

CERN Document Server Software: the integrated digital library

A. Pepe, T. Baron, M. Gracco, J-Y. Le Meur, N. Robinson, T. Simko, M. Vesely

CERN

CH-1211, Geneva 23. Switzerland

{alberto.pepe,thomas.baron,maja.gracco,jean-yves.le.meur,nicholas.robinson,tibor.simko,martin.vesely}@cern.ch

Abstract

CERN as the international European Organization for Nuclear Research has been involved since its early beginnings with the open dissemination of scientific results. The dissemination started by free paper distribution of preprints by CERN Library and continued electronically via FTP bulletin boards, the World Wide Web to the current OAI-compliant CERN Document Server.

CERN Document Server Software (CDSware) is a suite of applications which provides the framework and tools for building and managing an autonomous digital library server. In this paper, we discuss the design philosophy of CDSware and its modular, extensible, architecture. Each module comes as an independent entity embodying a specific aspect of digital library workflow.

By means of a flow-chart we present the operational workflow of the system, depicting its module interactions. Hence, some of the key features in the CDSware technology are introduced more in detail, namely metadata representation, acquisition and delivery, indexing and ranking techniques, user interface and personalization.

CDSware uses entirely freeware technology and it is available under the terms of the GNU General Public License. It is developed by the CERN Document Server team and is driven and validated by the CERN Scientific Information Service. In addition, CDSware has been installed and is in use by over a dozen institutions around the world.

A brief comparison with other existing free digital repository systems will also be made.

1. Introduction

1.1 Towards new digital library systems?

The main problem facing the Open Access (OA) movement [Suber2005] is the speed of feeding the existing repositories, either via OA publishers or OA institutional repositories. After ten years of effort, only 20% of peer reviewed articles are OA today [Harnad2005].

We consider that one of the most natural ways to increase open access to (scientific) information is to have libraries more involved in this process. They are traditionally mandated to keep and maintain the institutional production and to provide access to literature of interest for each institution. They are in the best position to open the gates for massive drive to OA. With initiatives like the recent Google Print Library project [GooglePrint], and a similar plan from the French National Library [Gurrey2005], books in the public domain are to become open access world wide. Referred articles, published either in periodicals or in conference proceedings, will undoubtedly follow the same road to OA.

In the area of particle physics, the necessity to share information between institutions worldwide led to the birth of the world wide web; at present, solutions for large scale OA systems are being challenged. The CERN Document Server software package is the result of ten years of organic growth aiming at merging the best of traditional library systems and the best of modern information retrieval technology. Driven and validated by users and librarians, CDSware has grown into a large software suite intended to cope with large collections (almost 1 million records at CERN), and with advanced library-type functionalities.

1.2 CDSware history

In 1993, the CERN Preprint Server started its life on the Web, aiming at collecting and disseminating all high-energy physics and related research documents. It was mostly used as an 'institutional repository', with two original collections, the CERN preprints and a SCAN series that was composed of physics papers received from the whole world and scanned by the CERN library.

In 1996, it became the CERN Library server (weplib), using the same software to provide access to periodicals, books and most of the material kept in the library.

In 2000, multimedia data, like photos, posters, brochures and videos produced at CERN were integrated in a new version of the application, called the CERN Document Server Software: CDSware. This package was made OAI-compliant and distributed in many places. It also started to be used in 2004 as a document management system by the CERN Directorate to handle all the incoming and outgoing documents passing by directorate offices.

Presently, the CDSware package can be used either as a general document management solution, a library system or an institutional repository. New developments are carried out through a partnership between CERN and EPFL (École Polytechnique Fédérale de Lausanne), and the software is regularly enriched with patches received from external contributors.

In parallel with the production of a digital library system, the CDSware team is also releasing a digital conferencing application, Indico [EUIndico], funded by the European Commission, and also OAI-compliant. CDSware is now becoming a suite, containing all necessary software to set up an efficient computing environment for digital library and conferencing. This paper only focuses on the library system assets but you can find more about the conferencing side at [LeMeur2004].

