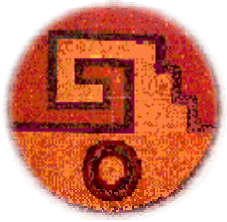
ACAT 2008 Erice, Sicily



WebDat: *Bridging the Gap between Unstructured and Structured Data*

Jerzy M. Nogiec,
Kelley Trombly-Freytag,
Ruben Carcagno

Fermilab, Batavia, Illinois

Outline

- **Management of Unstructured and Structured Data**
  - Structured vs. unstructured data
  - Data management challenge in R&D environment
  - Concept of integrated data management system
- **WebDat Design**
  - Design goals
  - Required functionality
  - Interaction with the system
  - Lifecycle of information access
  - Metadata
  - Contents
  - Automatic processing
- **WebDat Implementation**
  - Technologies
  - Deployment
  - User interface
  - On-line generated reports
  - Web service
- **Conclusion**
  - Features
  - Summary

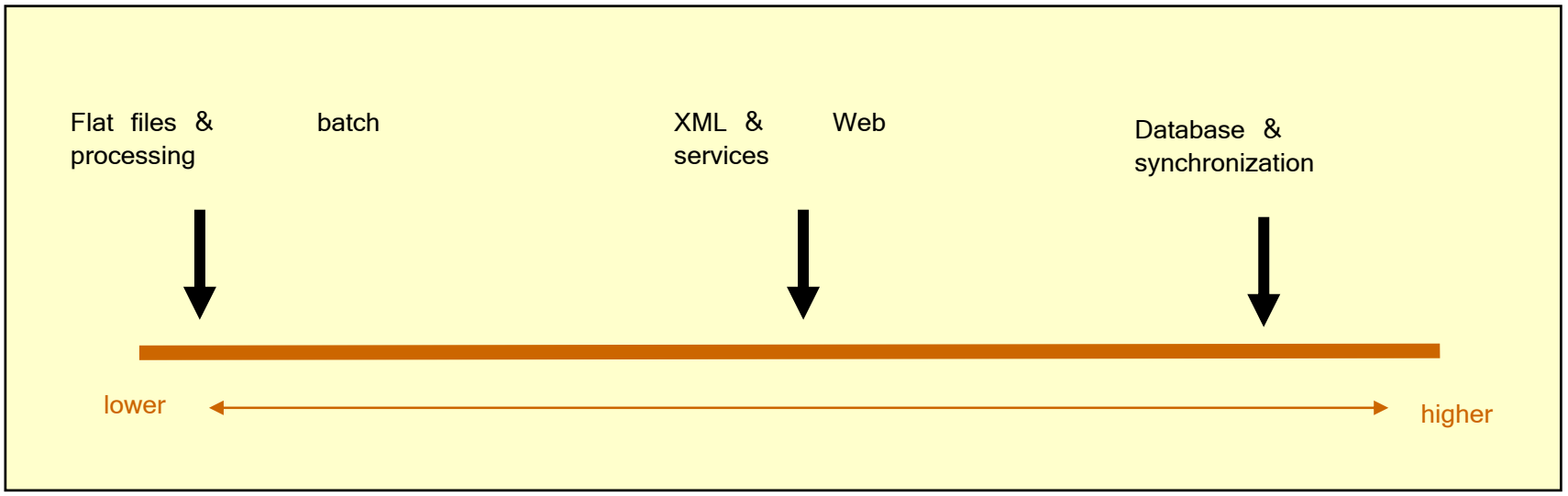
I. Management of Unstructured and Structured Data

Structured & Unstructured Data

- Information systems have grown around structured data made up of fields, columns, tables, indices; very predictable, static and ordered environment.
- The unstructured data have unknown organization and consist of documents, spreadsheets, rich media information; rather disorderly and fairly dynamic environment.
- Two major classes of data management systems evolved:
  - DBMS to manage homogenous, well-typed structured data.
  - CMS to manage non-traditional, heterogeneous, unstructured contents.
- Information regardless of its format, source, location has to be easily managed, searched and accessed. Access to all data is needed.

Systems that will provide an integrated, uniform access to heterogeneous information are needed.

Levels of Structure in Data



The level of structure of data provided by various data management and exchange models.

R&D Data Management Challenges

- There exist production systems as well as rapidly developed systems.
- Information ranges from well-structured, homogeneous, and stable (well-suited for DBMS) to unstructured collections of data or documents (well-suited for CMS).
- Data is owned and kept by several groups and individuals in various formats and locations.
- Systems have different level of completeness in data handling and configuration management.
- Information about DAQ systems varies since they range from configurable to single-purpose systems (possibly tailorable via parameter files).

Concept of Integrated System

- Create a set of common and necessary metadata that is configurable via UI.
- Use common metadata for unstructured data in files and structured data in databases.
- Define metadata first and create tests and upload data, contents and comments later.
- Allow for manual and programmatic access.

