Full citation

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

Kingdon, A.[1]*, Nayembil, M.L.[1], Richardson A.E.[2], and Smith, A.G.[1]

Affiliations:

1: British Geological Survey, Environmental Science Centre, Keyworth, Nottingham, NG12 5GG, UK

2: British Geological Survey, The Lyell Centre, Research Avenue South, Edinburgh, EH14 4AP, UK

* Corresponding author: aki@bgs.ac.uk

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 didn't 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.

1.0 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 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 (± 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.

2.0 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^{TM} 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 Boyse (1974) developed these principles into Structured Query Language (SQL). This was recognised as a standard of the American National Standards Institute (ANSI) in 1986 and International Organization for Standardization (ISO/IEC 9075). Today, SQL is accepted and used as the ubiquitous RDBMS language.

The relational database model is dependent upon the process of normalisation. This involves organising the attributes and tables of a relational database to minimise data redundancy. Normalisation involves decomposing a relation (or table) of data into smaller relations without losing information; defining foreign keys in the original table that references the primary keys of the new ones. Once this is achieved, then data is “isolated,” meaning that additions, deletions, and modifications of an attribute can be made in a single table and then logically propagated through the rest of the database using the defined foreign keys. Holding geological data as relational database tables has many advantages: the way in which data are stored, validated and served, including uniqueness (every item is only stored once), atomicity, security, constraints and vocabularies. Normalisation also enforces business rules on data and is essential for ensuring its integrity.

These principals have defined BGS’s data storage procedures since relational databases were first used for storing geochemistry data (Harris and Coats, 1992, Coats and Harris, 1995). This has allowed development of strategic databases which underpin many geoscience disciplines including borehole locations, borehole stratigraphy, geochemical sampling locations and analyses results. Relational databases constitute an excellent structure for primary data storage. Figure 1 and Figure 2 show the data models of BGS’s geochemistry and geotechnical databases respectively, demonstrating the full complexity. However, these normalised designs come at the cost of an apparent complexity and degradation in querying performance, particularly if directly queried by users and/or applications without a middle layer to pre-resolve joins and codified values used in dictionaries.

2.2 The application of databasing to diverse geoscience data

Managing datasets within normalised databases in an RDBMS enforces business rules on data, ensuring secure storage and making data updates auditable. This establishes the interrelationships between data, enforces referential integrity, eliminates unnecessary redundancy, facilitates easy and consistent update of the data, allows for uniqueness and secures data for storage. However, this also causes some disadvantages, particularly for data presentation and easy querying, which include:

- High costs of implementation
- Complex data structures which can be difficult for end-users to interact with either directly or through client applications
- Prioritising data storage concerns over providing tools for serving data to end-users

However carefully these designs are implemented, there is always a risk that end-users perceive a loss of data ownership. Generally these can be remediated provided that data access is fast and reliable. Performance depends upon the design and data volumes, so whilst some separation of data-user from data providers is inevitable, experts need to rapidly understand data output requirements. Whilst the RDBMS maybe optimal for primary data storage, alternative data structures may be required to optimise data supply to end-users.

2.3 Denormalisation to aid geological data access and visualisation

Connolly & Begg (2010) defined normalisation as a technique for producing a set of relations (tables) with desirable properties (attributes), given the data requirements of an organisation. The aim is to identify a set of relations with: the minimum number of attributes; attributes with a close logical relationship i.e. have a functional dependency; and minimum redundancy. Consequently multiple normalised tables need to be joined during querying, which presents significant performance burdens as the complexity of data structures and linkages increases requiring alternative approaches. Denormalisation is the reverse process where normalised tables are refined and consolidated into secondary simplified data structures with controlled redundancy to maximise performance (Connolly and Begg, 2010). In the authors' opinion, this is best suited to secondary data storage for querying by applications and users, where unsatisfactory query performance can be seen where table(s) are seldom updated and frequently queried.

