A distributed digital text accessing and acquisition system

Adolfo Guzmán Arenas, Victor-Polo de Gyves

SoftwarePro International [1]
a.guzman@acm.org

Abstract. BiblioDigital ® is a network of reservoirs (R) of text documents. Each document exists *primarily* in one R, with possible duplicates in other Rs. Each R sits in its own server. Each document in indexed in three ways: * by themes (vocabulary controlled by that R's librarian); * by *each word* in the document, * by the *concepts* which the document covers (using Clasitex ®). Each R contains the *global index* (of all Rs), so that each R can provide the following services: * browsing by themes; * by concepts; * by words; * by metadata; * by Boolean combination of above. Also, BiblioDigital * allows subscription to a personal News Services: through a user interest profile; * BiblioDigital combs the Web for documents that could fall in the themes or topics contained in its indices, and indexes them, thus enriching its knowledge content.

KEY WORDS: Digital library, distributed, concept classification, crawler.

1. Introduction

BiblioDigital, a distributed collection of reservoirs (R) containing full text documents is described. The system is already implemented and some small examples are given.

1.1 Executive summary

In addition to what the summary explains, other important features of BiblioDigital®:

- A reader can, through *any* R, have access to *all* its documents;
- A *librarian* (owner of an R) registers *authors;* readers (users) do not need to register; documents are primarily free and without encription;
- It allows document *versions,* auxiliary documents (tests, software…);
- Subsumes (absorbs full texts, and/or just indexes them) documents sitting in foreign libraries, thus allowing its full exploitation;
- It uses meta data (example: Dublin Core), if this option is on;
- Multimedia documents can be indexed, if they contain a text description;
- It handles documents in popular formats (plain text, PDF, Word, Excel...);

[11] BiblioDigital ® is property of SoftwarePro International. Adolfo Guzman is a researcher at CIC-IPN.

- Allows each librarian to have his own taxonomy of themes, and also uses its own global ontology of concepts imposed by Clasitex ®);
- Each R has a *cache* of frequent documents.

Features of a (yet to exist) second version of BiblioDigital®:

- Fault tolerance; damaged document correction;
- Servers can "get in" or "get out" of the mesh of Rs, *a la* peer-to-peer, without a root node (to be explained below);
- The global index will be distributed when too large for a single server.

1.2 Comparison to previous work

The field of digital libraries has made much progress; an early but still influential collection of articles is [1].

Most of the features of BiblioDigital can be found in other systems; it is the unique mixture of them, coming from the experience of the builders and some users in effective uses of text documents, that make BiblioDigital unique. Another unique feature of BiblioDigital is its use of Clasitex ® [3] to classify a document in the themes it talks about.

Single-server (not distributed) digital libraries are useful; in Mexico, Phronesis [2] is popular.

Federated libraries, or federated search, is handled often [4] by converting an initial user query in semantically equivalent queries expressed in other dialects, that "the other" libraries can answer directly. BiblioDigital does not use this approach, instead, it " milks" each document of a foreign library (§2.8) and indexes it, keeping the document in its original library.

To keep a replica of the global index (§3.2) in each R is a simplification of the peer-to-peer protocol, which we felt too complex to be of use now. In attention to the growth of the global index, tables are kept in R to migrate later to a more advanced distribution of the index, in which each server has only part of the total global index.

The mail service of personalized information (§2.4) according to a user profile is hardly new, but its use in digital libraries is somewhat of a novelty. Another novelty in digital libraries seem to be the *collections* introduced in §2.5.

Advanced search services (§2.7) can be found in some systems, like Amazon's book store; in previous experiences, we have found them quite useful, so that their implementation is coming (§3.5).

The handling of video files in BiblioDigital is possible, but due to limits in band-width, is not sponsored. Instead, the architecture in [5] is more appropriate this purpose.

2. Description

BiblioDigital® is a confederation of independent similar libraries, linked by a global index.

A node of BiblioDigital®, to be called R (reservoir) is a physical place (a computer) where text and image electronic documents are stored in an organized way, to be provided to users, which can access them through any computer connected to Internet.
The manager of an R is its *librarian*: he registers *authors*, *collections* and their *editors*; he defines the taxonomy of themes in that R. *Readers* need not be registered in order to use BiblioDigital.
Rs form a tree: each R (except the root, called *Adam*) has a parent R. Each document and each collection sits in (*belongs to*) exactly on a R. Each R lies in a PC (a *server* of BiblioDigital) with enough disk, no-break, antivirus… See figure 1.

