White Paper

Implementing corporate data standards in
the field using metadata-centric tools

Summary

A super major's data management group wanted to develop a generic solution for the adoption of their global data standards. It was intended that legacy Openworks projects would be measured for conformance and then migrated to fit with the company's new standards, with the minimum of disruption to operations and maximum automation of the process.

A prototype solution was sought, leveraging ETL Solutions' Transformation Manager product. The tool provides a largely automated method of migration that was successful in significantly reducing the time required to complete the end-to-end migration of the Openworks projects. Users are presented with a simple interface, consisting of Excel spreadsheets and reports. This enables projects to be quickly measured against the corporate standard, and errors or deviations to be rectified. Automatic generation of the transformations required to implement the migration of data can be achieved in even the simplest computing environments.

Introduction

Many companies are today engaged in the standardization of the data they hold in corporate data stores. In the Geoscience sector, where ETL Solutions is active, a great variety of data stores are used, such as OpenWorks, Corporate Data Store, Seabed, Finder and PPDM.

These data stores usually define formal catalogue controlled values to assist with enforcing standards. However, this data, usually held in database tables - which themselves contain catalogue controlled values - has often been extended or adapted for regional circumstances.

This causes two day-to-day issues:

1. Corporate reports assembled from such data can easily be misleading. An attempt to discover something as simple as the relative number of 'on shore' and 'off shore' wells may fail if one of the regional stalls has decided to classify wells as 'tidal', 'shallow water' and 'deep water' - or just as 'on' and 'off'.
2. Engineers trained in one region, or working for one subcontractor, can find it difficult to work in another region where standard reference values are chosen from a different set.

While these are awkward issues, it is typically difficult to assign a direct cost to any accounting centre and hence calculate the benefit of fixing it. A more pressing reason often arises when stores are to be upgraded to newer (and more consistent) versions, or when data held in legacy stores needs to be sufficiently cleaned so it can be migrated into corporate standard stores. Often, the day-to-day issues are tolerated until a migration is required by the business.

What makes this a tricky problem is the mismatch between the tools available to the average engineer - usually not much more than Excel and SQLplus - and the extreme complexity of a modern data store. PPDM 3.8, for example, contains over 1700 tables. A hand-tool approach to such a problem is going to be very expensive, tying up valuable staff for weeks on end and quite possibly failing.

It is in these circumstances that metadata tools in the right domain specialist's hands come into their own. ETL Solutions' Transformation Manager, for example, is capable of introspecting and interpreting the metadata from any one of the stores in less than a minute. Now it is possible to see the structure of the data in an easily understood Graphical User Interface (GUI) and indeed to navigate the data (or instance values) within this same user interface. However, much more usefully than merely viewing the data, it is possible to use the metadata to directly discover - and then manipulate - all the data that will be affected by changes in standards values.

Process

The client's pilot project uses a metadata-driven method to support the data standardisation and migration of legacy databases to the client's standards.

Transformation Manager has been used in a once-only 'Prepare Phase' to automatically create transforms that will be used during every project migration. These generic transforms will be run during the Extract, Verify and Update phases of the migration process (see Figure 1 below).