This work describes the application of denormalisation to relational databases to deliver data integration and improve access to subsurface physical properties by holding heterogeneous datasets in a single structure. A generalised structure for all property data that can be expressed in 3D space was implemented regardless of the source data structure and subject contents. This approach allows data output formats to be standardised but must be guided by very clear requirements statements and a high-level of interaction between developers and end-users to deliver a fit-for-purpose solution. Specific input property datasets include geochemistry, lithology / lithostratigraphy, and geotechnical data. Figure 3 is a schematic representation of the separate subject-domain databases that needed to be queried individually for all related property information prior to the implementation of the denormalised PropBase data structure.

2.3.1 Geoscience data requirement for a data warehouse

Inmon's (1992) commonly cited definition of a data warehouse "is a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision-making process."

The PropBase scoping study (Shaw, 2006) defined the need for an integrated single access system for disparate geoscience data resources (thus is subject oriented), that is interoperable and can be interrogated, time-variant (as data will change with time without loss as datatypes and records are accreting to existing datasets) and non-volatile (data added to this system will remain unchanged though the uses to which this is applied may alter). Therefore this meets Inmon's definition of a Data Warehouse.

BGS has therefore developed a geoscience data warehouse to explicitly address these requirements. However, since completion it has demonstrated much wider-reaching applicability to easy data discovery, download and use. This system, known as PropBase, extracts, transforms and loads data into a simplified data model from heterogeneous sources so that the data is compatible and accessible via a single interface. The data model is supported by a controlled vocabulary set based on a semantic layer, consolidated from multiple vocabularies across normalised databases. These structures are implemented in an Oracle® RDBMS (but are adaptable for other RDBMS environments).

The component parts are:

- Core Data Model
- Extraction, transformation and loading routines (Procedures)

- Data Access Layer (QueryLayer)
- Web services providing multiple data formats.
- Data discovery and download client (Data Explorer)

PropBase also includes a suite of search utilities that enable rapid data discovery and download of multiple data types so these can be synthesised simultaneously. Web services allow for machine-to-machine interaction, enabling other software to interrogate the combined dataset, primarily to visualise it.

The system has a significant impact by enabling multiple datasets to be integrated for scientific understanding, potentially across multiple clients, whilst ensuring that data is properly managed and available for future use.

2.3.2 Core Data Model

The core data model is based around the concept of a summary secondary data layer which can present complex data from multiple structured sources in a simple denormalised set of generic tables, and other programmable units, within a RDBMS. This summary data layer brings together 3D (x, y, z) and/or 4D (x, y, z plus time for temporally variable data) property information, from various databases each with their own relational structure and rationalises each into a generalised structure used as a single consistent point of access to this data. The data model is implemented within the same RDBMS as the source databases, facilitating easy loading of the data.

The data structure contains a main table PRB_DATA with the following attribution:

- a unique identifier
- the data source
- the unique identifier from the parent database for traceability
- the spatial location (x, y, z) or spatial / temporal location (x, y, z, t) and Coordinate Reference System (CRS)
- the property type
- the property value

- the units
- necessary qualifiers
- precision information and an audit trail

The data model also incorporates vocabularies to constrain some attributes including: the data sources, property types, units of measure and CRS. Vocabularies have been collated from the primary databases with unused values eliminated. The property vocabulary is a key component of the structure as this defines what properties and inherited hierarchies are to be coded and also guides the discovery process as to how these are extracted (searched) from the structure. Semantic interoperability is the ability of computer systems to exchange text data with unambiguous, shared meaning. In PropBase this is enabled by manually mapping related properties, for example porosity data from unrelated origin to common values; the ontologies necessary to enable a linked data approach do not yet exist in the geosciences.

PropBase is a spatially aware data structure incorporating Oracle® Spatial functionality to represent the various geometries recorded for easy access, integration and visualisation. Grid transformations are automated eliminating the need to store data in multiple CRS. Figure 4 shows the core data model of PropBase with its simplified structures.

2.4 Property data management and output

2.4.1 Extraction, Transformation and Loading (ETL) Routines

To ensure data integrity, given the heterogeneity of input data, the system requires a full audit trail, linkages to source records and regular updates. A co-ordinated technical approach is essential to keep the “warehouse” synchronised.