2. CDSware overview

The CERN Document Server Software is a suite of applications which provides the framework and tools for building and managing an autonomous digital library server. The software is readily available to anyone, as it is free software, licensed under GNU General Public Licence. The technology offered by the software covers all aspects of digital library management. Its flexibility and performance make it a comprehensive solution for the management of document repositories of moderate to large size. At CERN, CDSware manages over 500 collections of data, consisting of over 800,000 bibliographic records, covering preprints, articles, books, journals, photographs, and more. Besides CERN, CDSware is currently installed and in use by over a dozen scientific institutions worldwide [CDSDemo].

The software has undergone a constant incremental growth which has taken it from an early basic digital server to the current high-end repository system. Despite its increased complexity, CDSware has retained high performance, user friendliness and a high degree of customization, by enforcing compliance to an established modular architecture. From a technical point of view, CDSware runs on GNU/Unix systems on top of MySQL database server and Apache/Python web application server. The compile-time configuration is accomplished via GNU Autoconf and WML and the runtime configuration is done via MySQL configuration tables. The software is almost entirely written in the Python programming language, with some *ad hoc* modules and functionalities developed in PHP or Common Lisp.

The key feature of CDSware's architecture lies in its modular logic. Each module embodies a specific, defined, functionality of the digital library system. Modules interact with other modules, the database and the interface layers. A module's logic, operation and interoperability are extensible and customizable. A genuine overview of the software architecture can be grasped by looking at the diagram of Figure 1.

The diagram shows a top-to-bottom pictorial representation of the entire document workflow of CDSware. At the top, data acquisition is performed from three different sources: direct author submission (using email or the web interface), OAI and non-OAI harvesting. The metadata gathered is immediately converted into a standard internal metadata representation (MARCXML) whereas fulltexts are converted into PDF and directly submitted into the document server. Upon upload into the bibliographic server, metadata can be the subject of quality assessment procedures by library cataloguers. Metadata is additionally enriched with citation extraction from the relevant fulltexts. The bibliographic server can then be queried to generate indexes, ranks, clusters and formats of bibliography, suitable for fast retrieval. The information is finally delivered, at the bottom, to users and OAI service providers, through OAI-PMH requests, email alerts and the web search engine. In addition, the web interface offers access to personalized collections of documents (baskets), documentation and statistics. Most of the interactions between key modules and the persistence layer (the Database) are synchronized through a task scheduler (BibSched) that can also be used to run tasks in periodical daemon modes. A full account of the operational logic of each CDSware module can be found at [Vesely2003].