II. WebDat: Design

WebDat Design Goals

Develop a Web-based system using a database to organize data and documents pertaining to tests. The system will allow for sharing information and results. It will also:

- Adapt to various levels of integration and maturity of DAQ systems.
- Allow registering and cataloging data and related documents for each test.
- Keep data and analysis results organized, searchable and accessible (allow searching and navigating to data).
- Authenticate and control access (passwords, groups).
- Integrate documents and data.
- Preserve information about the DAQ system, test procedure, data acquisition system, data reduction, etc. (as much as supplied)

WebDat Functionality

Maintain information about systems and subjects

- Keep information about measurement infrastructure (facilities, stands, hardware, software, versions).
- Register new test types and series.
- Register new subject types, subject series and subjects.

Register new tests / upload data and documents

- Register tests (measurements) and relate them to infrastructure.
- Store data for a registered test (user file upload, programmatic file upload or data inserts).
- Store any documents pertaining to the test (test plans, reports, screenshots, configurations, etc.).
- Add comments to stored tests about documents, results, etc.

Retrieve / search / download data and documents

- Retrieve (download) all submitted documents and data files.
- Search tests using subject names, dates, test types, etc.
- View reports generated on-line from data (limited to 'installed reports').
- View statistical info (e.g., contents for given test, tests in last week, contents for given subject, contents for given test type).

Metadata

Test/Run

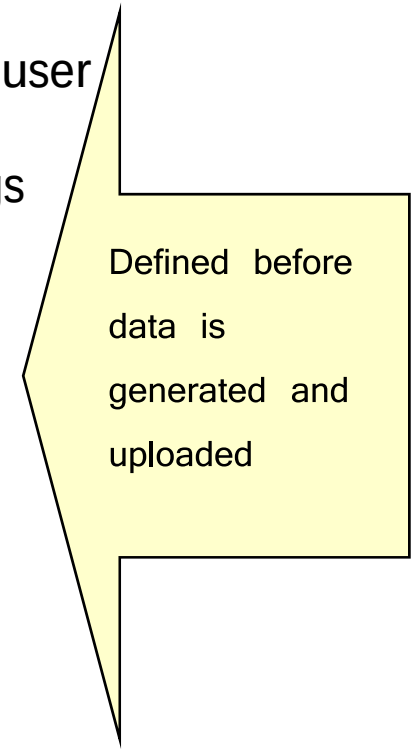
- **Test attributes**: required test parameters entered by the user
- **Test tags**: keywords characterizing test
- **Test types**: a type of test defined by its keywords and tags
- **Test series**: a collection (series) of tests

Subject

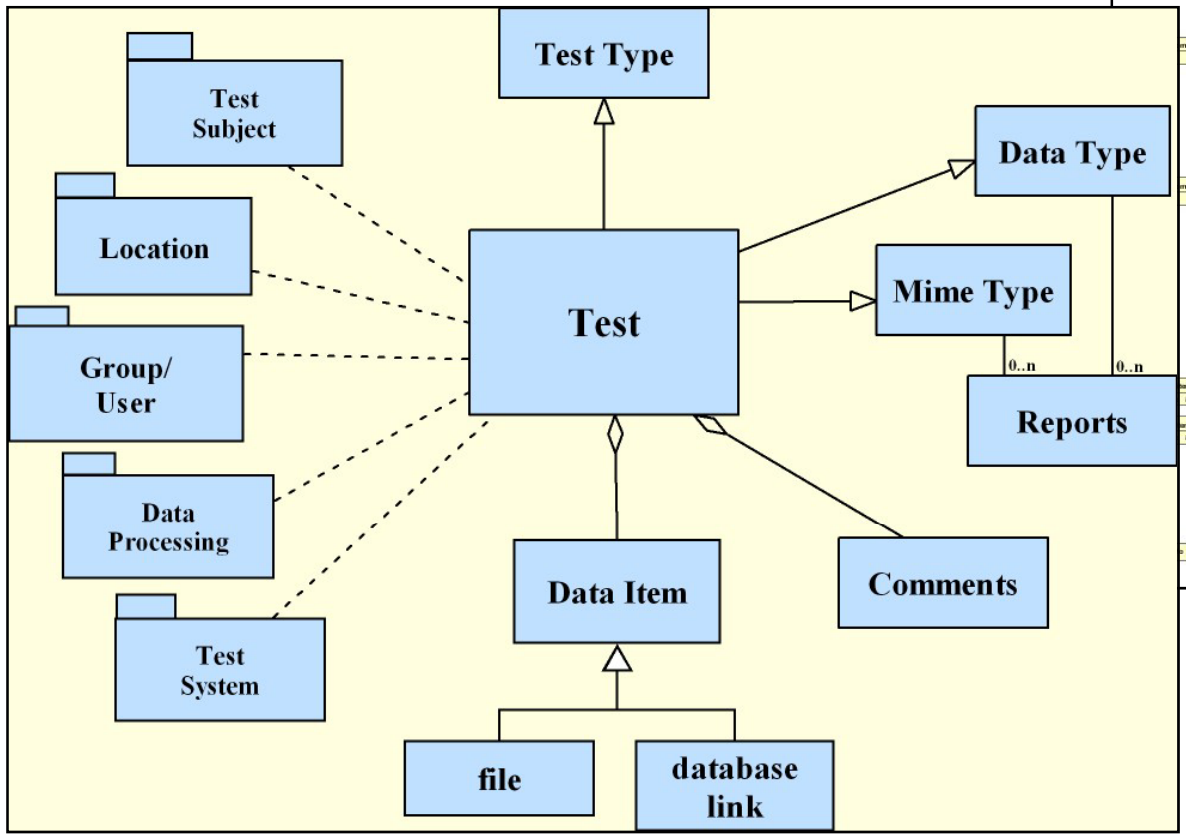
- **Subject type**: a type of subject (e.g., dipole, cavity)
- **Subject series**: a series of subjects