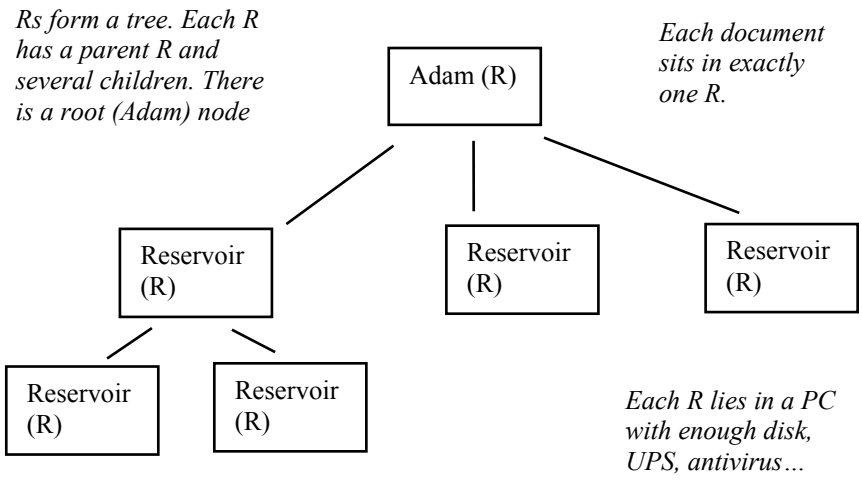

Fig. 1. BiblioDigital is a tree of physical reservoirs (R) holding electronic documents. Rs share a global index that is updated every night.

A reader, connected to any R, can access all documents in BiblioDigital, not only those of the R to which he is connected.
An author can (a) add new documents to the R to which he belongs; (b) update his documents; (c) add supplemental documents to (primary) documents previously entered into R. An editor of a collection can add (links to) documents to it, and update its *status*. Adding copyrighted material by an author can be illegal or punishable; a warning is posted at the upload window.

2.1 Access to a document

By theme. The thematic structure or taxonomy of an R is defined by its librarian. Each author classifies his document into one or more of those predefined themes (controlled vocabulary), including the theme "others."
By concept. The structure (ontology) of concepts is given by the system, which [automatically] classifies (using Clasitex ®) the document in the concepts covered by it.

By the words and special phrases ("In God we trust") in it. The structure is an alphabetical list; classification is automatic (by the system).

2.2 Browsing the documents

Two ways to browse (Figure 2) BiblioDigital's documents and collections:

- A reader (through any R) sees the tree of themes and navigates up and down the tree. He can see the titles in each theme, the summaries of the titles, and the full text of any document.

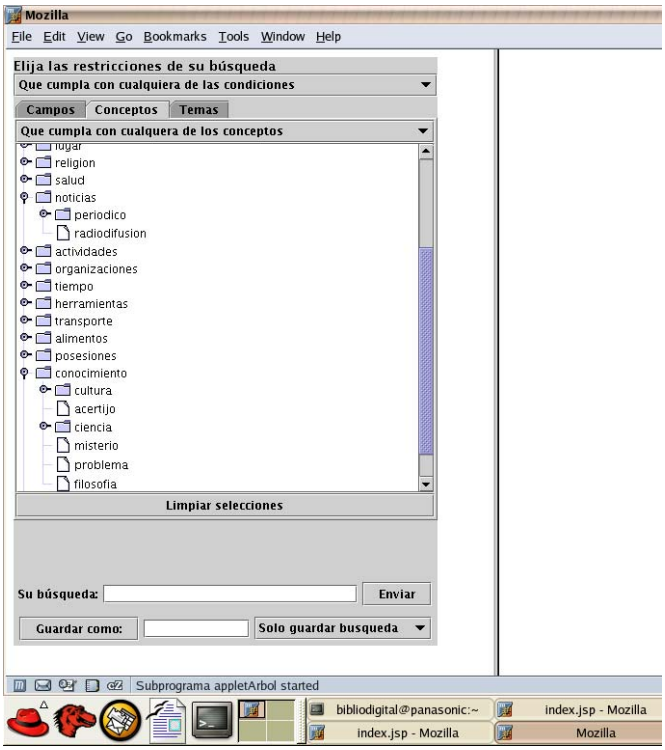

Fig. 2. The thematic tree (or the concept tree of Clasitex) appears to the left. A node can be opened to show its children. To the right, documents in the selected node appear: a title and a summary (metadata) for each. A click with the mouse brings the full text. Documents can be read, printed or copied.

