Computers & Geosciences 96 (2016) 87–97

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Research paper

A geodata warehouse: Using denormalisation techniques as a tool for delivering spatially enabled integrated geological information to geologists

Andrew Kingdon [a,*], Martin L. Nayembil [a], Anne E. Richardson [b], A. Graham Smith [a]

[a] British Geological Survey, Environmental Science Centre, Keyworth, Nottingham NG12 5GG, UK
[b] British Geological Survey, The Lyell Centre, Research Avenue South, Edinburgh, EH14 4AP, UK

ARTICLE INFO

Article history:
Received 1 September 2015
Received in revised form
22 July 2016
Accepted 25 July 2016
Available online 2 August 2016

Keywords:
Physical properties
Databases
RDBMS
Denormalisation
Geoscience data warehouse
Webservice

ABSTRACT

New requirements to understand geological properties in three dimensions have led to the development of PropBase, a data structure and delivery tools to deliver this. At the BGS, relational database management systems (RDBMS) has facilitated effective data management using normalised subject-based database designs with business rules in a centralised, vocabulary controlled, architecture. These have delivered effective data storage in a secure environment. However, isolated subject-oriented designs prevented efficient cross-domain querying of datasets. Additionally, the tools provided often did not enable effective data discovery as they struggled to resolve the complex underlying normalised structures providing poor data access speeds. Users developed bespoke access tools to structures they did not fully understand sometimes delivering them incorrect results.

Therefore, BGS has developed PropBase, a generic denormalised data structure within an RDBMS to store property data, to facilitate rapid and standardised data discovery and access, incorporating 2D and 3D physical and chemical property data, with associated metadata. This includes scripts to populate and synchronise the layer with its data sources through structured input and transcription standards. A core component of the architecture includes, an optimised query object, to deliver geoscience information from a structure equivalent to a data warehouse. This enables optimised query performance to deliver data in multiple standardised formats using a web discovery tool. Semantic interoperability is enforced through vocabularies combined from all data sources facilitating searching of related terms.

PropBase holds 28.1 million spatially enabled property data points from 10 source databases incorporating over 50 property data types with a vocabulary set that includes 557 property terms.

By enabling property data searches across multiple databases PropBase has facilitated new scientific research, previously considered impractical. PropBase is easily extended to incorporate 4D data (time series) and is providing a baseline for new "big data" monitoring projects.

1. Introduction

1.1. Changes to geological data delivery

In recent years three dimensional (3D) geological framework models (Kessler et al., 2009) have been replacing traditional maps as the primary mechanism for transmitting geological information. The ever more complex uses of the subsurface mean that models need to progress beyond representing only lithostratigraphy and structure to highlighting geological heterogeneity. This necessitates new relationships with the underpinning data.

Geological survey organisations (GSO) exist to provide governments, industry and the public with information to understand the subsurface, by identifying and synthesising data. The British Geological Survey (BGS) acts as a repository for all subsurface data from the United Kingdom landmass and continental shelf including many that describe the physical properties of the geosphere. Crucially BGS collects only a fraction of this; the majority is collected for other purposes, including civil engineering, natural resource extraction, (energy, mineral wealth, groundwater), and disposal of waste.

1.2. Geoscience data and decision making

Traditional geological maps and models are created from geologist's field observations combined with data sampled from

* Corresponding author.
E-mail address: aki@bgs.ac.uk (A. Kingdon).

http://dx.doi.org/10.1016/j.cageo.2016.07.016
0098-3004/

boreholes to delineate lithostratigraphic units, but treat the zones between these surfaces as homogeneous. UK geological maps depend upon nationally collated datasets (e.g. historic field observations and geophysical logs datasets from deep boreholes).

It is now possible to map the 3D spatial variability of lithology or physical properties rendered as voxels to fully understand both this heterogeneity and its impact on societal problems (e.g. Kearsey et al., 2015). This requires new data inputs.