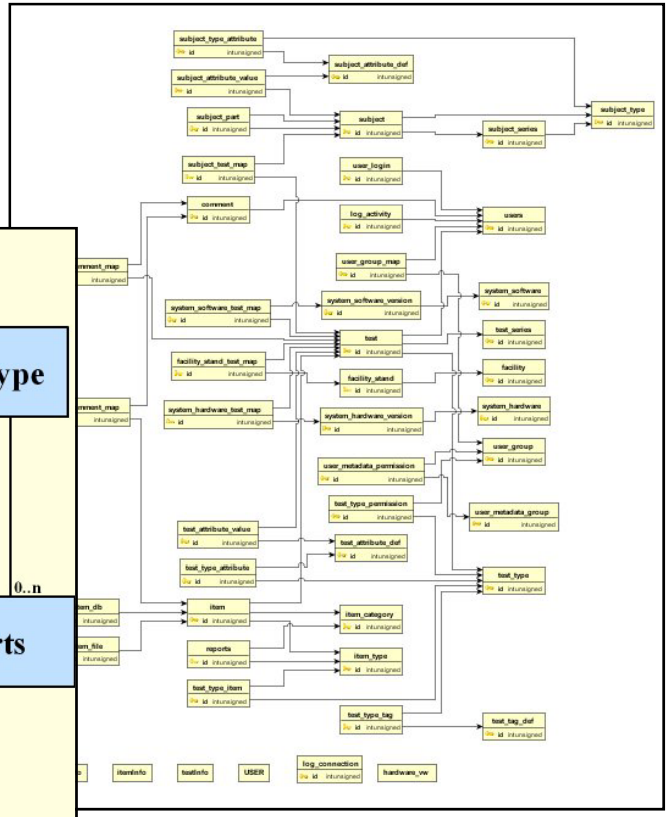
Test/Run environment

- **Location**: facility-stand pair
- **Software**: system and version
- **Hardware**: systems and version



Metadata


Test Type
Test Subject
Location
Group/ User
Data Processing
Test System
Test
Data Type
Mime Type
Reports
0..n
0..n
Data Item
Comments
file
database link

