PROCEEDINGS OF SCIENCE

WebDat: Bridging the Gap between Unstructured and Structured Data

Jerzy M. Nogiec[1]
Fermi National Accelerator Laboratory
Batavia, IL 60510, USA
E-mail: nogiec@fnal.gov

Kelley Trombly-Freytag
Fermi National Accelerator Laboratory
Batavia, IL 60510, USA
E-mail: kfreytag@fnal.gov

Ruben Carcagno
Fermi National Accelerator Laboratory
Batavia, IL 60510, USA
E-mail: ruben@fnal.gov

Accelerator R&D environments produce data characterized by different levels of organization. Whereas some systems produce repetitively predictable and standardized structured data, others may produce data of unknown or changing structure. In addition, structured data, typically sets of numeric values, are frequently logically connected with unstructured content (e.g., images, graphs, comments). Despite these different characteristics, a coherent, organized and integrated view of all information is sought out. WebDat is a system conceived as a result of efforts in this direction. It provides a uniform and searchable view of structured and unstructured data via common metadata, regardless of the repository used (DBMS or file system). It also allows for processing data and creating interactive reports. WebDat supports metadata management, administration, data and content access, application integration via Web services, and Web-based collaborative analysis.

XII Advanced Computing and Analysis Techniques in Physics Research
Erice, Italy
3-7 November, 2008

[1] Speaker

1. Introduction

Information systems have grown around structured data made up of fields, columns, tables, indices. As such, they present a very predictable, static and ordered environment. In contrast, unstructured data have unknown organization and consist of documents, spreadsheets and rich media information and, therefore, create a rather disorderly and fairly dynamic environment. To handle these two distinct types of information, two major classes of data management systems emerged: Data Base Management Systems (DBMS) to manage homogenous, well-typed structured data and Content Management Systems (CMS) to manage non-traditional, heterogeneous and unstructured contents. Despite this dichotomy between handling unstructured and structured data, all information, regardless of its format, source, and location, has to be easily managed, searched and accessed. Since access to all the available data is needed, systems that will provide an integrated and uniform access to heterogeneous information are sought after.

This is specifically evident in the R&D environment such as found at Fermilab's Magnet and Cavity Test Facilities. At those facilities, there exist various test and measurement systems ranging from production systems developed to test many similar magnets needed for big accelerator projects, to rapidly developed systems to perform a specific, one-of-a-kind test. The collected data ranges from well-structured, homogeneous, and stable (well-suited for DBMS) to unstructured collections of data files and documents (well-suited for CMS). Examples of the structured data are numerical results of repeatable measurements or calibrations, whereas examples of the unstructured data are test plans, analyst comments, measurement conditions, screenshots, plots and connection diagrams.

Creating a system to organize structured and non-structured data usually involves extracting metadata from contents and tagging the various components of data. This allows for organizing and retrieving information appropriately to address the specific needs of different customer groups. Another way to organize data is to create a specialized set of metadata that contains all necessary descriptions for the structured and non-structured information. However, systems that implement this must have specific domains and be limited to well-defined applications.

Both the general research trend to unify the management of unstructured and structured data and the specific situation at the Magnet and Cavity Test Facilities prompted the authors to work on an integrated system for managing test data. The specific and limited domain found here lends itself quite well to the specification of a core set of metadata to be used to integrate the unstructured data kept in files and the structured data kept in databases.

2. Design

WebDat is a unified data and content management system and data portal targeted at managing test (measurement) results, unlike other systems of this type that focus on general information management [1] [3]. The idea behind WebDat is based on metadata, common for unstructured and structured data for any given test.

The value of structured data lies in their completeness and in the relationships between data items rather than data items themselves. The existence of well-defined metadata in WebDat guarantees the presence and completeness of vital relationships between stored data items and the external conditions and factors related to each test.

2.1 Design Goals

WebDat has been developed in response to the desire to develop a Web-based system utilizing a database to organize data and documents pertaining to tests. The system, apart from allowing sharing information, was to be able to:

- Allow registering and cataloging data and related documents for each test.
- Keep data and analysis results organized, searchable and accessible.
- Authenticate and control access to data.
- Integrate documents and numerical data.
- Preserve information about the test procedure, data acquisition system, data reduction, and other data pertaining to the test as a part of the well-formed metadata.