PropBase uses procedural routines written in Oracle PL/SQL containing the logic to carry out data extraction, transformation and loading processes (data manipulation) and to keep the contents of the layer synchronised with the underlying databases. These routines can be run regularly (e.g. weekly) and/or invoked on demand. For each database three procedures have been developed to enable data insert, update and deletion operations.

2.4.2 Data Warehouse Layer or “QueryLayer”

The PropBase architecture includes a highly optimised data access layer to facilitate access by multiple applications. This is an additional denormalised table on the core data model, optimised as for query by PropBase Webservices. The tables in the core data model are never accessed directly by client applications, these are only used to collate and harmonise the data from the heterogeneous data sources to populate the QueryLayer. This was created by pre-joining the data tables that make up the core data model, with decoded vocabulary values and also spatially enabled to provide a single access object for property data. Figure 5 shows a diagrammatic representation of the simplified data access layer.

2.4.3 Optimisation of data warehouse layer (QueryLayer)

The QueryLayer is a key component of the PropBase architecture as it’s the single point of access for applications to the collated property data. This optimisation is the key element of the system as its generic structure allowing multiple datasets to be interrogated using common tools. Instead of the repeated columns that might be expected from “pure denormalisation” this system uses procedures to reformat data into a universal format. This allows new datasets to be added without bespoke engineering, by creating a universal template structure into which any spatially enabled data can be added. Whilst databases of physical properties are not new, PropBase’s unique innovation is this creation of a generic data structure capable of holding previously heterogeneous datasets in a homogenous format optimised for speed of data discovery, query and delivery. This ease of access facilitates new areas of scientific discovery that, whilst previously theoretically possible, were completely impractical due to the difficulties of collating and assessing the necessary data. To facilitate such rapid access by applications and direct querying, the QueryLayer is highly optimised as a query only object. The layer currently contains 28.1 million spatially enabled property data points from 10 separate primary databases containing 50 different data types.

Oracle (2015) describes the levels of optimisation needed in an RDBMS to make such a layer responsive to querying by applications and users. Consequently the layer uses normal and text indexing on key searchable columns, and also partitioning. The layer uses 48 numbered (hash) partitions with a partition key containing 13 key columns defined on dedicated table-spaces within the database.

The vocabulary set supports 557 different property types enabling the semantic interoperability. These are also presented as a separate indexed object to aid the response speed by property type queries. As a separate object, it provides greater flexibility for further optimisation, easily adding and/or removing data attributes/data and the ability to present data in different spatial transformations. Figure 6 shows an architecture diagram of the entire PropBase System.

2.4.4 Webservice

The PropBase web service component is created in Java™. It is designed as a tiered architecture, following the Model View Controller (see Figure 7) pattern to allow future flexibility to changes to the codebase. For example changes to either the output formats or to the database structures can be handled separately without impacting each other.

The Controller layer is the core of the Application and includes the business and Application logic required to receive end-user requests, pass these to the Database Access layer, and then call the appropriate part of the view layer. This Database Access layer has been optimised alongside the Oracle® Data QueryLayer for database access, within the PropBase warehouse, following the Data Access Object Pattern. This layer contains optimised querying and data paging strategies targeted to the database structures using SQL and additionally contains short-term memory caches of frequently used but largely static content, e.g. the vocabularies within the RDBMS. This reduces memory consumption and “object creation” thereby increasing throughput.

The Presentation layer within the discovery tool converts model objects into the output format requested by the end-user. Depending upon the format requested, this may use an existing library or be a piece of more complex output code to return the resulting object to the end-user. Multiple Data formats are available to the user in this way including, JSON, RDF, KML, Atom, ESRI Shapefiles, CSV and HTML.

Interactions with the end-user are handled using RESTful principles (Fielding, 2000) which is a methodology for HTTP (Hypertext Transfer Protocol) interactions, and wherever possible the output format's URL references back to the application used. This service additionally uses HTTP1.1 standard content negotiation and by setting the appropriate HTTP headers, enabling clients to perform their own caching strategies.

BGS maintains or uses a number of software libraries in order to allow this architectural pattern to be applied quickly into multiple applications. Open source components used include, but are not limited to, Restlet Framework, Guava and Apache Commons.