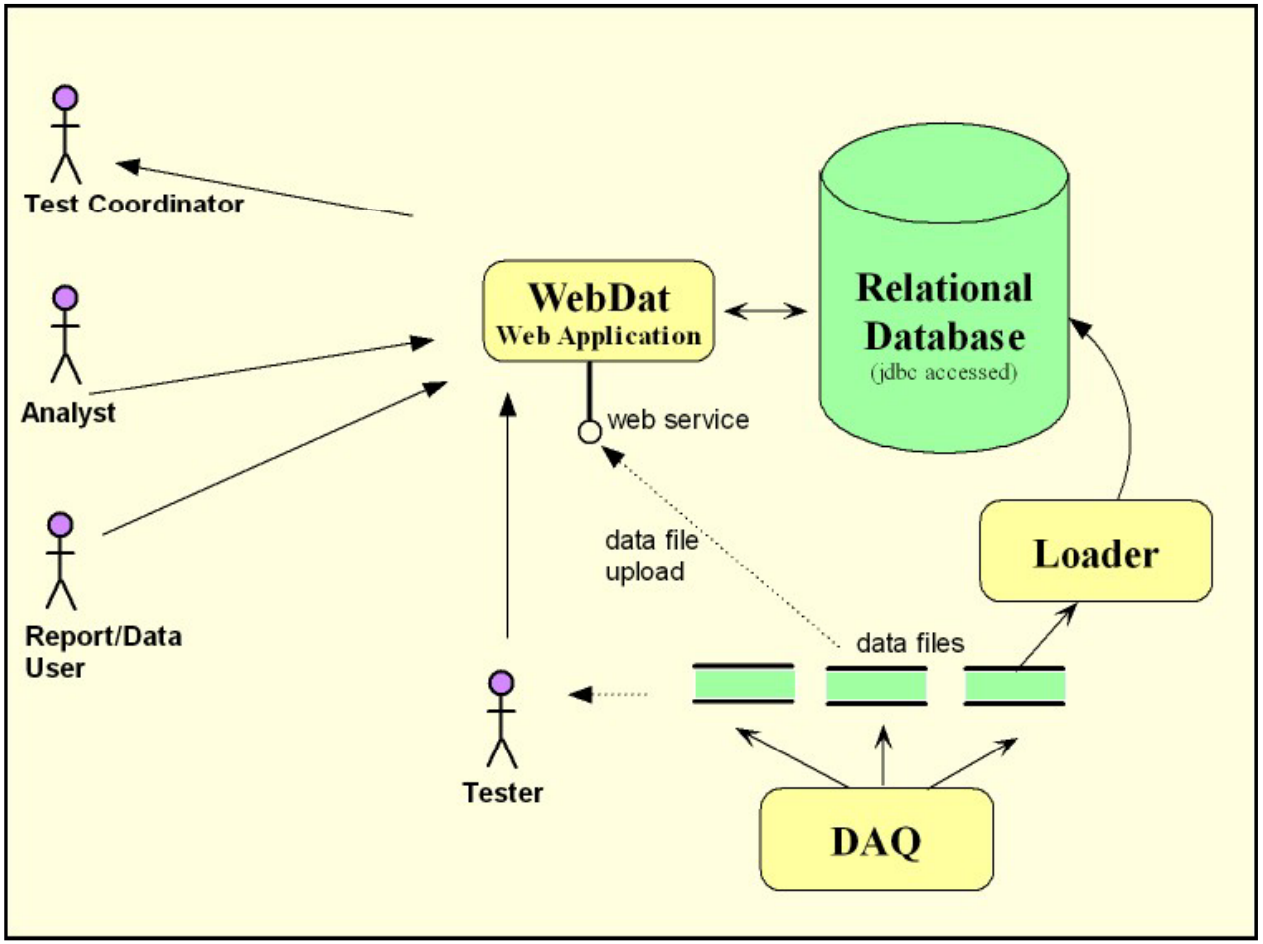


Unstructured and Structured Data

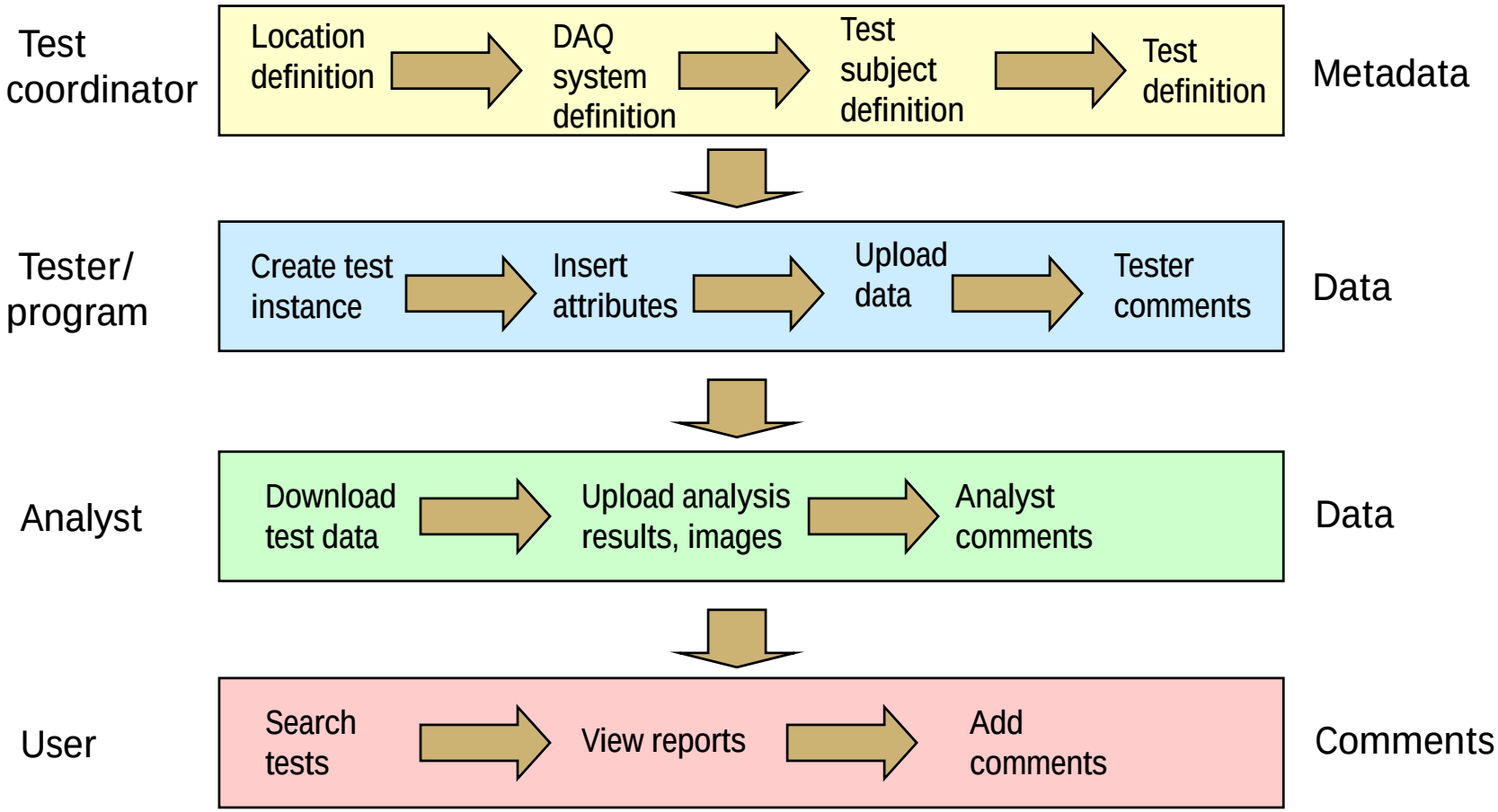
Test items – a collection of data and documents pertaining to a test (measurement).

