uoc.gr

[2] ggeo@csd.uoc.gr

[3] htsalapa@ics.forth.gr

4) Transparencies, images, and tables with data.

5) Bibliography and references.

6) Student projects, theses etc.

7) Computer programs.

8) Seminars.

Gray literature is essential for teaching almost 50% of the courses given in many universities in Greece, and this is probably the case in universities of several countries in which English is not the official teaching language. Material of this type has an unofficial status and is usually lost since it is not collected and indexed properly. Student projects are forgotten in various unexplored disk directories. Computer programs remain unpublished and undocumented. Class notes, transparencies, and images are stored in office-drawers. Exercises and their solutions are passed from student to student, constituting perhaps the grayest of gray literature.

Libraries of traditionally published material fail to incorporate the wealth of information presented in gray literature. It becomes apparent that the collection and organization of gray material will complement traditional libraries with valuable information. The printing cost, the availability (or lack of) of material, the variety of storage formats, and the continuous material modifications are some of the considerations leading to the conclusion that a digital library may be the only viable solution for the effective management of gray literature.

Furthermore, the electronic publication of gray literature through a digital library provides the material with a more “official” status. This has the following advantages. First, it creates pressure for high quality. For example, a lecturer may think that small errors in hand written class notes can be corrected during the lectures. The same lecturer would most probably hesitate to offer material for distribution through the Internet without a final proofreading. On the other hand, a digital library enables easy access to material by a large number of users, thus providing recognition to authors as well as an incentive for new publications. Another advantage of digital libraries is that the publication of material on similar subjects by a number of authors provides students with more complete information while at the same time it promotes co-operation between universities and prevents duplication of effort.

The special characteristics of gray literature influenced significantly the way we conceived our project and our implementation strategy. Some of these characteristics are discussed below:

1) "Gray literature" material is *dynamic*: class-notes and exercises are frequently reviewed, new figures are produced, programs become obsolete and are replaced by new versions, etc. Digital management of gray literature material enables efficient revision of material thus contributing to up to date information.

2) "Gray literature" is not printed in official copies that can be considered library "property". Material is not “owned” by the library, it is simply offered for use. Authors may withdraw their material at any time.

3) Authors participate actively in the indexing of their material by providing a correspondence between keywords and pages/sections. We cannot imagine a more effective way for correct indexing of gray literature material.

The above characteristics reveal that a digital library of gray literature is a lot more than a simple collection of documents. It has a lot of similarities with an **electronic publishing house.** To turn the idea of an electronic publishing house into a practically working everyday procedure we had to address the difficult task of orchestrating about 500 lecturers and professors. In the following sections we present our strategy and the tools we have been using during the past one and a half year to construct a federated digital library of gray literature among six Greek universities[4]. We keep both our feet on the ground and report for the sake of the reader the actual solutions we have been applying and seen working in practice.

Our project is named "S.K.E.P.S.IS.", an acronym which is also the Greek word for "thought".

[4] "S.K.E.P.S.IS." was financed by the Greek Ministry of Education through supporting funds from the European Union.

2. *Building a digital library from scratch*

To implement our library we had to concentrate on the following key tasks:

1) Orchestrate the participation of 500 lectures and professors from 14 departments of the University of Crete. Lecturers and professors from 5 more Greek universities participate in the project as well.

2) Convince professors to produce and distribute material in digital form.

3) Support participation of teaching personnel with various degrees of familiarity with computers.

4) Convert the entire procedure from a feasibility project into a systematic ***service*** offered by the University of Crete.

The above tasks had to be implemented in less than 18 months. To ensure success of the project given this time constraint we adopted a *stepwise* implementation strategy (for this reason in the following we refer to "transient" analysis). The number of implementation steps had to be kept small. On the other hand, the gap between them should be bridgeable. We spent a significant amount of time reviewing our approach. We believe that our persistence in the design of the implementation strategy was rewarded by the results of the project. In the following sections we describe in detail the implementation stages that in our opinion leaded to a successful digital library service offered by the University of Crete.

(2.1.) Acquiring material and converting it to digital form.

The first issue that we had to address, and perhaps one of the most challenging ones, was communication with about 500 lecturers and professors located in four buildings, in two cities. The professors had to be informed about our effort, and their trust and co-operation had to be gained. We started with public presentations, e-mails, and informative documents. However, the most powerful communication method proved to be direct person-to-person communication. We appointed a representative in each department that was responsible for collecting department course material. In fact, a substantial amount of the available budget was reserved for the compensation of the department representatives. We provided the representatives with sufficient details on our project and asked them to engage in similar personal discussions with their colleagues. The common language shared between the department representatives and their colleagues as well as the frequent follow up meetings between us and the representatives proved productive: a substantial amount of gray material was collected in only a few months.