2.4.5 Data explorer tool

The "PropBase Explorer" tool which consumes the webservice to provide data discovery, export and application integration functions to enable users interrogate the property data points.

The data explorer calls the main PropBase service with query URLs and asks for JSON format data. Its pages are instantly updated using JavaScript and are only ever reloaded to reset the OpenLayers maps. Figure 8 below shows details of the Javascript client and how it interacts with the Webservice on a request.

These search functions include: UK and world spatial querying; free text searches (akin to Google searching); a data wizard capability for step by step API URL generation, ability to drill down through the hierarchy of properties in the system; presentation of the relevant properties in a periodic table and data exports with the ability for application integration, are retrieved from the underlying webservice. These functionalities are shown in Figures 9 and 10.

3.0 Results of use case

Efforts were made to identify property data sets around a fixed point, the proposed ESIOS drillsite using boxes of 1 and 10 kilometres radius respectively. The aim was to identify:

- The total number of property datasets present in this area
- The number of points from each non-null dataset
- Output of these datasets formatted for incorporation into 2D Geographical Information Systems (GIS) and 3D models.

The first test used PropBase to interrogate and deliver these data. The second test used native SQL and pre-existing GUI to interrogate the source databases.

3.1.1 Results from PropBase data outputs

Comparisons of data delivery from these use cases are instructive.

	MIN BNGE	MIN BNGN	MAX BNGE	MAX BNGN	Records found	Total	Notes
1km radius test	344500	375700	344700	375900			
Data Sources					0	47	
Properties					0	296	
Measurements					0	>2500000	
Time Taken						0 seconds	No data
10km radius test	343600	374800	345600	376800			
Data Sources					8	47	
Properties					53	296	
Measurements					142	>2500000	
Time Taken						~5 minutes	

Table 1: Data identified for ESIOS drillsite using PropBase for two search boxes of radi 1km (top) and 10 km (bottom), Left-hand panel: bounding search coordinates in British National Grid (BNG). Right-hand panel, number of data sources, property types and number of lines of unique data identified.

The 1km box demonstrated a complete absence of data held in this area, as shown in Table 1. As results were returned instantaneously no further time was wasted on this. The second identified that data availability from 8 of a possible 47 datasets interrogated by PropBase and returned 53 datatypes from some 296 possibilities, and some 142 specific geolocated records. These were immediately available exported as a single file for use.

3.2 Comparison of data discovery without PropBase

Without using PropBase, discovery of spatial data is achieved using either GIS or legacy database Graphical User Interface (GUI). This requires knowledge of around ten separate applications (so overlooking data is likely) requiring extensive knowledge of BGS data holding. Each system takes approximately five to ten minutes to check if data exists, including the time to load system and to examine each for the target area, presuming they all have an existing GIS or GUI presence.

In contrast PropBase allows checking of all datasets in about the time needed to check one individually. Data completeness concerns are eliminated with near real-time update schedules.

3.3 Data Download

Where relevant GUI exist to identify data for download, retrieval takes about five minutes per dataset. For databases without GUI this would require scripting of native SQL data extraction queries. For example, querying geochemistry data requires joining four or five tables, a database expert might take an hour to research and build this. To many geologists lacking these skills data is inaccessible.

Finally, several hours would be needed to harmonise results including conversion to standard CRS and datum, and to complete common datatypes where they may be missing. Again specific domain knowledge of each system would be required.

In contrast the PropBase data can be extracted in seconds in a choice of standard formats allowing data to be immediately interpreted without significant further manipulation.

4.0 Discussion

4.1 Lessons from the use case

The ESIOS area data extraction use case highlights profound performance advantages provided by PropBase. This ignores effort already expended to deliver the system but given that building parametrised 3D models at different locations and scales is an ongoing activity this represents a worthwhile saving. The simplicity of both discovery and serving of data makes such studies immediately possible.

The first test highlighted the absence of data in this site, which is non-trivial as it saves further wasted time. The second test exported a set of data for incorporation into models and GIS that could then be used to underpin geological understanding.