- The same can be done using the concept tree. He can access the concept "England," expand it to go down to the concept "London," go up to "Europe"…

2.3 Search

Searching in BiblioDigital brings documents fulfilling some property (a Boolean expression) given by the reader (user).
Simple search. "Give me all documents about this AND that theme." "And that lie in certain Rs." "Of a given author." "Which talk about this OR that concept."
Complex search. "Having 'Irak' near 'invasion' (in the same paragraph).
The same Boolean expression can contain conditions about themes, concepts, words, special phrases and metadata (author, language of the document; date, type…).
A search can be stored. In fact, the system automatically stores the last 10 searches.

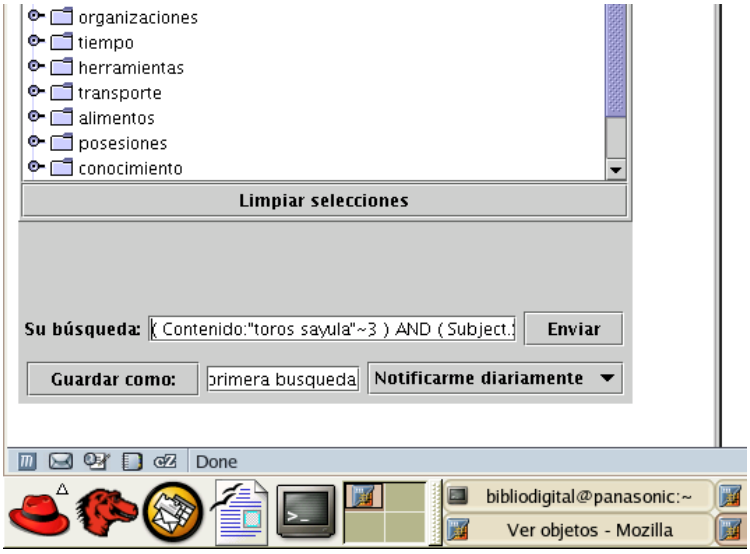

Fig. 3. A registered reader can ask for a periodic news mail service, indicating a profile of interest.

2.4 Subscription to a personalized news mail service

A reader can indicate his profile of interest, through a predicate (a query) containing themes, concepts, keywords and metadata. Then, the system will send him periodically (daily, weekly…) an email, a personalized news, containing the new titles and summaries matching his profile. For this, the reader needs to be a registered reader.
Documents in R are available to users of that R since the instance of its upload, and to readers of BiblioDigital (globally) in the next day. Updating the global index occurs every night.

2.5 Collections

The librarian of an R can register a new *collection* of documents, in charge of an *editor* (who needs to register with the Librarian as editor of said collection). In this manner, BiblioDigital can handle, for instance, digital journals. Each collection lies in exactly one R. A collection really contains pointers to documents already in BiblioDigital (in any R). A document in a collection can be in one of several *states,* defined by the *editor* at collection creation time. Example: received, in revision, accepted, rejected, accepted with minor changes…. Manually, the editor changes the state of a document when he so decides. In a future version, certain agents (e-mail, for instance) may trigger the transitions; thus, *collections* are a hook for workflow software. A document can belong to 0, 1 or more collections.

2.6 More on a document

A document (called now the main document.) can have: (a) versions; (b) associated or auxiliary documents: exercises, slides, solutions, additional examples, software…
Metadata. Each document has a small table (metadata) describing it: autor, date, language… which the author fills, with some fields pre-filled by R. Currently, we use Dublin Core.