Most of the material was either hand written or was available partly in hard copies and partly in electronic form. We insisted in a "*give-as-it-is*" policy to ensure participation of professors and lecturers that have little familiarity with digital documents. To convert the material into digital form we used several processing stages:

(a-stage) Scanning of manuscripts

(b-stage) Re-typing (to convert text into digital form)

(c-stage) Re-drawing schemata (to convert all material into digital form).

To ensure high quality of scanned material pages were scanned in 300dpi resolution and gray-scale. This high level of resolution requires significant storage space. Furthermore, the size of the generated files is prohibitive for network transfers. For the above reasons, scanned material is stored on CD-ROMs for backup purposes. After being backed up, the material is transferred to the digital library. There the material is converted to "*.pdf* " format. The internal compression of this standard reduces the size of the scanned material significantly, thus allowing the resulting files to be downloaded in an acceptable amount of time. Furthermore, the *".pdf"* formatted files can be viewed on-line through *Acrobat Reader*®, a user friendly, practical, cross-platform, ground-gaining and free browser/viewer.

Some material was already available at (b-stage) form, that is, text was available in digital form. However, only a small number of professors provided material in fully digital form (c-stage). Naturally, these were professors whose disciplines allow familiarity with computational techniques: physics, economic studies, and, of course, computer science. Having established a process for acquiring gray material the next step was to design an *indexing* method.

(2.2) Indexing material.

Indexing requires the establishment of a correspondence from keywords to place holders in the document. Even when a catalogue of keywords is available, it is quite time-consuming to create this correspondence. Frequently the result is not so useful: for example, not much is gained when a word is linked to a page of a manuscript soon to be converted to digital form. Moreover, and needles to say, hard copies cannot be indexed automatically by the commonly used computational tools. A correspondence between keywords and placeholders in a document was more than we could ask from professors to contribute.

We selected four stages of detail for mappings between keywords and placeholders:

(a-stage) Mapping of keywords to an entire document. This is a very rough solution but allowed us to link a document to our collections very easily. It is performed through a search engine and is discussed in further detail in later sections.

(b-stage) Mapping of keywords to chapters or sections of a document. These are few in number and easily identifiable. Moreover, the map can be based on the table of contents that most authors do provide.

(c-stage) Mapping of keywords to pages of a document. Professors often provide this correspondence for documents that use a lot of terminology.

(d-stage) Mapping of keywords to the exact points of their occurrence in a document. We consider this last stage to be of lesser importance since gray literature documents are usually browsed page by page.

Having a way to index our material our next task was to design a user-friendly access mechanism:

(2.3) Accessing the stored material.

The organization of digital material so that it can be easily located is not a trivial task. As the saying goes "a computer does not give you what you want, only what you request". In our system we provide three methods for locating information:

1. *A tree-like classification:*

Classes are organized in a tree-like hierarchical structure that includes three levels: the 1st level includes universities participating in our federated library. The 2nd level includes the participating departments for each university. The 3rd level includes courses offered by each department. Navigation of the hierarchy is performed through a tree-based user interface where users explore available information by expanding and collapsing relevant tree nodes.

It is worth mentioning that the initial web pages of our digital library software (which is heavily based on the Dienst system developed at Cornell University [3], [4]) adopt a simple yet functional design. The Human-Machine Interfacing group of FORTH-ICS designed a more elaborate user interface for our system.

2. *Keyword searching:*

Naturally we provide the usual method of keyword searching. Our material is described through a set of metadata search fields. The system search engine matches keywords against pre-built indexes on the metadata to locate information of interest. Our ontology is simple: material is organized in "courses". The material of each course in our system can be retrieved through a set a web pages: a home page for the course and additional pages referencing subsections such as class-notes, exercises, projects, bibliography, etc.

3. *A thematic index:*

Keyword search introduces two interesting issues: a) Users initiating a query must “guess” a keyword that accurately describes the information they are looking for. A search engine can aid users in keyword selection. This can reduce significantly the number of queries that return no results due to inaccurate keyword selection. b) Professors and users may describe the same material through different sets of keywords.

We decided to build a vocabulary of terms for our system and offer our users the possibility of employing keywords from this vocabulary in the queries. We consider this to be a first step in addressing accurate keyword selection. A keyword list for university courses can be based both on the department course guides that include short descriptions of offered courses and on the indexes available in course textbooks. We decided to proceed in three stages:

(a-stage) As a first step we gathered a list of essential terms from the short descriptions of courses included in course guides. The guides are easily accessible both by students

and us. Furthermore, their use prevents contradictory descriptions and relieves authors from the task of providing keyword lists for their material.