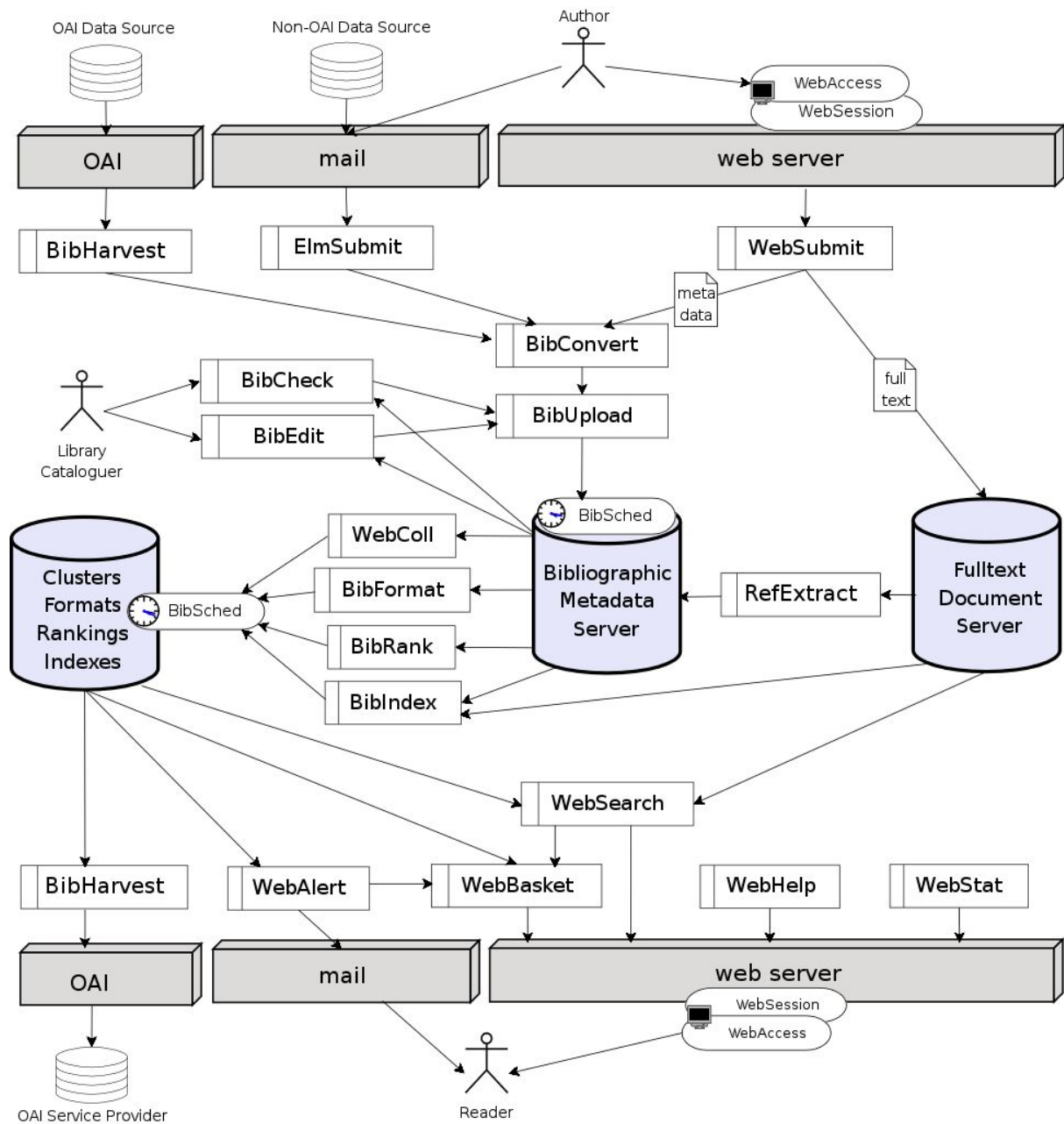


Figure 1. Document workflow in the CDSware system: from acquisition to delivery.

3. A glimpse into CDSware technology

In this chapter we intend to highlight some of the technological features that distinguish CDSware from other freely available digital library or repository software, such as DSpace [DSpace] or EPrints [GNUEprints]. A more in-depth discussion of the software's capabilities together with the procedures of installation, configuration and administration can be found in the online user and admin-level documentation guides [CDSAdmin].

3.1 Metadata representation

All the bibliographic data in CDSware are internally represented in the MARC 21 format. This metadata structure is the chosen internal representation for a number of reasons:

- it is a well-established standard in the library world - used since the 1960s;
- it blends well with modern mark-up technologies, such as XML; CDSware uses recently standardized MARCXML format, provided by the Library of Congress;
- it is flexible enough to guarantee long-term reliability;
- it can be thoroughly extended to adapt to any metadata structure; at CERN, current MARCXML schema includes more than 150 metadata fields.

Institutional repositories with homogeneous document types usually do not feel the necessity to go into a full MARC cataloguing system. In this case, they can use CDSware's default markup scheme that presets the most commonly used metadata fields. A non-exhaustive sample list is shown in Figure 2.

METADATA CONCEPT	MARC 21 REPRESENTATION
Abstract	520 \$a
Author, first	100 \$a
Author(s), additional	700 \$a
Email	8560 \$f
Keywords	6531 \$a
Language	041 \$a
OAI identifier	909CO \$o
References	999C5 \$* [many subfields]
Primary report number	037 \$a [unique throughout the system!]
Additional report number(s)	088 \$a
Title	245 \$a
URL (e.g. to fulltext)	8564 \$u, \$z
etc.	

