A Digital Archive of Monitoring Data

Fábio Costa
Faculty of Engineering, University of Porto
fabiopcosta@fe.up.pt

Gabriel David
INESC TEC, Faculty of Engineering, University of Porto
gtd@fe.up.pt

Álvaro Cunha
Faculty of Engineering, University of Porto
acunha@fe.up.pt

Rua Dr. Roberto Frias 4200-465 Porto, Portugal
+351-225081400

ABSTRACT
The change of status of data files from mere stepping stones to build other research products into publishable documents raises the question of how to organize data repositories appropriate for dissemination of such publications outside of the original research group. If the repository is to be used for the on-going research by the research group, it assumes the role of a digital archive. In this paper, a metadata model for the special case of projects relying on monitoring data is proposed and a prototype digital archive is described that has been built according to that model. This metadata is critical to preserve the context of production of the data at the organizational and technical levels and the meaning of each value. The digital archive offers several services for ingestion, visualization and dissemination that are essential for the effective adoption of the system. The method followed has been focus group work with a research group on structural health monitoring during the metadata specification phase, and an iterative development approach during the prototype construction phase of a digital archive for the same group.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: Collection, Dissemination, Systems issues

General Terms
Documentation, Design.

Keywords
scientific data repositories, experimental data streams, structural health monitoring.

1. INTRODUCTION
The investment and amount of effort put in setting up scientific experiments and collecting data from them fully justify that these data sets be properly preserved and eventually made available to the scientific community. This way, the research results may be cross-checked and used by other researchers for further investigations in what is being called e-science.

Due to the size of many of these data sets, it makes sense to organize their publication in data repositories, along with the required metadata to assert their meaning and authenticity. The metadata comprises contextual aspects on the entities involved and the purpose of data production, on authorship, on details about the scientific accuracy, on technical aspects of the digital support, and on integrity and preservation. The Core Scientific Metadata Model [5] covers most of these aspects. However, due to the diversity of scientific data, it is rather complex and may hinder the essential cooperation of the researchers in contributing the metadata elements.

The purpose of this paper[1] is to present a simplified model for monitoring data, which has been developed in dialogue with a team working on structural health monitoring; and a digital archive designed according to it, which is now being used as the research group's main data repository.

2. MAIN PROJECT GOALS
The data collection phase in research activity has mostly been considered a private concern of each project while the papers, reports and prototypes were the sole outcomes deserving to be published. Therefore, the collected data sets were organized in nonsystematic ways and, after being used, they were kept in the personal backups of the researchers and eventually discarded.

This understanding has been changing for several reasons. Some experiments are so expensive that it is not feasible to replicate them, as happens with high energy physics. The recording of natural phenomena, as in astronomy, is in many cases inherently unique. The advances in data acquisition systems led to the availability of huge data sets, in parallel with the capacity to process them. The development of the Internet turned the cooperation of research teams practical. All this has represented a strong push towards sharing not only the research results but also the data sets across the Internet. The creation of the Web itself has been a response to the need for cooperation in scientific research. Following the trend, several funding agencies adopted the policy of requiring the publication of the data sets produced within funded projects, which became research outcomes themselves [1]. The expertise required to properly design the experiments, decide and install the equipment and clean the data from defects and abnormal conditions in the acquisition is so high that the data sets can be seen as being authored by the researchers in charge of those tasks. Adding authorship to the data sets is a way to raise the personal responsibility of the researchers in properly taking care of the data sets, rewarding them by acknowledging their role in these scientific outcomes, and increasing the contemporary and future trust the data sets deserve.

Publishing means that the data sets will outspan the projects where they were born and even the research group. To make the

[1] Research supported by project "DYNAMO - Advanced Tools for Dynamic Structural Health Monitoring of Bridges and Special Structures", PTDC_ECM_109862_2009, funded by FCT/MCTES (PIDDAC) and FEDER through COMPETE/POFC. Gabriel David is co-financed by the North Portugal Regional Operational Programme (ON.2 – O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF).

data usable by researchers with no direct knowledge of the originating project circumstances, enough metadata must be added to the data sets. The metadata must fulfill two main roles: adding to the meaning of the data and improving the confidence on its accuracy and authenticity. So, it must include not only descriptions of the variables and measurement units, but also information on the process and equipment for data acquisition, on the researcher in charge of each phase of the project, and on the host institution, and details on the events that may affect the interpretation of the data, on the integrity of the data and on the processing the data may have been subject to.