- A document
  - An uploaded file (Word, Excel, jpg, gif, ...)
- Data
  - Files
  - Data kept in relational model (loaded to tables)
- Comments (on-line entry)

Interactions with WebDat



Lifecycle of WebDat Information



Data Insertion & Retrieval

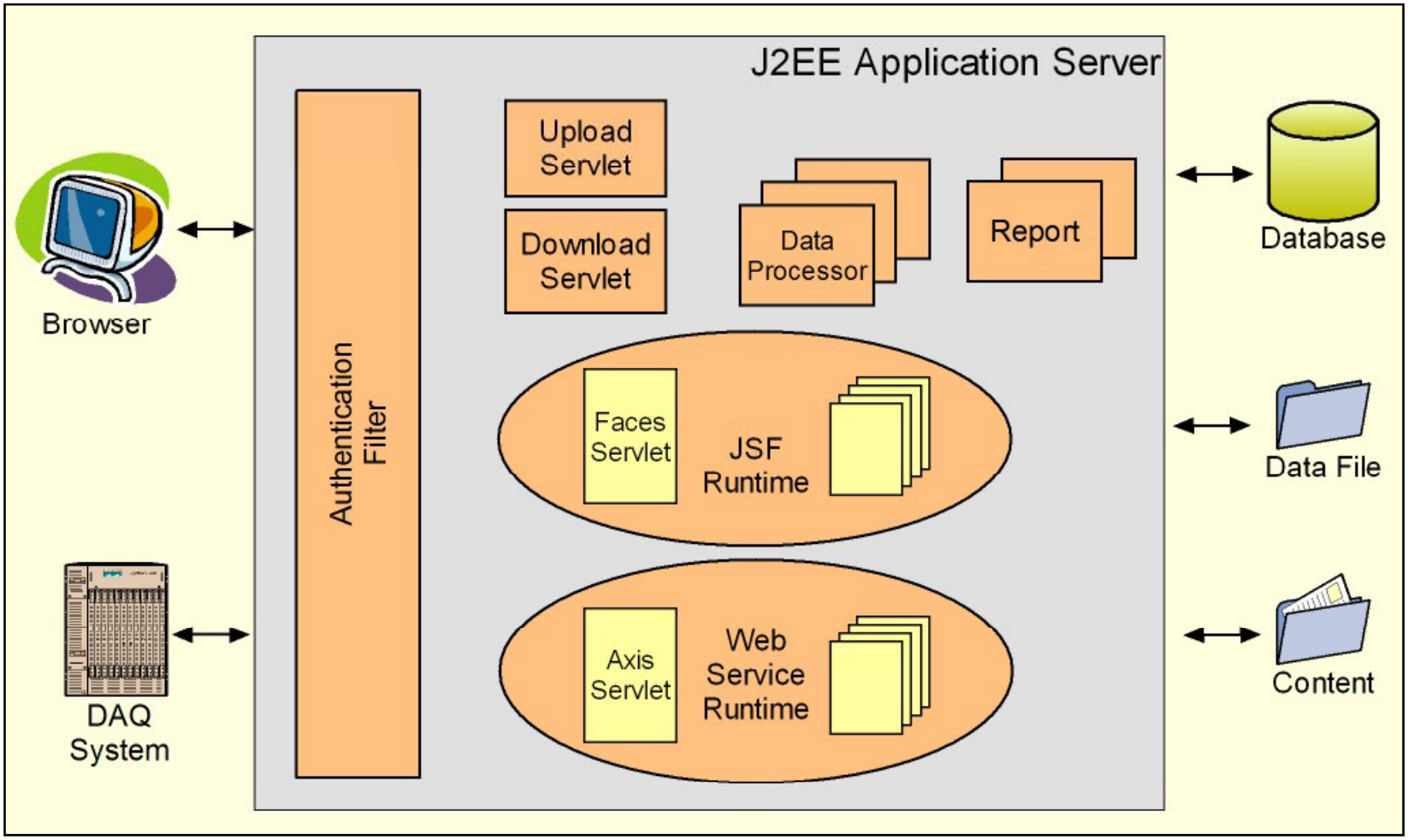
- Web-based interface (interactive access)
  - Insert (upload) data (drag & drop or browsing)
  - Search for tests based on metadata (location, DAQ, test subject, tags, attributes)
  - Download data
  - View comments, preview data
  - View reports
- Web service (programmatic access)
  - Upload of data
  - Query for test information
  - Data retrieval