2.7 Advanced search or Markov search

These are based on the dynamics of the reader, as he jumps from a set of documents to the next, or from document to document. Examples. "I offer you documents similar (in concepts content) to those you have been reading." "I offer you documents that other readers with your same dynamic reading path have been reading." "70% of readers that read document A and then B also read document C; here is C." "Give me all the documents read last month by Carlos Fuentes." "Or by the Engineers Association." "And about the NAFTA agreement."
Some of these searches, although technically possible, are not available since they go against the privacy of readers.

2.8 Access to other existing digital libraries

Documents in existing digital libraries can be indexed and served (shown) by BiblioDigital:

- if they possess metadata, by the items in such metadata;
- In any event, by concepts and by word content;
- If they have a summary, it will be used by BiblioDigital.

Each document can be shown to the user by calling the original software (that is, calling the *other* digital library). The foreing documents are kept at their original site; BiblioDigital does not acquire (import) their full text. See figure 4. It is also possible to import those documents in full text, duplicating them.

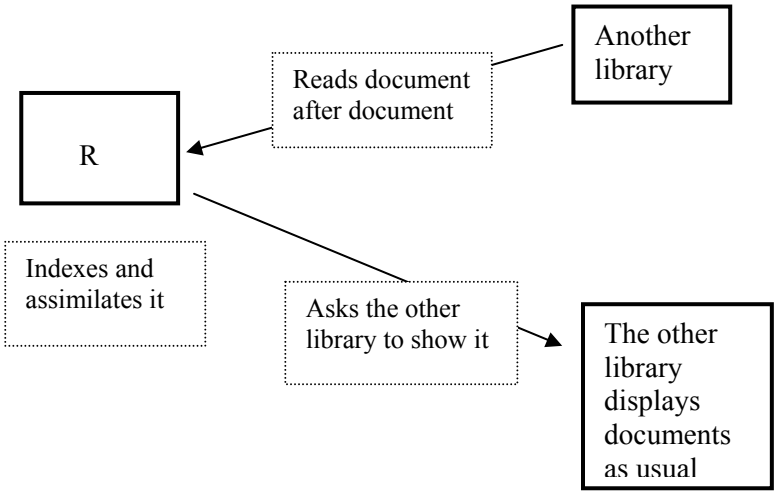

Fig. 4. From BiblioDigital it is possible to access and serve documents in other existing digital libraries, if these provide the two APIs shown.

2.9 Cache of recent documents

Automatically, BiblioDigital keeps in the disk of the local server R a cache area with the recent documents most used by readers logged in that R. This area is updated automatically. This increases access speed to those documents.

2.10 Modifying the taxonomies

Changes to the taxonomy of themes defined by a librarian for his R are infrequently allowed. To add new themes (nodes) initially empty is no problem, except that some documents belonging to the parent (of the new node) now belong more appropriately to the new node (the new son). So, these documents must be "moved down" by the author from the father to the new son. This provokes additional work for the authors, which they will tolerate if infrequent.
In general, repositioning the documents from an outgoing (old) part of the thematic taxonomy to the incoming (new) corresponding part, is done by the librarian as follows:

1. The old part of R (with all its documents) is brought down ("erased.")
2. The new part of R (new part of the taxonomy) is brought up ("created"), initially empty.
3. Each author takes his documents deleted in (1) and adds them to (some of) the new nodes of (2),

Notice that no changes are needed to the concept taxonomy. With respect to the Full word index, the librarian can introduce more "stop words;" that is, words that should not be indexed.

2.11 Modifying the themes of a document

A document may belong to several themes, as defined by the author at upload ("entering") time. The author can change his mind and reposition the document in the nodes of the thematic taxonomy. For this, the author brings down (erases) his document and brings up (enters) the same document, taking due care to index it in the new themes. It is an intentionally awkward procedure.

2.12 Protection against inexperienced librarians

Some frequent errors of librarians and how BiblioDigital copes with them:
Frequent changes to the thematic taxonomy. They will not be possible, since this is an intentionally painful procedure. Cf. §2.10.
To register an excessive number of authors. This may be allowed up to disk capacity, or there could be limits imposed by BiblioDigital (none at present).
Badly constructed taxonomies, where the grandfather is brother of the grandson. There are guides, some of them accessible inside BiblioDigital, about how to construct good (solid, sound) taxonomies. If the librarian produces a bad taxonomy, it is his responsibility. BiblioDigital makes no further checking or advising.

