Managing, manipulating and preserving data…for always

Gavin Stilgoe

Gavin Stilgoe[1], Catherine Huf[2]

[1]Department of Jobs, Precincts and Regions, Melbourne, Australia, gavin.stilgoe@djpr.vic.gov.au
[2]Department of Jobs, Precincts and Regions, Warrnambool, Australia, catherine.huf@djpr.vic.gov.au

THE PROJECT

The project was improving the FAIR of information created by Geological Survey of Victoria (GSV). The resource funding was included as part of the Victorian Gas Program (VGP). Information created as part of the VGP was used as a case study for the GSV way forward. This project was conducted by the Geoscience Information (GI) team. The system used for long-term storage of tabular data (in this instance laboratory data analysed internally and externally as well as observations conducted by geologists) was an Open-road system created and maintained by GSV.

MANAGING

Improved information management (IM) at GSV included cultural change, high management approval, policy creation and the implementation of data management plans (DMP).

Cultural Change

Cultural change included a variety of approaches both subtle and specific and visible. Cultural change included: formation of an IM Working Group (IMWG) as a mechanism for discussions; introduction of information management terms into day-to-day vocab; creation of IM policy, including a policy for DMPs.

High-level management approval

High-level management approval was also both subtle and specific and visible. The director started with "scientific evidence that is findable and reusable" and eventually became familiar with the term FAIR. The term FAIR is now used freely by high-level management. IM policy is ultimately to be approved by the Director. IM policy must be driven by other high-level management (outside of GI team) and hence various policy currently sits in draft until seen as needed by high-level management. It is a case of them identifying a need and then GI team delivering "yeah we can do that, here's one we prepared earlier".

Data Management Plans

DMPs were identified as being crucial for achieving the aim of FAIR. An initial template included everything needed to achieve a FAIR result, however this overwhelmed people outside of the IM area. Hence the DMP was cut back to include information to the point of publication, not for long-term storage etc. This resulted in resistance dissolving and successful collaboration occurring.

MANIPULATING

Tools

Two freely available ETL tools were trialed, namely Metabase and Streamsets.

Metabase was trialed first and was used to allow people outside of GI team. i.e. the geologists. Assistance was provided by the GI team to tailor the outputs to the needs of the geologists. Metabase was also used as a visual aid to help geologists determine what fields (existing and new) were needed for the importation of new tabular data into the existing system.

New tabular dataset

The VGP included a large amount of sampling. The sample numbers were large, the sample types were varied, and the samples were sent to 11 different laboratories. External laboratory analysis alone was expected to include over 250,000 chemistry results. There were also ~1000 palynology samples and ~100 stygofauna samples tested by external laboratories. Air samples were analysed internally.

The timeframe and resources at hand determined that one agreed format was needed to cover the tabular 20 data subsets of the VGP. The approach used was to collaborate to find the commonality. Collaborations included using the IMWG as a mechanism for conversation and Confluence, the wiki used by GSV. All VGP team members involved in sampling or use of sample results were invited to participate in the discussion. This was extremely successful and results in one agreed list, broken down into three types producing three templates: “location”, “sample” and “analysis”.

Location information was already largely in the system, as most samples used wells or boreholes already in the GSV system. The exception to this were the water samples. Sample information was started from scratch for the project (to avoid time wasting over terms) but was later altered to utilise many existing fields within the system. Examples of system information included the depth a sample was collected at, sample type (e.g. rock chip, water etc.) and project name. Analysis information included the analyte, the unit (e.g. mg/L), result, laboratory used, analysis method, lower limit of detection and comments.

The existing system did require a few new tables (for the water samples) as well as some new options for existing fields such as new laboratories, new analytes and new units of measure.

Three of the external laboratories were resourced enough to be able to send or resend results in the new “analysis template”.

PRESERVING

Tools used to help provide a common way of accessing and ingesting the data into our corporate RDBMS are Metabase and Streamsets both opensource packages.

Metabase allows for the creation of standard SQL based queries that can be made available for users via the web it can also provide API's and provide the data via JSON objects for B2B processing

Streamsets is a highly flexible ETL tool web based and has a very easily understood GUI pipeline creation screens and can perform quite complex query and data transformation actions during both data import and export. Streamsets is also capable of providing API's for B2B transactions. Streamsets allows us to ingest data from Web Services enabling us to utilize services external to GSV for access to IGSN (International Geo Sample Numbers) as we store the sample data into our systems.

The current process we use is to ingest the data from them templates and the data is then immediately available via the Metabase screens.

This method greatly reduces the requirements of specialist software or high level of programming ability to ingest and transform complex data.

REFERENCES

1. Streamsets available from https://streamsets.com
2. Metabase available from https://www.metabase.com
3. IGSN https://www.igsn.org