1.3. Property data requirements

Whilst the land surface can be easily sampled, subsurface properties can only be sampled directly by drilling, returning either samples or geophysical measurements to surface. GSO outputs incorporate physical property data sources including:

- Laboratory analyses from core and geological samples, collected by GSOs, geotechnical, oil and water industries
- Geophysical logging data and derived proxies of properties
- Civil engineering geotechnical testing

Each dataset will have been collected at different times using distinct conditions for different purposes and archived in dataset specific structures.

1.3.1. 3D modelling of physical property data

Understanding the variation of properties in the geosphere in 3D requires the building of models. Previous work has highlighted many issues in BGS with discovering, interrogating, extracting, collating and serving the physical property data necessary to solve these problems.

Building physical property models therefore requires multiple data extraction operations and levelling from multiple inputs. Existing systems to extract data from database tables one dataset at a time have proven overly burdensome to users. Therefore there was a requirement to develop a new methodology for data storage, discovery and extraction.

1.3.2. Attributes of physical property data

Regardless of origin, all property information has a minimum set of common properties:

- A location measured in 3 dimensions ($\pm$ timestamps)
- Measured values, units and margins of error
- Metadata describing the acquisition, analysis, storage and processing.

Provided data can be expressed in this format, multiple datasets can be integrated. This allows understanding of the impacts of environmental processes at specific locations, thereby developing new scientific insights. Physical property investigations are constrained by access to the type and volume of data available. Achieving this efficiently requires this data to be served in a standard format and analysed / visualised using common tools. Few datasets available to GSOs are pre-conditioned and fully attributed with metadata, many have no standards for metadata description and use different spatial reference fields.

The PropBase scoping study (Shaw, 2006) identified requirements for data that were needed to allow the study of subsurface physical, chemical and other properties. The data required to deliver this vision were held in various data storage locations and formats. Integration of these therefore represented a significant technical challenge and was not a practical deliverable in most cases.

1.4. Solution for spatially enabled geoscience data storage

This paper describes the methodology, data structures and data access tools of a “data warehouse” for aggregating geological data from multiple interconnected databases. This achieves the objectives of the PropBase scoping study. Given the diversity of data that needs to be understood and the many tools used to study these, a single conventional subject-domain based normalised database could not hold all such data and is neither a practical nor robust solution. The structures described in this paper allow multiple datasets to be integrated in a single data structure without the loss of data integrity.

The paper focusses on the requirement to be quickly and effectively access a broad range of geoscience information. This has been achieved by transforming data held in existing relational databases and loading them into a new denormalised data structure “data warehouse” using a combination of procedural routines and database jobs, in a manner analogous to materialised views. This data structure is then further denormalised by pre-resolving all joins and codified vocabulary values into a single QueryLayer object. This layer is optimised using a combination of techniques to include normal and text indexing of key columns and data partitioning. This QueryLayer is a summary data structure for three-dimensional (3D) property datasets.

The structure allows multiple datasets to be searched and visualised together to facilitate a new understanding of subsurface properties. The ease of data discoverability and download maximises their value as well as supporting visualisation in multiple software tools. The denormalised data access layer is required because of the heterogeneous input data, the number of output data formats required and the need to aggregate data in a single homogenous data structure with common semantics so they can be accessed together.

1.5. Use case of data provision

A use case has been identified to test the effectiveness of this system in providing data for 3D modelling. The BGS Energy Security and Innovation Observing System for the Subsurface (ESIOS) project will develop a subsurface energy research centre. Modelling of the proposed site at Thornton in Cheshire needs to be undertaken to understand distribution of physical properties in 3D providing a test of the advantages or disadvantages of a new data structure. Therefore all available property data around the proposed ESIOS test site has needs identifying allowing collation of datasets from many sources each currently held in discipline specific relational databases.

2. Methodology

2.1. Introduction

This paper describes the creation of data structures analogous to those used in data warehouses. These are implemented in a relational database management system (RDBMS) using tables, materialised views, procedures written in Oracle PL/SQL™ procedural language and associated infrastructure to provide a standardised access to physical property data derived form multiple subject-domain databases.

2.1.1. Relational databases

Codd (1970) established the basic principles of the relational database, subsequently codified as the 12 rules of databases (Codd, 1985). Chamberlin and Boyce (1974) developed these principles into Structured Query Language (SQL). This was recognised as a