2.13 Protection against inexperienced authors

Some frequent errors of authors and how BiblioDigital copes with them:
Controllable by the librarian:

- An author uploads pornographic or irrelevant documents (music, pictures). This can be tolerated or prohibited by the librarian.
- An author enters too many texts. The librarian can set a limit in number of documents, or in megabytes.
- An author assigns the wrong themes to his document. Fixable by the author.

An author assigns to a document of him themes that do not exist in the thematic taxonomy. This is impossible, since the themes are selected from a menu. The only "new" theme is the theme "others." Also, an author can request from his librarian the addition of a new theme to the thematic taxonomy of their R.

2.14 Several authors write a document

This is simple in BiblioDigital:

- The librarian defines one of the authors as the editor of a (new) collection.

- An author writes his parts and sends them to that collection in R. He also sends comments and criticisms to the other parts,
- The editor accepts, rejects or modifies parts and criticisms.
- When finished, the editor erases the collection and creates (enters) the document as a new document. Or the final form of the document is kept in the collection.

3. Handling foreign documents

No matter how many documents can sit in all Rs, there always be more documents *outside* (in the Web). To tap these foreign riches, BiblioDigital reads and indexes (by concepts, and by words) the documents "outside BiblioDigital."
For this, librarians provide a set of sites (URLs) where there exist indexable documents. BiblioDigital divides this set into subsets, one for each R. The crawler of each R will search Web pages in each subset for suitable documents, to be added to that R (the document is not imported into R, but is kept in its original site).
To avoid work duplication (an spider or crawler accesses node NSF, and another spider is doing the same), there is a procedure where these crawlers share and synchronize themselves for time to time, avoiding overlap in search.

3.1 Other documents: audio, images

They can be indexed, as long as they have metadata or a written (text) description or introduction. Only certain kinds of formats can be stored in BiblioDigital: (TXT, HTML, XML, PDF, PS, DOC, MPG).

3.2 High performance

- More than 100 queries/second (with 5 servers)
- The themes, concepts and words *are already indexed.*
- Each R has the total index and *all* summaries of every document (of all Rs).
- Normally, a user connects to an R of themes intesting to him:
  - A physician connect to the medical R;
  - This diminishes traffic between Rs;
- Automatic caching of frequently read documents; a cache for each R;
- I can order a query the night before.

3.3 Module to manage taxonomies

BiblioDigital comes with an editor for the librarian to arm and maintain his taxonomy: update, add and delete nodes. Every change to a taxonomy already in site (active, with documents) will affect the indices and introduce re-indexing, much of this of manual nature. It also comes with a manual of "good manners to form a taxonomy."

Recommendation: think and test a taxonomy before enabling it.

3.4 Mail from readers, to authors, editors…

BiblioDigital allows a reader to send a document to a friends. There is also communication with the autor, librarian and editor of a collection. Also, a reader can add small comments to an article that he has read. An author, editor or librarian can add in BiblioDigital a pointer to his Web page.

3.5 Status

Version 1 is running since January 2004, it is a development of the authors for SoftwarePro International. More information at: a.guzman@acm.org Version 1 also handles audio files, as well as it monitors the principal news in (electronic) newspapers of national coverage. Version 2 will have the features of §§2.7-2.9

References

1. Computer, Vol. **32,** number Two. Feb. 1999. Digital Libraries. IEEE Computer Society.
2. David A. Garza-Salazar, Juan C. Lavariega, Martha Sordia-Salinas. Information Retrieval and Administration of Distributed Documents in Internet. The Phronesis Digital Library Project, in *Knowledge Based Information Retrieval and Filtering from Internet*, Kluwer Academic Publishers, Boston, MA. 2003
3. Guzman, A. Finding the main themes in a Spanish document. (1998) *Journal Expert Systems with Applications,* Vol. **14,** No. 1/2, 139-148, Jan./Feb Phronesis
4. Bruce Schatz, William Mischo *et al.* Federated search of scientific literature. In [1], pages 51-59.
5. Howard. D. Watclar, Michael G. Christel *et al.* Lessons learned from building a terabyte digital video library. In [1], pages 66-73.