Figure 1: Migration phases
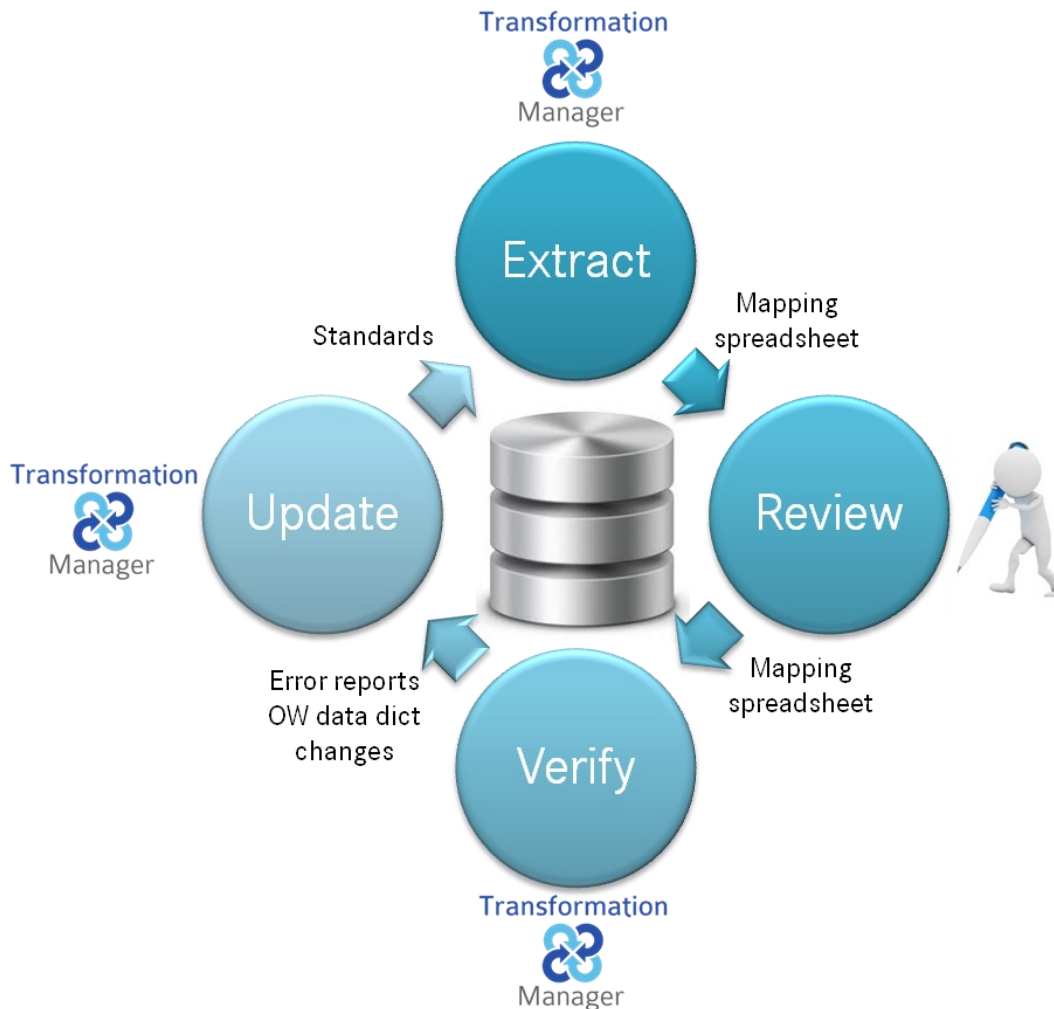
The client's standards are defined using two spreadsheets; this is the starting point for each migration of a legacy project.

The standard values spreadsheet (refer Figure 2 overleaf) contains a worksheet for each table that includes their standard values. These tables contain a number of verification code tables (_VC), a number of reference data tables (_R), and other tables, e.g. OW_DATA_SOURCE, that contain standards data. Each worksheet consists of a header row specifying the fields defined by the company's standards, followed by rows which provide the set of permitted values.

Figure 2: Standard values spreadsheet

The data dictionary spreadsheet, shown in Figure 3 below, defines data to be entered into the OW Data_Dict Table to specify default values or data ranges covered by the company standards.

Figure 3: Data dictionary spreadsheet

The data dictionary spreadsheet is used to set default values, which may be null, and a range of values that each field must conform to, expressed either as a numeric value range, e.g. for CASING_SIZE '0..10e+4', or as a set of value strings.

Any project can be updated by running a command line script during an Extract phase to determine any non-standard values used by the project. This automatic process reviews the project tables, determining any values currently used which require updating to conform to the client's standards. A set of mapping CSV files is created with values currently used which are not to standard.

A mapping file (.CSV) is created for each table and its natural key field, as in Figure 4 overleaf, which contain values that are not present in the client's standards.
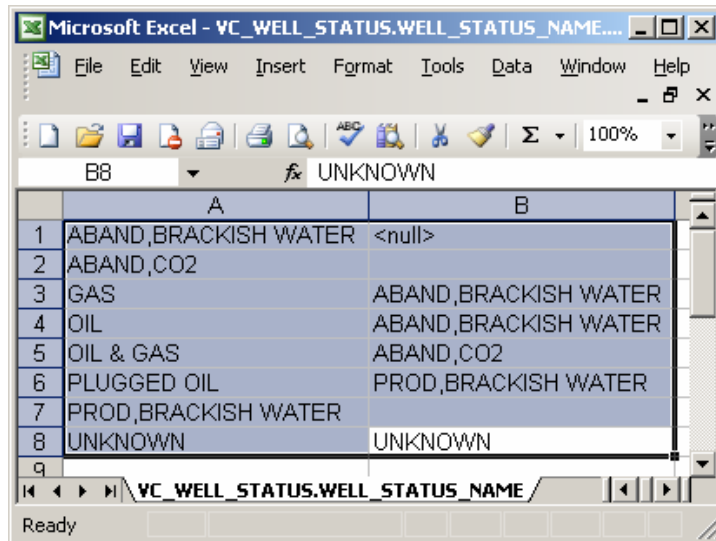
An absent mapping spreadsheet indicates that all current values are consistent with the standard values. A mapping file initially contains the set of existing values used in the project and forms the interface for a user to enter a standards value to be used instead. When mapping spreadsheets have already been created for a previous project, these can be provided and will be updated during the extract process. In this case, the spreadsheet will contain some new standard values as recommendations for the project migration.