Trying to encompass the diversity of situations that may occur in different research projects, leads to a complex metadata model. The context of the current paper is a research group working on structural health monitoring. The projects involve monitoring structures like bridges and other large Civil Engineering structures and their environmental and operating conditions. The monitoring is done through carefully designed and installed data acquisition units able to record, for instance, accelerations, temperatures, or wind speed. The data is collected in files every 30 minutes and sent by a data link to a computer of the research group. These raw data files are then pre-processed in order to clean possible malfunctioning situations and the cleaned files are also stored. One or more sophisticated algorithms are afterwards applied to the latter to calculate the evolution of relevant dynamic characteristics. The whole process may last for several years, resulting in a large number of similar and relatively simple files. The data in the results or processed files may be more complex, but it can be recalculated.

The importance of structural health monitoring is manifold. Keeping track of the behavior of bridges, dams, or large building under actual operating conditions of load, wind, earthquakes, etc. is important to study those structures and prevent incidents, to detect the effect of ageing and to help on repairing and compensating.

The main project goals are: (1) specify a metadata model; (2) design and build a digital archive, according to that model, able to store and organize the monitoring data as well as the processed results of the on-going projects; (3) improve data reliability through an integrated backup strategy; (4) create a Web interface able to browse and search the digital archive metadata and to visualize the data and download it; (5) set up a simple user management system and an access control policy; (6) automate the ingestion procedure of the data files into the digital archive.

Attaining these goals means a more reliable and systematic data life cycle, reduced researcher time on data management activities, support for data sharing in research cooperation, and a way to fulfill possible requirements of data publication. Furthermore, several important steps are taken towards preservation when insisting on collecting contextual and technical metadata and on organizing data in a systematic way.

3. METADATA MODEL

The metadata model is organized in three levels: Context, System and Data. There are two support classes related to the three levels, namely Person and Document. The Context package represents the information on the project itself and the hosting institution as well as the target structure being monitored. The project designs and installs specific data acquisition systems and chooses or develops specific software, both of which are described in the System package. Each data stream coming from an acquisition system is described in the Data package that also records the corresponding list of data files. The data files are organized in the file system.

According to the relevance given to authorship and good documentation, the support classes Person and Document are omnipresent in the model. It is possible to document the project and the structure, the data acquisition system, and the data sets with several types of documents, including technical descriptions, papers, pictures and diagrams. The documents have authors but, besides authorship, persons are associated to several components of the model, under different roles.

3.1 Context level

Contextual information (see Figure 1) is essential to know who and why has produced the data and under which circumstances.

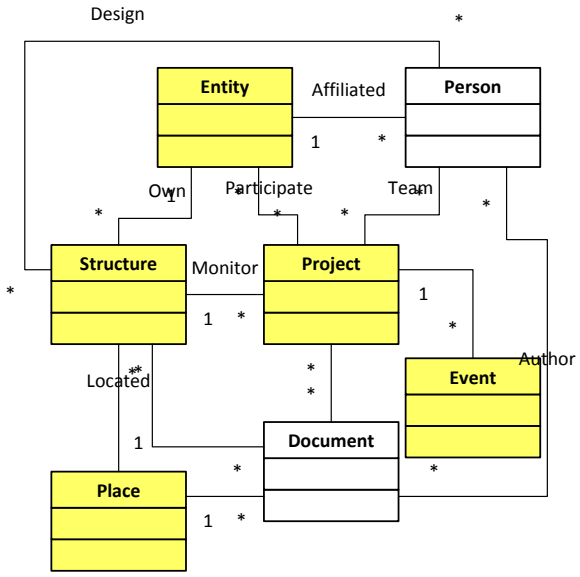

Figure 1. Contextual metadata.

All data sets are organized inside a project. So the central class is Project, including identification, type, life span and related entities. These entities (class Entity) may play different roles, like participating entity, funding agency, or owner of the monitored structure. The Project references a single Structure it monitors. The class Structure gathers information about the monitored object, like identification, building date, description, owner, location, and designers.

To improve on contextual information, the class Place records addresses of entities and geo-references structures and documents, in particular, pictures. There is also a class Event associated to Project that is meant to record any kind of event that may affect the monitored object or the data streams. Examples of events are earthquakes and strong winds, but also power shortages or maintenance actions on data acquisition systems. Events have an interval to enable setting a window on the data streams.

Structures and Projects may have associated documents of several types. Any kind of document may be associated but the recommended formats are PDF, PNG, JPEG and TIFF or any open document format, for preservation reasons. Besides the title, description and dates of creation and last update, technical details are recorded like the generating application, file type and size. It is possible to associate a place, especially meaningful for pictures. Documents relevant for structures are design summaries,

historical notices, and illustrative pictures. With respect to projects, the project proposal and a global diagram of the monitoring approach can be helpful. Each document has usually one or more authors, who are represented in the class Person.

Like documents, persons connect to the model in several points. Persons have identification and contacts and are affiliated to one entity. They are authors of documents and designers of the structures. And they group in teams for each project, with a certain role and during an interval. Persons belonging to teams and other designated collaborators have access credentials.

3.2 System level

The second level is about the technical system information (see Figure 2).