4.2 Merits of the PropBase structure

PropBase has answered a set of data access problems that were previously treated as intractable. Figure 6 shows a diagrammatic representation of the entire PropBase system, including both main tables and data access layer. The data now used to parameterise 3D models with physical properties are not new but had previously been difficult to synthesise and output together because the complex normalised databases that each data type had to be queried individually. This made data extraction time consuming and difficult to repeat for multiple locations.

When data was accessed from primary tables, querying was time-consuming with multiple data queries being created separately. PropBase's combination of simplified data structure and QueryLayer optimised for rapid data export, means that the Data Explorer front-end can deliver outputs very rapidly. Choosing between data types simple involves toggling between them. Since the implementation of PropBase the querying effort is almost eliminated ensuring that data is identified quickly and consistently, without further interactions with developers. This enables geologists to parameterise models to answer scientific questions, and research new relationships between datasets immediately as these can be visualised, assessed and accepted or discarded quickly. Previously this would have been too time-consuming to undertake.

4.3 Simplified data input

All datasets discussed thus far are those already held in RDBMS where the issue is improving discoverability and access. However, many scientific datasets held in GSO are collected and stored only as spreadsheets. These enable data to be quickly interrogated and utilised by individuals (or small project teams) but at the cost of loss of metadata, data security, lack of integration with other datasets, and wider access. These are critical if data is to be reused and its value maximised. Such datasets are typically stored in a “single row per data object” style, including fundamental metadata (e.g. sample location) embedded within it. Creating fully normalised databases to hold complex datasets are developer intensive activities. For small static, often legacy datasets, this burden cannot be justified, yet these still need to be held securely, made discoverable with other datasets and available to end-users.

PropBase has allowed the development of an intermediate data structure as a transitional step to a full corporatized standard. Spreadsheets can be imported into the PropBase structure using volatile tables and procedures. Data can be held and used dynamically as part of PropBase, but without the burden of full normalisation. Such datasets transition from private collections to corporate assets, are secure and ready to be interrogated.

4.4 “Big data” vs. “long-tail” data

There has been extensive discussion of the impact of “big data” on environmental science investigations (e.g. .Sellars et. al, 2013) typically defined by attributes such as “volume”, “variety”, “velocity” (Laney, 2001) and sometimes “veracity” or “value”. Environmental big data can be defined as:

- Spatially extensive
- Containing suites of individual measurements
- repeated measurements to create time-series

Examples include large remote sensing datasets and geophysical surveys which are particularly important in nations containing large tracts of inaccessible open-space like the USA, Canada and Australia (Wyborn et.al, 2014). Such data are consistently formatted but large in volume requiring significant computing power to process them.

Most geoscience information sources do not conform to this definition of big data. The UK landmass has been extensively sampled through direct mapping for over 200 years by multiple industries for diverse purposes. Subsurface data are acquired individually (e.g.

boreholes) and the high acquisition costs make repeat sampling unlikely. Therefore such "long-tail" data may be of greater relevance today than "big data".

Subsurface analysis of the UK is complex not because of data volume but availability. Such datasets are small in volume (MB - GB scale) but highly heterogeneous. PropBase gains value from "long-tail" data; this can be used to ground-truth time variant "big data" studies.

4.5 Semantic interoperability

Within the PropBase system, data types can currently only be associated if the names are sufficiently similar to one another to easily allow them to be matched. Therefore "Bulk Density" measurements, geophysical "density" logging and measurements from geotechnical testing are easily linked by similar names. These indicate a single physical property but calculated at different scales, resolution and accuracy.

These issues have been addressed by the PropBase vocabulary layer which includes properties that are derived and compiled from the source tables. Whilst this does not yet represent full semantic interoperability this provides a strong basis for the development of a semantic layer to allow for deeper data mining. Ultimately, analysing datatypes collectively requires complex semantic understanding including relationships such as "is related to". This will require a semantic architecture to understand these relationships based on as yet unavailable ontologies. Summarising of property types identifies the route towards development of formal ontologies.

