A Search Engine for the Engineering and Equipment Data Management System (EDMS) at CERN

A. Tsyganov, S. Mallón Américal, S. Petit, T. Pettersson, A. Suwalska
CERN, TS Department, CH-1211 Geneva 23, Switzerland
Andrey.Tsyganov@cern.ch

Abstract. CERN, the European Laboratory for Particle Physics, located in Geneva - Switzerland, is currently building the LHC (Large Hadron Collider), a 27 km particle accelerator. The equipment life-cycle management of this project is provided by the Engineering and Equipment Data Management System (EDMS [1] [2]) Service. Using an Oracle database, it supports the management and follow-up of different kinds of documentation through the whole life cycle of the LHC project: design, manufacturing, installation, commissioning data etc... The equipment data collection phase is now slowing down and the project is getting closer to the “As-Built” phase: the phase of the project consuming and exploring the large volumes of data stored since 1996. Searching through millions of items of information (documents, equipment parts, operations...) multiplied by dozens of points of view (operators, maintainers...) requires an efficient and flexible search engine. This paper describes the process followed by the team to implement the search engine for the LHC As-built project in the EDMS Service. The emphasis is put on the design decision to decouple the search engine from any user interface, potentially enabling other systems to also use it. Projections, algorithms, and the planned implementation are described in this paper. The implementation of the first version started in early 2007.

1. INTRODUCTION

CERN, the European Laboratory for Particle Physics, is currently commissioning the LHC (Large Hadron Collider), a 27 km particle accelerator. The equipment data life-cycle management of this project is provided by the Engineering and Equipment Data Management System (EDMS [1] [2]) Service. Using as data repository an Oracle database, the Service supports document and equipment data management throughout the entire life cycle of the LHC project: design, manufacturing, installation, commissioning data etc... The requirements for the development of a new EDMS Search module was perceived now as the LHC approaches its commissioning and operational phases. The EDMS today manages about 1’000’000 documents with technical information about the CERN accelerator complex and about the same amount of data about specific equipment installed. To easily find accurate information in this still rapidly growing repository of information a fast and flexible search engine is needed.

1 To whom any correspondence should be addressed.
The CERN EDMS uses two commercial products in a unique combination (Agile and Datastream7i, both are using an Oracle database as their basic repository); the document and equipment data is stored in the internal structures of these systems. To abstract their internal architecture and be able to “easily” absorb any possible product evolutions the EDMS common layer was created (figure 1).

![Figure 1. The EDMS Common Layer](image1)

Until recently the system abstraction was implemented using Oracle Views and PL/SQL [3] procedures. However, due to the normal evolution of the system and its intense usage, certain of these views have become very complex, difficult to maintain and keep efficient (figure 2).

![Figure 2. The EDMS architecture](image2)

In addition, the number of complex queries in the system based on these views using often slight, but subtle, different aspects kept growing in a worrying manner leading to problems with the overall EDMS performance. The arrival of the LHC “As-Built” phase with its new sources of equipment data and documents related to the commissioning phase of the accelerator finally launched the development of a new, more powerful, search engine.
2. System overview

2.1. Scope

The objective of the new search engine is to provide a fast and flexible equipment data and document search functionality. The objective is to replace the many statements in the various, diverse and complex abstraction views by one single, dedicated engine which uses simple but efficient database instructions (figure 3).

![Figure 3. The EDMS new search architecture](image)

Re-engineering the many abstraction view statements that have been added to the system over the last 10 years also improves the software development process. Once the central Search Engine has been correctly tuned many of the previous SQL queries scattered around the system can be removed, thus eliminating sources of inefficiency, old obscure and hard-to-maintain code as well as potential performance problems. Having a single central search module offers benefits to several types of clients and opens the scope to new applications: software API’s, other PL/SQL stored procedures and even Web Services.

