Standardising geochronological data

Keith Sircombe
Geoscience Australia
Cnr. Jerrabomberra Ave and Hindmarsh Drive, Symonston, ACT 2609
Keith.Sircombe@ga.gov.au

SUMMARY

Although there are several resources for storing and accessing geochronological data, there is no standard format for exchanging geochronology data among users. Current systems are an inefficient mixture of comma delimited text files, Excel spreadsheets and PDFs that assume prior specialist knowledge and force the user to laboriously – and potentially erroneously – extract the required data manually. With increasing demands for data interoperability this situation is becoming intolerable not only among researchers, but also at the funding agency level.

Geoscience Australia and partners are developing a standard data exchange format for geochronological data based on XML (eXtensible Markup Language) technology that has been demonstrated in other geological data applications and is an important aspect of emerging international geoscience data format standards. This presentation will discuss developments at Geoscience Australia and the opportunities for participation.

Key words: Geochronology, data management, metadata, standards.

INTRODUCTION

Geochronological data is a fundamental part of geoscience research. Radiometric methods provide an absolute age of geological events that enables the assembly of geological models on all scales from outcrop to continents. As data-rich integration becomes the norm for geological research, the need to precisely describe and exchange those data between researchers, service providers and industry clients is becoming critical. On a broader scientific level, research funding agencies within the Australian Government are also beginning to examine how the output from publicly funded science is managed in order to maximise the utility of those data in the emerging e-Research framework (Arzberger et al., 2004; Department of Education, Science and Training, 2005).

A number of resources now exist to provide geochronology data, e.g. the national geochronology database OZCHRON (http://www.ga.gov.au/oracle/ozchron/frames.html) and related state survey databases (http://www.geoscience.gov.au/). However, current transmission of geochronology data from those resources is often inefficient and troublesome. Enormous amounts of data are effectively inaccessible because their current format prevents easy sharing among researchers and industry clients.

The development of a geochronology data standard is a subset of other standardisation projects across geosciences such as the XMML (eXtensible Mining Markup Language, https://www.seegrid.csiro.au/twiki/bin/view/Xmml/WebHome) and GeoSciML and will interact with related projects to build a fully integrated standard.

THE PROBLEM

The current exchange of geochronological data is grossly inefficient. This inefficiency occurs at two levels: mode of transmission (e.g. file types), and data semantics (e.g. the 'headers'). The inefficiencies also cause large data management demands with the need for format translators and maintenance requirements.

There is currently no broad definition of data labels (e.g. 'column headers') beyond local usage. In some cases the data labels are left out of a data file entirely on the assumption that end-user already knows the format. For instance, a simple label like "*6/38*" may mean the raw measured ^{206}Pb/^{238}U ratio to one user, a corrected ^{206}Pb/^{238}U ratio value to another user or even the determined age based on the ^{206}Pb/^{238}U ratio to yet another user. A major component of the standardisation project will be establishing a standard format where labels have a precise definition agreed by consensus in the geochronology community.

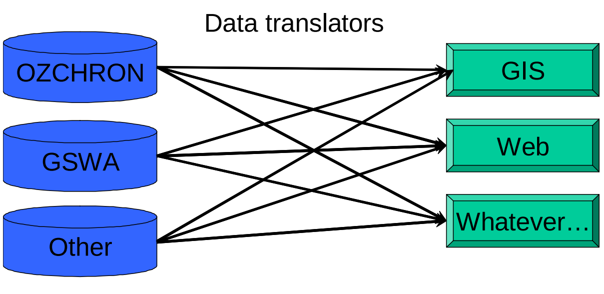

Figure 1. Illustration of the current complex arrangement between geochronology databases on the left and potential applications using geochronology data on the right.

The current lack of standards means that transmission of data from a database to an application requires an individual data translator for each pairing (Figure 1). In current usage many of these translators are in effect manual with a user cutting and pasting data between files, spreadsheets and applications. These translation procedures are often created ad-hoc and are rarely documented leading to inefficiencies and reduced productivity. Although in some instances these translators can be automated on a localised level with macros, this introduces maintenance overheads where the translators require updating whenever the relevant file formats and applications change. At present, a seemingly minor change in database output or application input requirements can have a significant impact.

PROPOSED SOLUTION

The solution proposed by Geoscience Australia is the establishment of a community agreed standard format for geochronology data. The format would employ XML technology to embed information about the data (e.g. the column headers) within the data text file.

For instance, a geochronology XML datum may look like:

<206Pb-238U-corr>*0.4598***</206Pb-238U-corr>**

which would indicate that the number 0.4598 is the corrected ^{206}Pb/^{238}U ratio. A database programmer can then create a process by which the equivalent of that field in their database, say "*206-238c*" in the "*U-Pb*" table, is mapped to a **<206Pb-238U-corr>** field when a user asks for that data from the database. Likewise the developer of the user's application (e.g. a Geographic Information System, GIS) can program the application to look for the **<206Pb-238U-corr>** field in the data and process it, for instance, and use the ratio to calculate a determined age to plot alongside the sample location on map.

With such a common data format, the task of maintaining the links between databases and applications is also simplified (Figure 2) so geochronologists can get on with the job they were trained to do – analysing samples and interpreting the geological meaning of those analyses.