4.6 Consideration of Alternative Technical Solutions

Consideration was given to other technical solutions for delivering this functionality using, for example, NoSQL. This was not mature during the initial system design. NoSQL presents a powerful tool for facilitating access to unstructured datasets. However, NoSQL is not relevant to the problem of optimising data storage to maximise delivery efficiency of already structured spatially enabled datasets. PropBase's key advantage is improved data access provided by the common data structure and consequent massively improved output performance which facilitates answering of geological questions which it would previously have been impractical to ask. Future iterations of this design will consider incorporating NoSQL structures, in order to further optimise data retrieval.

5.0 Conclusions

Development of the PropBase warehouse has facilitated a new ease of access to physical property data for both discovery and serving. Data are no longer simply archived a subject-oriented databases but are immediately available for integration with similar data from other domains and thus have been used more frequently to deliver geological insight.

PropBase has facilitated enhanced delivery of physical properties data by:

1. Speeding up data discovery for users.
2. Providing a quick look at data both by type and by spatial availability (e.g. what rock strength data is available in Birmingham)
3. Identifying areas of poor data resolution thus allowing prioritisation of data acquisition and/or capturing appropriate proxies in those locations
4. Allowing physical properties data to be served efficiently for 3D modelling and visualisation (Google Earth, GIS) software
5. Allowing related data to be reviewed together quickly and reliably
6. Creating “mash-ups” of previously unrelated data to identify new relationships by allowing simple extraction of multiple data sets for the same locality.

.

The PropBase implementation is not static. Firstly it is subject to the constant maintenance and updates of BGS’s core datasets held in primary data objects. Secondly it can be extended as required to include new datasets to support new decisions; new data are being added continuously. The flexible design means that provided data meets the generalised requirements (that is location details in 3D and property attributes) it can be harvested and ingested into the PropBase structure immediately once specific procedures for data extraction, transformation and loading are adapted and any additional vocabulary values added to the semantic layers.

The structure ensures that primary databases are maintained, ensuring data security and integrity, with secondary data warehousing facilitating data export and visualisation. The optimised QueryLayer, is a “flattened” structure (akin to a large file), optimised using different types of indexing and data partitioning to facilitate the most likely export queries. This structure is now being extended to underpin BGS’s 4D (time variant) monitoring of environmental change through installed sensors.

Acknowledgments

This paper is published with the permission of the Executive Director of the British Geological Survey, funded by NERC National Capability.

The authors wish to thank our reviewers and the following colleagues: Kate Royse, Garry Baker, and Mark Fellgett.

References

Chamberlin, D.D. and Boyce, R.F.,1974., SEQUEL: A Structured English Query Language. Proceedings of the 1974 ACM SIGFIDET Workshop on Data Description, Access and Control (Association for Computing Machinery): 249–64.

Coats, J. S. and Harris, J. R. Harris, 1995,. Database design in geochemistry: BGS experience Geological Society, London, Special Publications, v. 97:25-32, doi:10.1144/GSL.SP.1995.097.01.04

Codd, E.F., 1970. A Relational Model of Data for Large Shared Data Banks. Communications of the ACM 13 (6): 377–387.doi:10.1145/362384.362685.

Codd, E.F,.,1985. Is Your DBMS Really Relational? ComputerWorld.

Connolly, T.M, and Begg, C.E, 2009. Database Systems: A Practical Approach to Design, Implementation and Management. 5th edition. Addison Wesley pp 1400. ISBN: 978-0321523068

Fielding, R. T., 2000. Architectural styles and the design of network-based software architectures· Doctoral Dissertation, University of California, Irvine ©2000. ISBN:0-599-87118-0

Harris, J.R. and Coats, J.S., 1992. Geochemistry Database, Data Analysis and Proposed Design. British Geological Survey Technical Report WF/ 92/5 (BGS Mineral Reconnaissance Programme Report 125).

Inmon, W.H. (1992). Building the Data Warehouse. Wiley. ISBN 0-471-56960-7.

ISO/IEC 9075-1:2008: Information technology – Database languages – SQL – Part 1: Framework (SQL/Framework).