Figure 2. An excerpt of CDSware's default MARC representation scheme

In other circumstances, users may want to adopt CDSware's defaults and implement additional markup tags for local specific metadata concepts. Pushing it even further, certain users may want to define very particular data objects and thus, implement a new markup scheme of their own. This extreme markup extensibility together with CDSware's configurability allows the system to handle virtually any type of metadata concept (e.g. museum objects or multimedia presentations). A comprehensive guide to MARC representation can be found at [MARC].

3.2 Metadata Acquisition

Metadata acquisition is performed by automated and semi-automated procedures of document harvesting (BibHarvest module) - by applying either standardized approaches such as using the OAI-PMH compliant metadata harvesting [Vesely2002] or ad-hoc procedures such as shallow Web harvesting. Document submissions may be done directly by authors over the Web or e-mail using the WebSubmit and ElmSubmit modules. In both cases metadata is gathered in raw form, converted into the native CDSware metadata representation and finally uploaded into the bibliographic metadata server.

This conversion is done within the BibConvert module that allows conversions between various sequential (e.g. ISO2709) and semi-structured (e.g. XML) formats, between various metadata formats (MARC21, DublinCore, RFC1807, etc.) and features detailed text formatting including regular expressions specified in the BibConvert Configuration Language.

The BibConvert Configuration Language provides a specification of the syntax and semantics of the metadata conversion description that is encoded in a set of conversion templates. Templates consist of the data source extraction template that provides a description of a source record, data source template that provides a description of each field extracted from the source and data target template that describes the layout of the target record. An extensive documentation on the usage and configuration of BibConvert can be found at [CDSAdmin].

In addition BibConvert features matching functionality that allows to match gathered records against the bibliographic metadata server content reducing the risk of database double entries by multiple upload. Records marked as ambiguous or refused within the matching step are then to be treated manually by the metadata acquisition administrator. Confirmed metadata is subsequently uploaded in the MARC21 representation in XML as described in 3.1.

Metadata conversion through BibConvert allows a high degree of automatization: metadata records from several different sources can be easily imported into MARCXML and immediately entered in the system by simply using a handful of standard configuration files. In order to establish the authenticity of the metadata entering the system, library cataloguers can perform quality assessment through module BibCheck.

3.3 Indexing and Ranking

The indexing and ranking modules are at the core of the CDSware system. Specially designed indexes were introduced in order to provide Google-like speed for repositories of about 1,000,000 documents (see Figure 3). Moreover, on top of the metadata searching, CDSware can index and search fulltext files and document references in one go, providing the possibility of a combined metadata/fulltext/references search. For example, a query like *find all documents written by Ellis in 2002 that mention the term Higgs boson in the fulltext and that refer to Physical Review D 1997 papers* is very possible.

CDS database size	~650,000 records
CDS collection size	~450 collections
word indexing time	~2 days
word index size	~5 GB
word index size	~3,000,000 words
global word index growth rate	~3 words per record
title word index growth rate	~0.1 words per record
search speed for a word query `cern' (223843 hits)	0.07 sec
search speed for a boolean query `of cern' (109635 hits)	0.10 sec

Figure 3. Word index statistics for the CERN Document Server database, as of 2004. Note that the word indexes were designed with the aim of providing fast search times perceived by the end users at the expense of slow indexing times perceived by the administrators.

The results retrieved by the search engine can further be ranked according to several criterias. The default CDSware installation includes the classical word-frequency based vector model that permits one to retrieve similar records. Furthermore, a ranking method machinery based on specific metadata values is included: as an example, the journal impact factor ranking method which ranks documents using a configurable knowledge base of journal titles and their respective journal impact factors. Finally, the new experimental ranking features in CDSware include the possibility to rank by the number of citations and the number of downloads.

The search results are clustered into collections. The administrator of the system has the possibility to define regular and virtual collection trees to ease the navigation in the document corpus. Furthermore, automatic classification studies are under development with the aim of providing an intelligent, automated, result clustering and navigation.