(b-stage) As a next step we used the gathered terms to build a vocabulary of keywords. This vocabulary will be available "on-line" to both authors and readers as an aid to query keyword selection. In a second phase this vocabulary will be enriched with words from textbooks' indexes.

(c-stage) As a last step we organized this vocabulary into a thesaurus. We consider this transformation to be an important issue. We think of a "thesaurus" as a vocabulary in which words/terms are interrelated in two ways: (1) by "equivalence" relations (2) by "generalization" relations.

The equivalence relation introduces two issues. First, material may be characterized through the first of two equivalent words and be accessible only through the second. The second issue is the support of multilinguality: equivalent terms may be available in more than one language.

The generalization relation offers a very useful capability: through a single query users can locate material that matches the keywords they entered as well as material described by more “general” (or more “specific”) terms. For example, if we declare in our thesaurus that the word "economy" is more general than the word "money" then a user employing the keyword *"economy"+"more specific"* will be able to locate material characterized either by the word "economy" or by the word "money".

The Information Systems group of FORTH-ICS has developed a set of tools for building, browsing, and merging thesauri. The system is called Thesaurus Management System - Semantic Information System [5], and uses of client-server architecture. Tools for linking this system to Dienst have also been developed.

3. The system.

3.1. System architecture overview.

The SKEPSIS system uses a distributed architecture to enable single point of access, retrieval, and presentation of heterogeneous material possibly located on remote servers.

The architecture is based on the Dienst digital library software developed at Cornell University.

SKEPSIS is comprised of a number of autonomous server nodes, each managing one or more underlying repositories. A SKEPSIS server consists of three services, the indexer, the user interface, and the repository service:

The indexer service locates information of interest to the user through pre-built indexes on metadata descriptions of the objects available in the underlying repositories. The indexer builds one index for each metadata field off-line. At querying time, the indexer matches keywords submitted by users against the pre-built indexes to determine availability of relevant information.

The user interface service is the intermediary between the user and the system. It receives user queries and presents search results. Through the user interface, users can select the repositories to be queried. The user interface service propagates queries to all selected repositories in parallel, collects the query hits from each repository, and presents the merged results to the user.

The repository service retrieves digital objects in a format that is suitable to the user needs. In a distributed digital library like SKEPSIS, each object may be available in a variety of formats, e.g. postscript, PDF, TIFF, HTML, etc. SKEPSIS supports several commonly used storage formats and is easily extensible to support additional formats.

In addition to the above three services that are available in each SKEPSIS server node, the distributed SKEPSIS architecture includes a single, centralized metaserver service. The metaserver provides the autonomous SKEPSIS server nodes with information about the nodes, services, and repositories available in the distributed system.

Users contact and query the SKEPSIS system through a web browser. They can submit simple queries, in which case SKEPSIS matches query keywords against all metadata field indexes. In addition, users can submit fielded searches by specifying keywords to be matched against specific metadata field indexes through an HTML query form. As a result to their queries, users receive a list of references to matched objects. Each reference includes a description of the object, which is dynamically retrieved from the object metadata. In addition, it includes links through which the object can be retrieved in one of the available storage formats. Furthermore, when reviewing a retrieved object, users can obtain additional "similar" objects available in the system effortlessly by clicking on a button. Finally, users can browse the system collections through a tree-structured graphical interface that navigates the hierarchy of available repositories and documents.

3.2. Query support for heterogeneous repositories

SKEPSIS supports the definition and indexing of arbitrary metadata field sets [1]. This allows participating organizations to describe the material they are contributing through customized metadata definitions; each repository/collection in SKEPSIS may use a separate metadata definition. Furthermore, metadata fields are typically used in a variety of ways. Certain fields, e.g. author and title, are typically used in queries for locating information, while other fields, e.g. the location through which the object can be downloaded, are more frequently presented to the user as information on an object. For this reason, SKEPSIS supports the indexing of a subset of the fields of a metadata definition.

Distributed queries on repositories whose objects are described by different metadata definitions are supported on the intersection of the metadata field sets of the repositories. The SKEPSIS user interface dynamically computes the intersection of the metadata field sets of the repositories selected by the user performing a distributed query and builds a customized HTML query form, thus alleviating the user of the responsibility of identifying the metadata fields of the SKEPSIS repositories. In addition, SKEPSIS supports queries on the union of all metadata field sets.

While the SKEPSIS architecture supports the definition and indexing of any metadata field set, a metadata set tailored to university gray literature has been defined for the purposes of the SKEPSIS digital library. The metadata set focuses on the description of the thematic area and the content of a university class as well as the class subsections (e.g. exercises, projects, papers, exam questions, and computer programs).