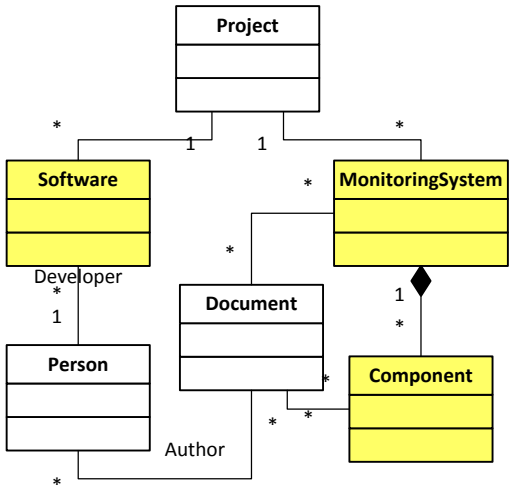

Figure 2. System metadata.

A project may use one or more data acquisition systems, simultaneously or in different periods. The main class is MonitoringSystem and it identifies a data acquisition system, its manufacturer and supplier, the period of operation and includes a description. A single data acquisition system usually possesses several transducers or data acquisition devices, with specific characteristics that are relevant to physically interpret the data streams. So, a Component class provides the details for each data source, like type, manufacturer, configuration, and positioning. Monitoring systems and components may have associated documents, like detailed installation diagrams, data sheets, or operating instructions.

Very often, the data acquisition process includes processing steps using commercial tools or specifically developed algorithms. The corresponding information, recorded at the system level in class Software, is the type, product, version, and the manufacturer or the developer, along with a description of its function.

3.3 Data level

The third level describes the data streams produced by the systems of the second level (see Figure 3). Each data stream is associated to a monitoring system and is represented by the Dataset class. It contains attributes describing the data stream, the acquisition method and the specific parameters used to obtain it. There is also a description of a possible processing step and a reference to the corresponding software. The intended results are summarized. A set of temporal attributes is also included like the period from one data file to the next, the number of files per day, and the sampling period and frequency. Some summary attributes include the time of the first and last reading in the data stream and the daily and global data volumes. The actual data location is recorded in the directory attribute. The types of foreseen data streams are: raw, for the data files as they are received from the acquisition system; pre-processed data streams correspond to a cleaned version of the data, after spurious values have been removed or errors have been fixed; and results data streams are those obtained by applying specific algorithms. The implicit genealogy of data streams is recorded in a many-to-many association.

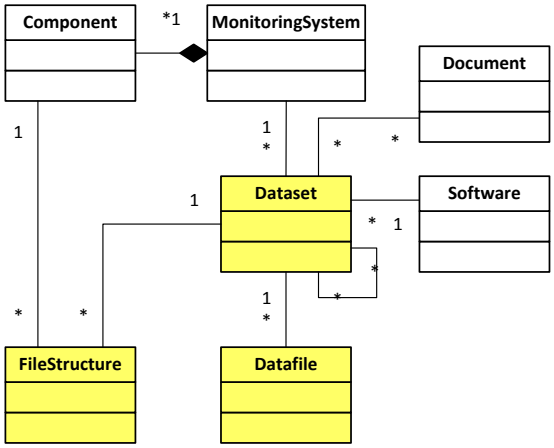

Figure 3. Data stream metadata.

A single data stream corresponds to a sequence of data files, all with the same record structure. The structure of a record is a sequence of columns. Each column is described by the class FileStructure. A column is associated to the channel of the component of the monitoring system producing that specific variable. The column has a number, two names (allowing for grouping similar columns), information on the type of variable and measurement unit, and a data type (integer, float, double, string). There is also an optional description. This information should be enough to understand and process the data files.

Finally, the Datafile class keeps track of the actual data files for each data stream. The main attributes are the filename, the file type, the file creation date, the start and end timestamps, the number of records, the file size and the compressed file size, a status (if the file is damaged) and a comment.

4. THE DIGITAL ARCHIVE

The metadata model of the previous section has been tested in the development of a digital archive for a research group in structural health monitoring[2]. The typical situation for projects in this area is to produce tens of thousands of datafiles.

4.1 Repository organization

The first problem on building a digital archive for these volumes is to establish an organization for the files. The decision has been to have a root directory for the digital archive, with a subdirectory for each project containing one directory for the documents associated to the project and one directory for each data stream. The latter is then hierarchically organized with directories for each year, each month and each day.

[2] ViBEST: http://paginas.fe.up.pt/vibest

4.2 Services in the prototype

The prototype that has been built is now in the first phase of use[3]. It implements the metadata model but in a certain sense is more than a data repository as it includes several services helping the research group to manage large amounts of data, use them in their day-to-day research and make them accessible to external researchers.