Automatic Processing

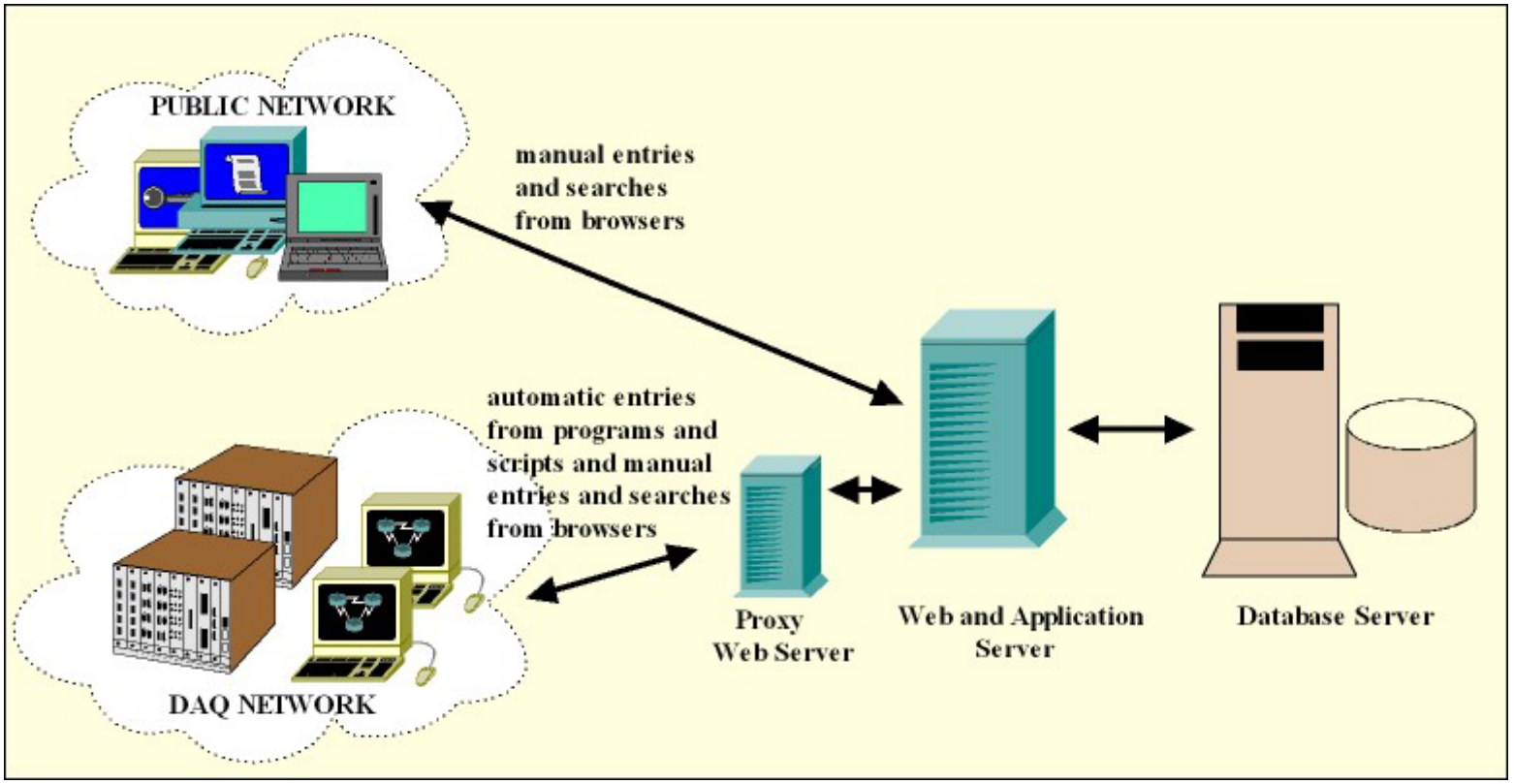
- Automatic processing of data upon upload (pre-processing). Examples include:
  - Format change
  - Format and contents verification
  - Automatic analysis
  - Insertion to a database
  - Compress data
- Automatic processing of data upon download (post-processing). Examples include:
  - Export/conversion to CSV or Excel format
  - Uncompress data