2.2. Search engine architecture

The EDMS manages different classes of information and each class of information has its own, specific semantic and specific structure. The LHC engineering data can grossly be considered as consisting of data describing equipment, data describing functional positions and data describing locations in a tridimensional space. An additional, complex, class of data constitutes the Job class. A Job consists of a series of elementary work steps that each has to be completed before the given task, the Job, is done. Each job step has attributes and status indicators which are of interest when searching for non-conformities and accelerator performance problems with a given equipment or equipment installed in a suspected given volume in space.

For each information class simple search modules are designed and implemented, each module is then carefully tuned to give the best possible performance. These modules are known as atomic searches and they will constitute the basic elements of the backbone of the new search engine.
The equipment data atomic search will be used as an illustration of the approach. This module has search parameters that are specific to the equipment description data structures and an equipment specific query builder. The module can not only handle simple requests such as: find all equipment whose identification code begins with HCMB\(^2\), but it can also handle more complex requests: find all equipment whose identification code starts with HCMBA or HCMBI or HCMBC. This feature is supported by a special light-weight semantic analyzer, which is able to build non-trivial database requests using its own specific semantic rules. This design approach offers the possibility to do complex, programmed, searches on equipment data without having to know in detail the structure of the data. Similar atomic search modules are being built for the other classes of LHC information managed by the EDMS.

The semantic analyzer is common to all the atomic search modules and uses the same syntax description for all atomic searches. By using a large variety of atomic searches the system is now able to efficiently retrieve items from selected subsets of LHC information classes.

Another dimension can be explored by combining several atomic searches. An example describes the new approach: find all equipment descriptions with an identification code that begins with HCMBA and that are linked to documents which are classified as “manufacturing Non-conformity”. This search example uses a combination of the equipment atomic search module and the document atomic search module. When using such a combination of several atomic search modules in sequence a search strategy decision is necessary. In the example given, is it more efficient to search on the document data first and only then on equipment description data or should the search be executed in the reverse order?

It is thus seen that an efficient search process has to be able to combine the search results of atomic searches in a uniform way. To obtain the required level of abstraction the Oracle Object Types (OOT) \[4\] was used. All the different, LHC information class specific, atomic searches are built using the same model and the same output type – OOT (figure 4).

![Figure 4. Internal architecture of the EDMS atomic search](image)

The combination of the atomic search modules and the search strategy optimizer which communicate through collections of Oracle Objects, constitute the EDMS search Engine (figure 5).

\[2\] These codes refer to specific types of elements in the LHC accelerator as defined in the LHC Quality Assurance Plan.
3. CONCLUSION

The benefits of the new EDMS search engine can be summarized as follows:
By federating several well tuned and smaller, subject specialized, search engines, a wider search spectrum can be scanned better.
Thanks to its internal level of abstraction based on object oriented paradigm, a more in-depth processing of information having different semantics and structures is possible.
Adding another subset of information to be scanned has become relatively easy thanks to the same object paradigm. Being central, the new search engine offers wide connectivity possibilities: it can be integrated into different advanced user applications or into various graphical user interfaces or Web Services that are in rapid evolution. Although the internal data structures are hidden, some of the modules' input parameters can be used to specify complex searches with the help of light weight syntax. New requirements on data mining can now be implemented by developers who do not need to know specifics of the low level architecture.
Detailed user requirements for the LHC “AS-Built” project are still in the specification stage but the objective is to provide a general purpose LHC information data search engine able to cope with new requests and new information classes as they appear in the future life-cycle phases of the LHC.

References
[1] EDMS, https://edms.cern.ch
[2] Pettersson T 2003 https://edms.cern.ch/file/370320/1/micado1.doc
[3] Oracle Corporation 2002
   http://download.oracle.com/docs/cd/B10501_01/server.920/a96524.pdf
[4] Oracle Corporation 2002
   http://www.oracle.com/technology/products/oracle9i/pdf/simple_strat_for_complex_rel2.pdf