2.2 WebDat Functionality

The design goals specified above have been translated into a set of functional requirements, which have been allocated to several user categories, or roles: the coordinator, the tester, the analyst, and the data user (see Figure 1).

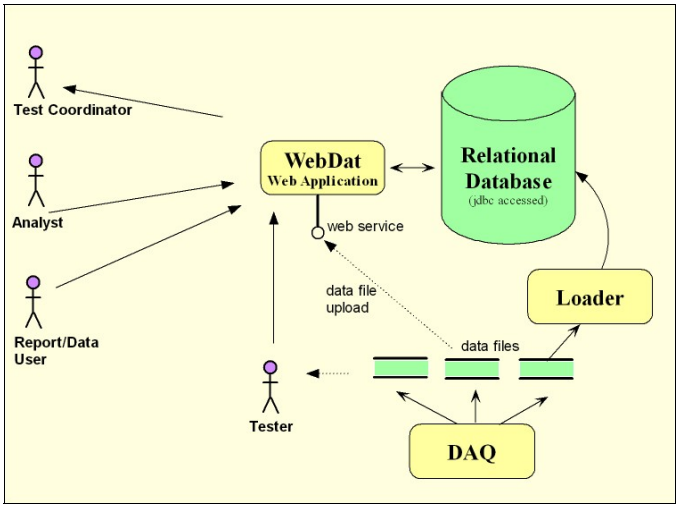

Figure 1: Use of the system.

The test coordinator is in charge of organizing test processes. He maintains information about measurement infrastructure (facilities, stands, hardware, and software systems), available test types and test series. He also maintains metadata regarding test subjects (magnets, cavities and other accelerator components), which includes registering new subject types, subject series and subjects.

The tester is a person responsible for conducting the test. She registers new tests with the system and relates them to the infrastructure, available test types and subjects. She is also responsible for storing the acquired data, which can be accomplished automatically or manually, and storing any other contents pertaining to the test (test plans, reports, screenshots, configurations, etc.). During the test, she may also record comments.

The analyst is in control of conducting the analysis of data and/or verifying the results of automated analysis. He stores the results of his analysis as well as his comments and conclusions.

The end user can search for a particular test, view dynamically generated reports as well as check available statistical information (e.g., contents for given test, tests performed in the last week, contents available for a given subject, contents available for a given test type). She can also download all submitted documents and data files in their original formats and perform her own analysis.

2.3 Metadata

All the data managed by the system is related together by the metadata. The term metadata has been coined for "data about data" and in the case of WebDat describes the information pertaining to tests that allows for correlating data collected from the DAQ and analysis systems with the information about the testing infrastructure, test subjects, and test procedures. The relation between the test and other data is shown in Figure 2. In order to simplify the diagram related tables have been grouped together and depicted as packages.

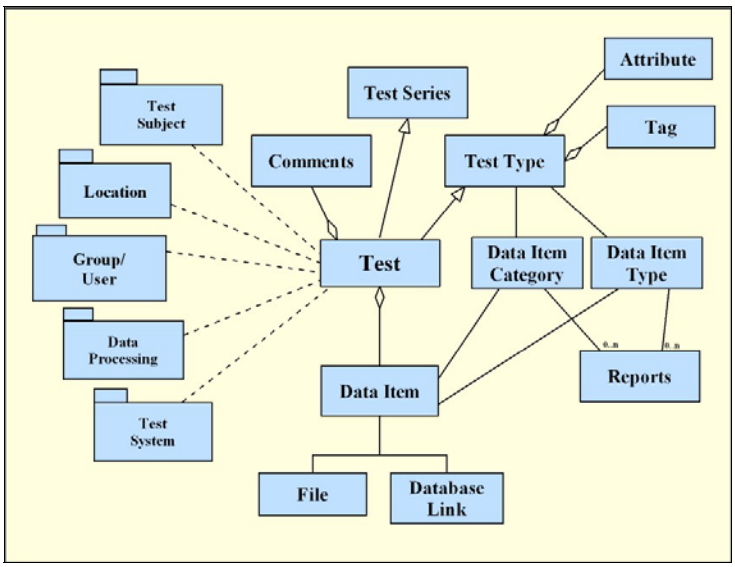

Figure 2: WebDat metadata.