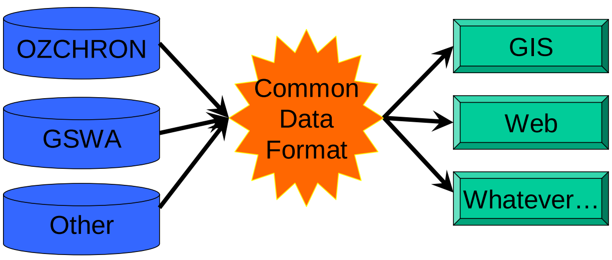

Figure 2. Illustration of how a standard data format would simplify data transmission between databases and applications.

Standard data formats are built around schemas, which are in effect online dictionaries held at a commonly accessible point on the internet. As well as listing the fields available for use, these schemas also aid data quality control. For instance, the number in a **<206Pb-238U-corr>** field should always be positive and this can be defined in the schema and any data file can be checked for compliance.

The formal definitions established in schemas can also formalise the conceptual relationships among data. Searching for and defining these commonalities is a critical component of standards development. This is will particularly be the case for geochronology where a bewildering array of systems and methods are united by the radioactive decay equation. Thus it is vital for all geochronological methods to be involved in developments.

THE STANDARDISATION PROCESS

Standard data formats in XML schemas are created by consensus within a user community. In the case of the geochronology community this is complicated by the broad range of methods and uses, but is nonetheless achievable. The steps can be broadly described as follows:

1. Establish User Community
2. Gather User Requirements
3. Consideration and consensus
4. Modelling and programming
5. Iteration and communication
6. Implement and test

The standardisation project at this stage is in the early phases of steps 1 and 2. Geoscience Australia welcomes any input from the geochronology community and the broader geoscience community that uses geochronology data.

Development of a geochronology XML is also closely related to developments of standards across geoscience in general. For example, a XMML model already exists for defining the geological timescale and Global Stratotype Sections and Points (Cox and Richard, *in press*). Classes of data in this model will describe Date Determinations and Dating Procedures.

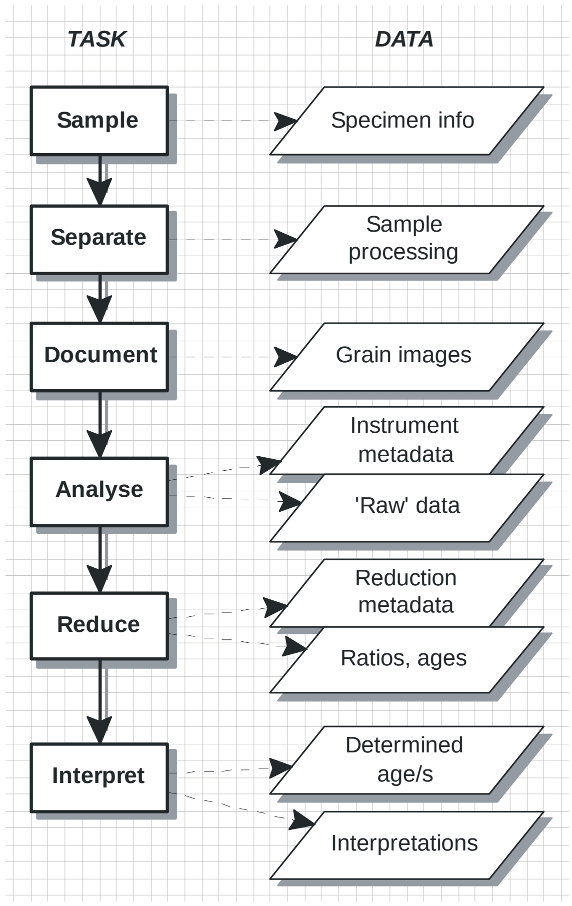

Figure 3. A typical geochronology analytical workflow illustrating the types of data generated.

FIRST STEPS

Geochronological analyses generate a broad range of data as illustrated in Figure 3. Many of these classes of data, e.g. specimen information, are common to all geoscience disciplines and related data standardisation projects are already underway to develop the appropriate standards.

After consultation with stakeholders, the approach Geoscience Australia is following is to begin with building a community consensus about XML field tags as typically seen in published geochronology data. This is in effect the *Reduce* and *Interpret* steps of Figure 3 and is starting on the 'outside' of the geochronology data workflow and working toward the details of analysis. From this development commonalities will emerge that will aid the future development of more generic schemas for all of geochronology.

REFERENCES

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P. and Wouters, P., 2004, Promoting Access to Public Research Data for Scientific, Economic, and Social Development. Data Science Journal, 3, 135-152. (http://journals.eecs.qub.ac.uk/codata/journal/Contents/3_04/3_04pdfs/DS377.pdf , accessed Sept. 2005)

Cox, S.J.D. and Richard, S.M., *in press*, A formal model for the geologic timescale and GSSP, compatible with geospatial information transfer standards; Geological Society of America, Special Paper, Cyberinfrastructure for Earth history: challenges and promises.

Department of Education, Science and Training, 2005, An E-Research Strategic Framework: Interim Report: http://www.dest.gov.au/NR/rdonlyres/9526F5BC-B079-4903-AE0B-87140165F37F/8496/InterimReport.pdf (accessed Dec. 2005).