3.3. Multilinguality support

Material published in a federated digital library may be available in a variety of languages. Furthermore, users of a digital library may be of different nationalities. Both material publishers and users prefer to use the system in their native language. For the above reasons, it is important for a digital library system to include multilinguality capabilities. SKEPSIS supports multilinguality both at the user interface and at the indexer level [2].

At the user interface level, SKEPSIS supports multilinguality by providing an HTML interface in a variety of languages. Users select the language they prefer for communicating with the system through a specific menu available in the system home page. The language of the text for all subsequently used HTML pages is determined from the initial language selection.

The SKEPSIS indexer is easily configurable to support indexing and querying of strings in any ISO character set. Furthermore, if metadata is available in more than one language/character set, SKEPSIS indexes the metadata in all languages it is available in. Thus, users can locate information by querying in several languages. Typically users prefer to query the system in their native language. However, in certain situations terminology may be more expressive in a language other than the user's native, as for example may be the case with computer science terms.

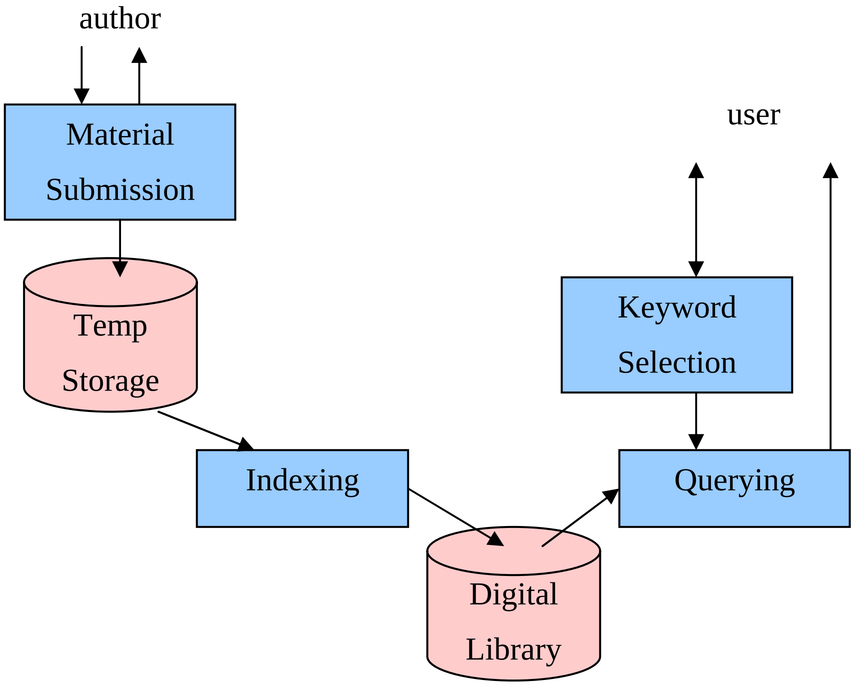

Figure 1. Overview of SKEPSIS procedures.

The Distributed and Parallel System group of FORTH-ICS implemented the extensions to the Dienst software discussed above.

4. Epilogue

We have presented our experience on building a federated digital library for gray literature among six Greek universities. Our approach goes one step further from simply collecting material to setting up the procedures required for the smooth operation of an electronic publishing house. Our presentation focused on the description of the "transient" process we followed in order to take our library from non-existence to a degree of development that is sufficient for the initiation of a gray literature digital library service offered by the University of Crete. Our overall experience is very encouraging for the future success of the service.

Digital libraries are expected to flourish in the near future. However, it is difficult to predict the exact areas that digital libraries will cover as well as where, when, or how they will present their full power and flexibility. For example, will digital libraries cover the full extend of university teaching material or will they be limited to an essential part of it? Time will tell. For the moment we are pleased by the fact that we have paved a way towards it - a way as smooth and efficient as was possible for the human and capital resources we had available.

References

1. S. Kapidakis, P. Alexakos, H. Tsalapata, “Parameterization of Dienst metadata search fields”, FORTH-ICS TR99-0249, March 1999.
2. S. Kapidakis, I. Mavroidis, H. Tsalapata, “Multilingual extensions to Dienst”, FORTH-ICS TR99-0248, March 1999.
3. C. Lagoze, E. Shaw, J. Davis, D.B. Kraft, Cornell University TR95-1514, May 05, 1995.
4. J. Davis, C. Lagoze, “A protocol and server for a distributed digital technical report library”, Cornell University TR 94-1418, June 24, 1994.

5. M. Doerr and I. Fundulaki, "SIS - TMS: A Thesaurus Management System for Distributed Digital Collections", Proc. 2nd European Conference, ECDL'98, September 1998, Heraklion, Crete, Greece.