3.4 User interface and personalization

CDSware can handle virtually any electronic material thanks to the flexible MARC format, as outlined above. In order to display these various document types properly and accordingly to the specificities of each format, a

flexible output formatter with the possibility of automated link generation to external resources based on the record content is used.

The user interface proposes a number of personalization and collaborative features. The end users who register can set up their personal collection of documents (baskets) and periodical notifications about newly added documents in their areas of interest (alerts). Groupware collaborative features include the possibility of declaring certain baskets public and sharing their content with other users. More collaborative social-software features such as commenting and reviewing of records are currently being worked on.

CERN being an international and multi-cultural environment, the internationalization of the CDSware is another asset worth mentioning. The search interface has been translated to 13 languages (Czech, German, Greek, English, Spanish, French, Italian, Norwegian, Portuguese, Russian, Slovak, Swedish, Ukrainian) enabling the end user to dynamically select the language of her choice. Most of the translations are being maintained by the members of the developers team and by the CERN users. About a third of translations were contributed by administrators of other installations of CDSware in the world.

4. Conclusions

We have presented CDSware, an all-inclusive application framework which allows to run an autonomous digital library server. We have introduced the software by illustrating its typical operational workflow, highlighting its modular architecture. We then focused on the distinctive features of the software which may make it an interesting solution for the management of a document management system, especially in the context of large, heterogeneous repositories.

The effort to keep the software capabilities at the bleeding-edge by retaining a high degree of customization has resulted in an increase of complexity. For this reason, the first approach with CDSware's installation and administration may be a little demanding. The effort spent in setting it up and running it is rewarded with extreme configurability and the capability to meet the specific needs of virtually any kind of data repository.

References

- [CDSAdmin] CERN Document Server. Admin Area and Hacking corner. From <http://cdsware.cern.ch:8000/admin/> and <http://cdsware.cern.ch:8000/hacking/>
- [CSDSDemo] CERN Document Server. Demo and Production Sites. From <http://cdsware.cern.ch/demo/>
- [DSpace] DSpace Federation. DSpace. From <http://www.dspace.org/>
- [EUIndico] EU Indico Project. Integrated Digital Conferencing. From <http://www.cern.ch/indico>
- [GNUePrints] GNU EPrints. The eprints.org Software. From <http://www.eprints.org/>
- [GooglePrint] Google Print. Library Project. From <http://print.google.com/googleprint/library.html>
- [Gurrey2005] Gurrey, B. & de Roux, E. (2005). Jacques Chirac veut promouvoir un projet de bibliothèque virtuelle européenne. From <http://www.lemonde.fr/web/article/0,1-0@2-3246,36-401828,0.html>
- [Harnad2005] Harnad, S. (2005). Fast-Forward on the Green Road to Open Access: The Case Against Mixing Up Green and Gold. From <http://www.ariadne.ac.uk/issue42/harnad/>
- [MARC] HOWTO MARC Your Bibliographic Data. From <http://cdsware.cern.ch:8000/admin/howto/marc.html>
- [LeMeur2004] Le Meur, J-Y., Sanchez, H., Baron T., Gonzalez J., Turney V. (2004). Indico - the software behind CHEP 2004. From <http://indico.cern.ch/contributionDisplay.py?contribId=282&confId=0>
- [Suber2005] Suber, P. (2005). Open Access Overview: Focusing on open access to peer-reviewed research articles and their preprints. From <http://www.earlham.edu/~peters/fos/overview.htm>
- [Vesely2002] Vesely, M., Baron, T., Le Meur, J-Y. & Simko, T. (2002) Creating Open Digital Library Using XML : Implementation of OAi-PMH Protocol at CERN. In: Intl. Conference on Electronic Publishing, Karlovy Vary, Czech Rep.
- [Vesely2003] Vesely, M., Baron, T., Le Meur, J.Y. & Simko, T. (2003). CERN Document Server : Document Management System for Grey Literature in Networked Environment. *Publ. Res. Quarterly* 20, 1, 77-83.