One of the principal activities the user undertakes, referred to as the Review phase, is to check the mapping spreadsheets and provide any standard values that are required to be used instead of the non-standard presently used. Once this has been completed, the user again runs a command line script to verify that once the client's standard values are applied, the project will be consistent with the standards. Transformation Manager transforms are run to verify that this is the case. The output is a set of reports; one of these as viewed by Excel is shown in Figure 5 below.

Figure 5: Example review report
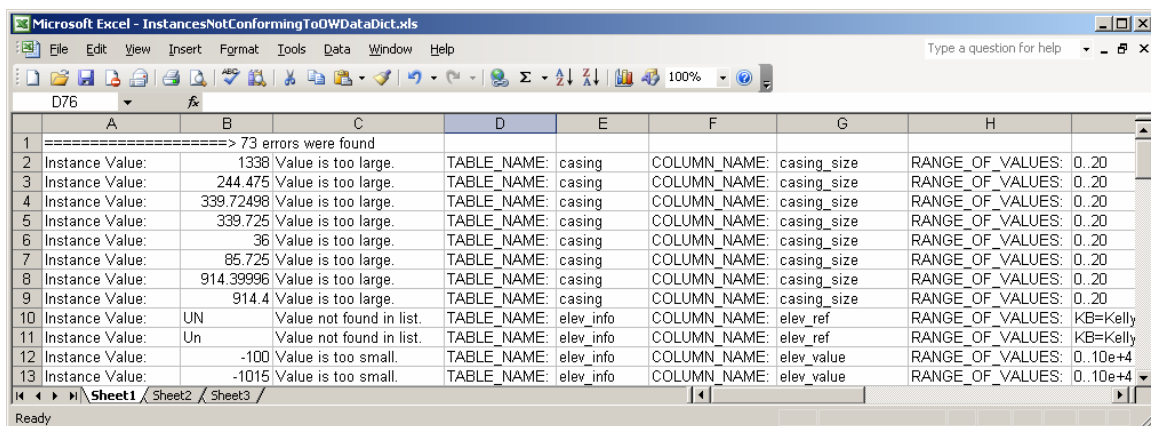
During the Verify stage, the user will correct all non-standard values and data dictionary errors until the reports indicate that a completely consistent project will be produced if migration proceeds.

Up to this stage, the project data has been accessed in read only mode and the project has been available to all other users. It is not until the final Upgrade phase of the project migration that exclusive access to the project is required.

Once verification has been completed, the final Update phase is again run from the command line and completes the migration process.

The Extract, Review and Update phases implemented using Transformation Manager are completely automatic, and provide warnings if the migration process is not completely compatible with the client's standards. The main user input is in the Review phase, where a user with knowledge of the project specifies which standard values shall be used to replace any non-standard values presently used.

Throughout the migration phases, the user operates in a simple user environment, with minimal involvement of system or database administrators. The project remains online until the final Upgrade phase, which requires exclusive access to the project database for a relatively short duration.

Benefits

The pilot achieved full compliance to the client's standards. Transformation Manager took as input the company's standards spreadsheets, read metadata and ensured instance data was consistent. The user was presented with an interface using Excel and reports with which everyone is familiar. The speed of the migration enables even the largest of projects to be quickly converted to standards.

The prototype and pilot proved how, using Transformation Manager, the migration process can be implemented with greater business control and transparency. Migrations can be flexibly deployed, minimising effort and impact on the operations and significantly reducing the cost and risk of the project. The solution can also continue to be used to measure and maintain conformance.

Conclusions

If the business wants to leverage the benefits of corporate data which conforms to a defined set of standards, using metadata-driven standardization tools will enable the business to truly implement and maintain the standards. Also, such a programme can be delivered in a fraction of the effort, time and risk that often adversely affects even well supported initiatives.

Transformation Manager delivers corporate data standards, ensuring they are implemented, enforced and inherently more flexible to meet changing needs. Finally, in the future, the migration process can easily be extended to other Geoscience areas of interest, and to other types of database stores in virtually any sector.

Contact us

ETL Solutions Ltd, Menai House, Parc Menai, Bangor, LL57 4HJ, UK
+44 (0)1248 675070
info@etlsolutions.com
www.etlsolutions.com