Each test is of a particular type and can belong to a collection (series) of tests. A test type has associated with it attributes and tags. An attribute is a required parameter of the test that must be supplied upon creation of the test. Tags, similar to keywords, characterize the test and aid in searching for similar categories of tests.

A tested item, called subject, is of a particular type and can belong to a series (collection) of subjects, e.g. a concrete production series of magnets.

The infrastructure connected with the test is described by the versioned location details (e.g., test stand, measurement hall), and the versioned software and hardware systems used to acquire and process data.

Finally, the test has associated with it data items, which are stored in a database or a file system and reports that can be ran on those data.

3. Implementation

WebDat has been implemented with the J2EE technology using JavaServer Faces, MySQL, Apache Tomcat and Axis, and JasperReports. At the core of the system is a database that stores metadata together with structured data and references to externally stored contents. The database is accessible exclusively via the application server. Users can interact with the system using their browsers, whereas applications use a Web service. Apart from the JSF and Web service runtimes, the application server also hosts an authentication filter, upload and download servlets, data processors and reports (see Figure 3).

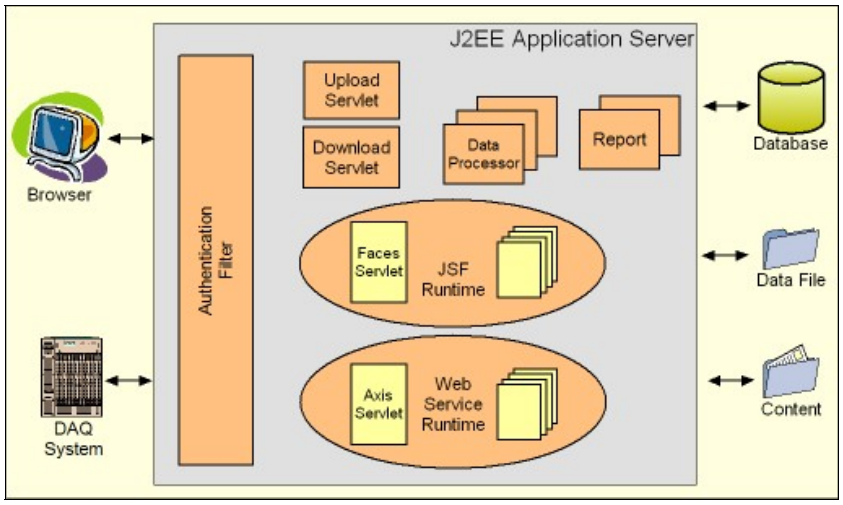

Figure 3: WebDat architecture.

Both the user interface (browser) and the application interface (Web service) offer similar functionality, including upload and download of data, test searches with metadata, retrieval of comments, and access to statistical information. In addition, the user interface allows for drag & drop upload of individual files, sets of files or whole folders with files, and viewing reports.

The system provides automatic processing of data, either upon data upload (pre-processing) or download (post-processing). The examples of automatic processing include: format change, contents verification, automatic analysis and compression of data.

4. Summary

As it is evident from the literature on the subject, there is a need and interest in uniform management of structured and unstructured data. The R&D environment is one of the areas that would benefit from such an approach. The WebDat application developed at Fermilab addresses this need by providing a system based on a common set of metadata for both structured and unstructured data. It provides a relationship between otherwise unrelated data items and allows for consistent and uniform access to file and database data sources. It supports interactive (Web-based user interface) and programmatic (Web service) insertion and retrieval of data as well as automatic processing of data upon their insertion or retrieval. The WebDat is especially suited for organizing loosely coupled data coming from collections of heterogeneous systems in dynamic environments.

References

[1] G. Goth, A Structure for Unstructured Data Search, IEEE Distributed Systems Online, January 2007.

[2] W.H. Inmon, A. Nesavich, *Tapping into Unstructured Data: Integrating Unstructured Data and Textual Analytics into Business Intelligence*, Prentice Hall, 2007.

[3] D.A. Maluf, P.B. Tran, *Managing Unstructured Data with Structured Legacy Systems*, Aerospace Conference, 2008 IEEE, March 2008.

[4] D. Crampton, *How to...deal with structured and unstructured data*, Techworld, January 2004.