Healthcare Dataset Integration and Management

Gerry Reilly?, Adam Milward, David Milward, Courtney Irwin and Susheel Varma?

HDR UK / MDW / OU.

Email: david@metadataworks.co.uk

**Abstract.** Metadata is key in library systems in helping to search and organise data. The same is true for research datasets, however automated data catalogues are a relatively new innovation and it is only recently that key medical research centres in the UK have been able to integrate their research data under the remit of Health Data Research UK. In carrying out this work metadata has been imported from 37 different sites into a central data catalogue to make it available for all. This has resulted in the development of a new schema for health metadata and new methods for integrating the descriptive and technical metadata. This paper describes our approach and result in carrying out this task.

1. Introduction

An adequate account of data semantics and provenance is important in data management and analysis: to facilitate re-use and help ensure compliance. This importance increases with the value and the complexity of the data. This paper shows how a simple metadata modelling language and metamodel has been used to capture and implement a new descriptive metadata schema, enabling 37 research organisations to share datasets using a centralised metadata registry. In addition we document the key points needed to federate the system into a federated metadata network, enabling new members to join and allowing different schema types to be updated and managed in real-time.

A fundamental problem in the UK and elsewhere is how to make clean data from heterogeneous healthcare datasets available to researchers easily and quickly. Much of this information is from research sources such as clinical trials and electronic patient records, however currently researchers can spend more time wrangling and cleaning the data than is spent in analysis, some reports put this at between 60-80% of time taken in analysis tasks[2]. Analysis generally involves python and R scripts which are unique to the researcher that wrote them, if new information is not in the form that the researcher anticipated then the text has to be re-written, and checked for accuracy. In an ideal world data would be input to a data warehouse in a form that allows the same query to be run repeatedly, because the data that comes into the repository is guaranteed to be in the same format, and if the format changes the dataset and query can be updated to take account of this without a lengthy script re-write.

One approach to tackling this problem is the use of the standardized dataset, the idea being is that a set of data items are defined in the standard, and all reporting of any such data-items is made to confirm to the standard. This will enforce some simple rules such as a patient identifier in the NHS needs to be an integer of a certain length, conforming to a certain set of rules, very often encoded with a regular expression. This approach helps enormously, but its application hasn't been entirely successful to date for several reasons. Firstly, it is impossible currently to mandate that everyone uses the same standard, or set of standards. Secondly, some standards are strong in some areas and weak in others. Thirdly standards evolve.

1. Background

Metadata registries are becoming commonplace in data warehouses, in order to classify and help enhance search and find capabilities. Traditionally, healthcare data has relied on multiple diverse datasets being integrated by domain experts. These integration processes are executed with a high degree of human involvement, integrating datasets can be time-consuming and can result in the introduction of errors into the data. This paper describes work to build an integration toolset for healthcare datasets based on the DCAT metadata standard and drawing elements of the ISO11179 Standard for metadata registries in addition.

Please follow these instructions as carefully as possible so all articles within a conference have the same style to the title page. This paragraph follows a section title so it should not be indented.

* 1. A
  2. B
  3. C

1. Purpose

One of the main purposes behind this work is to allow researchers to see what datasets are available from member organisations, to examine the metadata of these datasets to establish if it is worth downloading the dataset, to see what constraints are attached to the datasets and establish whether it is worth downloading a particular dataset for use in a particular study.

Researchers do not want to have to examine each dataset in depth before deciding whether to use the dataset, they need a way of accessing the key attributes of a dataset remotely so that a decision can be made before getting to the detail of the data itself.

1. Approach

The use of sections to divide the text of the paper is optional and left as a decision for the author. Where the author wishes to divide the paper into sections the formatting shown in table 2 should be used.

* 1. Basic Descriptive Metadata Schema
  2. Basic Technical Metadata Schema

1. Results

Footnotes

1. Discussion and Conclusion

Each .



1. References

[1] Aderhold J, Davydov V Yu, Fedler F, Klausing H, Mistele D, Rotter T, Semchinova O, Stemmer J and Graul J 2001 *J. Cryst. Growth* **222** 701

[1] Information technology — Metadata registries (MDR ISO/IEC 11179 – 1) — Part 1: Framework

[2] Information technology — Metadata registries (MDR ISO/IEC 11179 – 3) — Part 3: Registry metamodel and basic attributes.

[3] Josefsson, S., “The Base16, Base32, and Base64 Data Encodings,” RFC 4648, October 2006 (TXT).

[4] Williams, N., “On the Use of Channel Bindings to Secure Channels,” RFC 5056, November 2007 (TXT).

[5] Altman, J., Williams, N., and L. Zhu, “Channel Bindings for TLS,” RFC 5929, July 2010 (TXT).