The technology used is the Postgres database management system, the Vaadin framework for Java Web applications [2], the Apache http server running on Ubuntu operating system and a few libraries for specific operations. The application has a Web interface automatically displaying pictures of each of the structures being monitored.

The generic information about each structure and its monitoring projects is publicly available. To go into more details about the monitoring systems and the data streams, authentication is required. So a simple user management module has been implemented, granting access rights at the project level.

In order to allow for manual as well as automatic ingestion of the data files, a background job has been created that periodically checks whether new data files have been received after the last checkpoint in each data stream and updates accordingly the records on the data files.

The design of the interface follows a compact style, trying to concentrate the most information on a single page. So, there is one page for the project and the corresponding structure, another for the monitoring system and a third one for the data stream (see Figure 4).

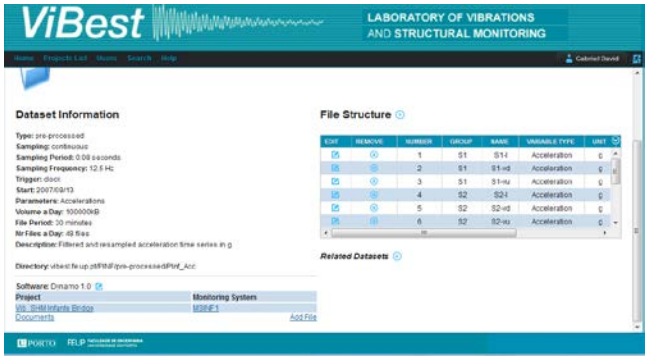

Figure 4. A data stream page.

However, due to the typical number of data files, these are only accessible through a selection form to choose an interval or an associated event. When arriving at a specific set of data files two options are given to the user: download them or visualize the data. The visualization option includes selecting which columns in the data files are to be included in a graph.

Two more aspects, related to dissemination, deserve mention. Due to the large number of files, downloading them one at a time is not feasible. So, a zipping facility has been prepared to combine all or, at least, chunks of the selected data files on a single zip file.

The second aspect is the addition of the OAI-PMH protocol [4] to enable the aggregation of the information in the digital archive by specialized repositories. A Dublin Core view on the metadata has thus been defined to support interoperability. This protocol works fine with the public information on projects and structures. However, the policy of the archive requires authentication for access to the second and third levels of metadata and to the data files. So, an extension to the OAI-PMH protocol has been prepared to allow authenticated users to keep using it at the level of data sets and data files.

At the same time that it improves the current research conditions, the prototype sets up the conditions for some preservation steps. The raw and pre-processed data files are zipped text files that will remain straightforwardly accessible. The metadata explaining the meaning and units of each variable, the sampling conditions, and the context of the experiment is collected in a relational database. The metadata is then preserved using the SIARD Suite [3] to convert the database into an XML representation.

5. CONCLUSIONS

The goals set up for the project have been achieved. In particular, the metadata model, although a bit demanding for the researchers asked to input the required information, proved to be enough to describe monitoring data for special Civil Engineering structures. As just few specific details of the structural health monitoring area have been used, the model is believed to be useful for general monitoring data projects.

With respect to the digital archive application, the size of the problem prevented the use of naïve approaches and forced some fine tuning of the http and application servers.

The main point still under analysis is related to the visualization of result data files not in a tabular format (scattered graphs, two or three dimensional matrices, etc.). Although it is possible to store these files in corresponding data streams, more work is needed in order to find an appropriate description for those data files. Probably, an XML representation will be chosen. A mechanism to visualize the diverse data formats needs to be devised.

6. REFERENCES

[1] European Commission. 2012. *Scientific data: open access to research results will boost Europe's innovation capacity*, press release IP-12-790. Brussels 2012-07-17. http://europa.eu/rapid/press-release_IP-12-790_en.htm?locale=en

[2] Grönroos, M. 2012. *Book of Vaadin*, 4th ed., pp. 466, Vaadin Ltd, Finland. https://vaadin.com/download/book-of-vaadin/vaadin-6/pdf/book-of-vaadin-pocket.pdf

[3] Heuscher, S., Järmann, S., Keller-Marxer, P. and Möhle, F., 2004. *Providing authentic long-term archival access to complex relational data.* In *Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data*. European Space Agency.

[4] Lagoze, C., Sompel, H., Nelson, M. and Warner, S. 2002. *The Open Archives Initiative Protocol for Metadata Harvesting, Protocol Version 2.0 of 2002-06-14.* Document Version 2008-12-07T20:42:00Z. OAI. http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

[5] Sufi, S. and Mathews, B. 2004. *CCLRC Scientific Metadata Model : Version 2*, CCLRC Technical report DL-TR-2004-001, 2004. http://epubs.stfc.ac.uk/bitstream/485/csmdm.version-2.pdf

[3] Digital archive URL: http://vibest.fe.up.pt/shm