II. WebDat: Implementation

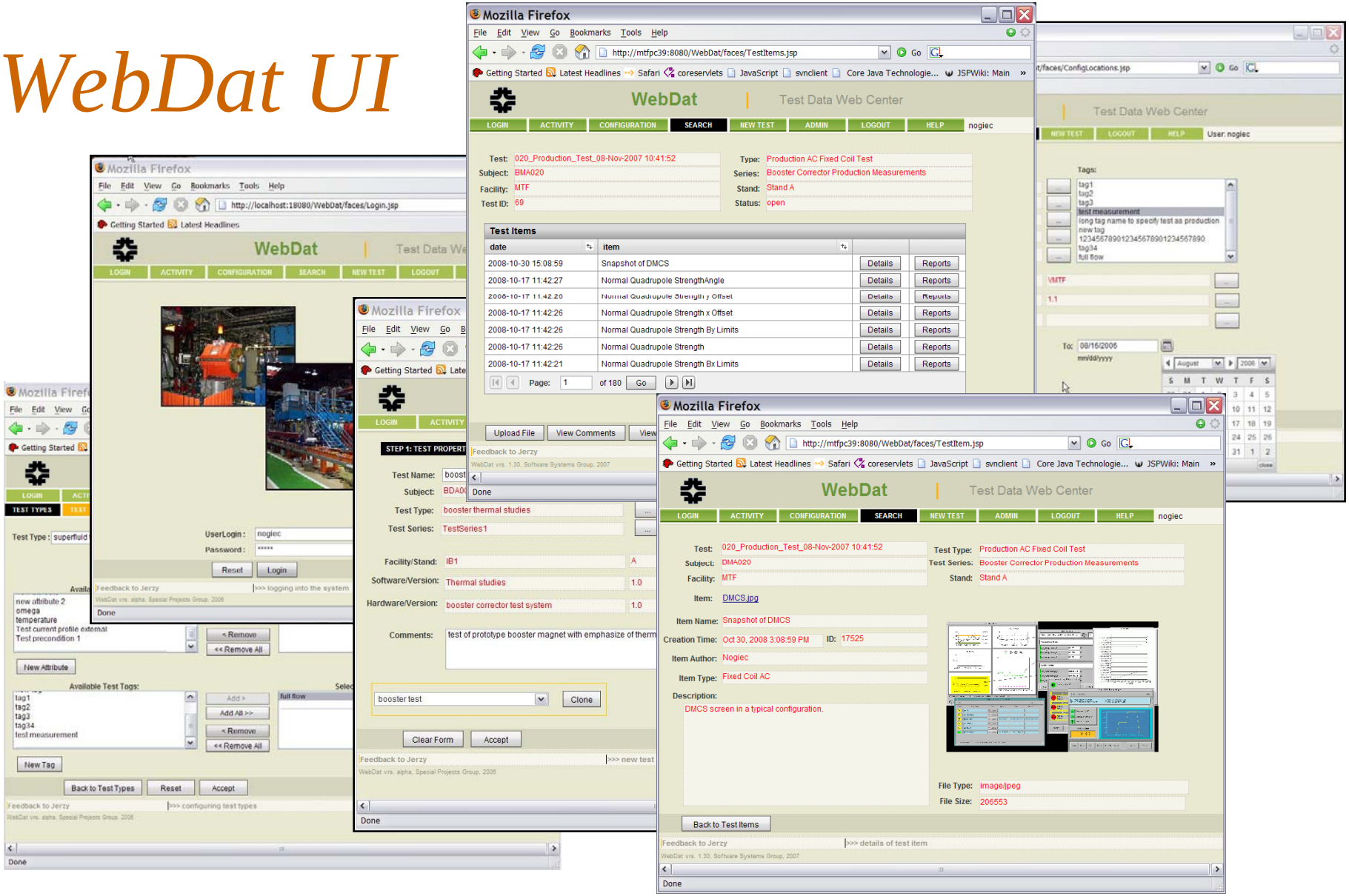
Architecture



WebDat Deployment

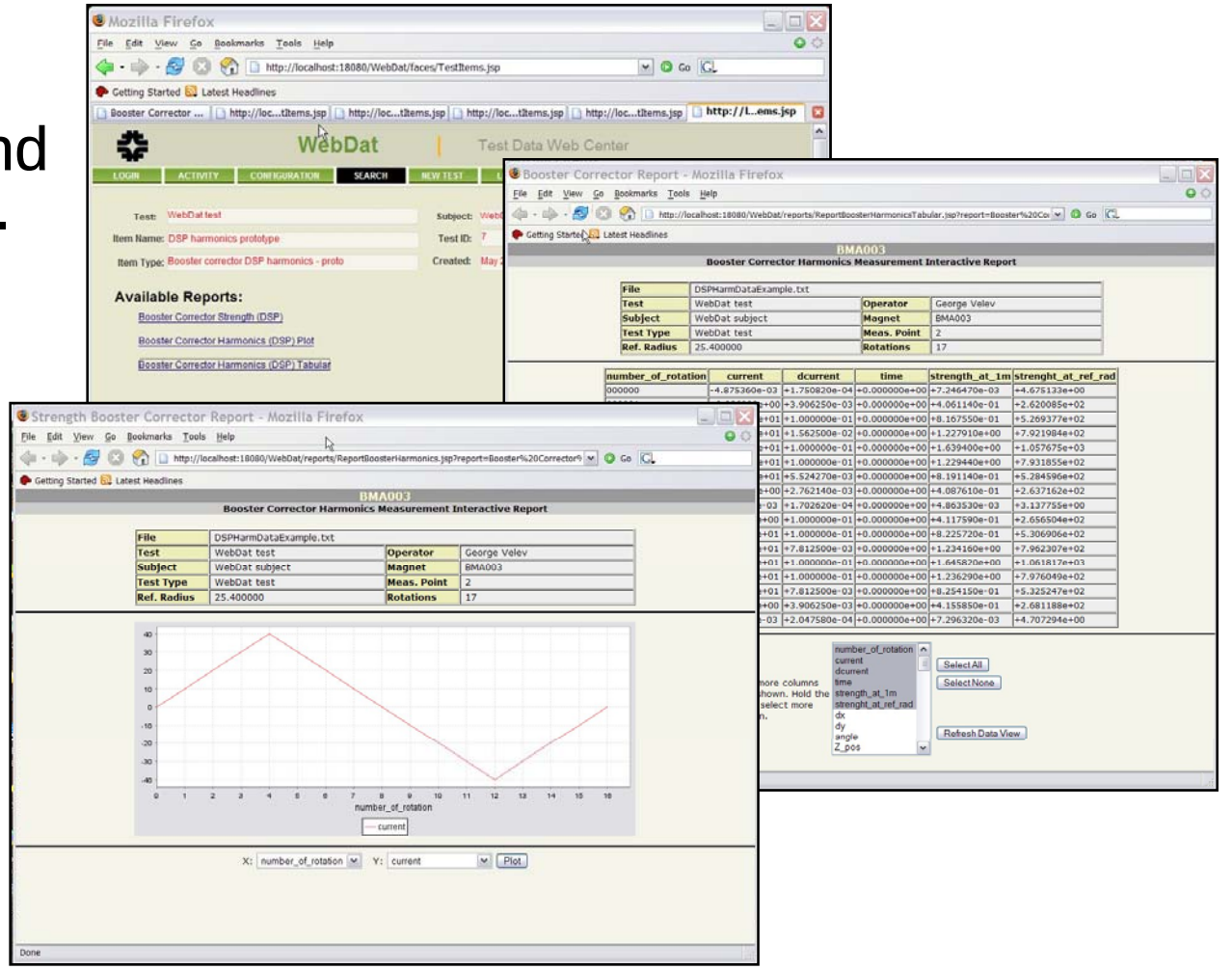


WebDat UI


Mozilla Firefox
WebDat
Test Data Web Center
LOGIN
ACTIVITY
CONFIGURATION
SEARCH
NEW TEST
ADMIN
LOGOUT
HELP
Test Items
Upload File
View Comments
Back to Test Items
Clone
Accept
Reset
Login

Reports

- Interactive and static reports.
- Ability to produce reports from file and database sources.



Other Features of WebDat

- Secure authentication and role-based authorization.
- Administration of users and groups.
- Test lifecycle management.
- On-the-fly compression of data.

Technologies

- JavaServer Faces
- MySQL database
- JasperReports
- Apache/Tomcat
- Apache Axis (Web service)
- Eclipse & Sun Java Studio Creator

Summary

- There is a need and interest in uniform management of structured and unstructured data.
- R&D environment is one of the areas that would benefit from such approach.
- WebDat addresses this need by providing a system based on a common set of metadata for structured and unstructured data.
- WebDat provides consistent access to file and database data sources.
- WebDat provides for interactive (JSF-based UI) and programmatic (Web service) insertion and retrieval of data as well as pre and post processing of data.