Kearsey, T., Williams, J.D.O., Finlayson, A., Williamson, J.P., Dobbs, M.R., Marchant, B.P., Kingdon, A., Campbell, S.D.G.,2015. Testing the application and limitation of stochastic simulations to predict the lithology of glacial and fluvial deposits in Central Glasgow, UK. Engineering Geology, 187. 98 - 112

Kessler, H., Mathers, S.J. & H.G. Sobisch. 2009. The capture and dissemination of integrated 3D geospatial knowledge at the British Geological Survey using GSI3D software and methodology. Computers & Geosciences, 35, 1311-1321. http://dx.doi.org/10.1016

Laney, D., 2001. 3D Data Management: Controlling Data Volume, Velocity and Variety. Gartner. (http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf)

Oracle, 2015. Database VLDB and Partitioning Guide. In: Oracle Database Online Documentation 11g Release 1 (11.1). Data Warehousing and Business Intelligence (http://docs.oracle.com/cd/B28359_01/server.111/b32024/partition.htm)

Shaw, R.P., 2006. PropBase Scoping Study. *British Geological Survey Internal Report*, IR/06/088. 35pp.

Sellars, S., P. Nguyen, W. Chu, X. Gao, K. Hsu and S. Sorooshian, 2013. Computational Earth Science: Big Data Transformed Into Insight, Eos Trans. AGU, 94(32), 277. DOI: 10.1002/2013EO320001

Wyborn, L.A, Evans, B.J.K., Pugh, T., Lescinsky, D.T., Foster, C. and Uhlherr, A. , 2014 Collaboratively Architecting a Scalable and Adaptable Petascale Infrastructure to Support Transdisciplinary Scientific Research for the Australian Earth and Environmental Sciences. AGU Fall Meeting Abstracts, Vol 1, 6pp.

Figures

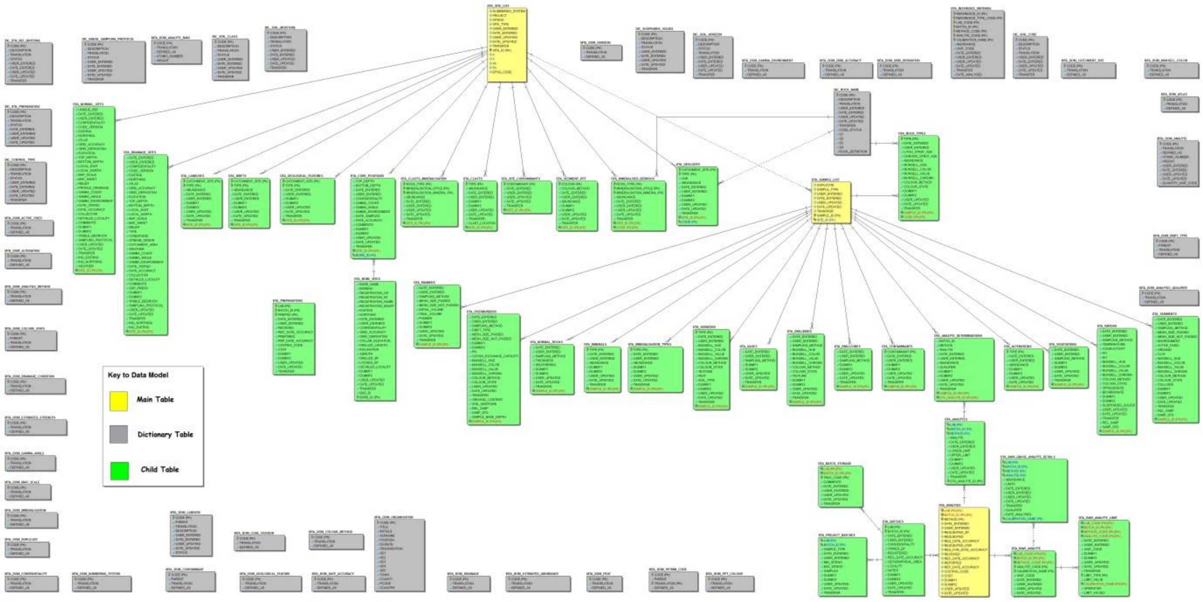


Figure 1: Data model of BGS GBase Geochemistry database as an example of the implicit complexity of a fully normalised database including main and child tables, plus dictionary tables, and entity relationships.

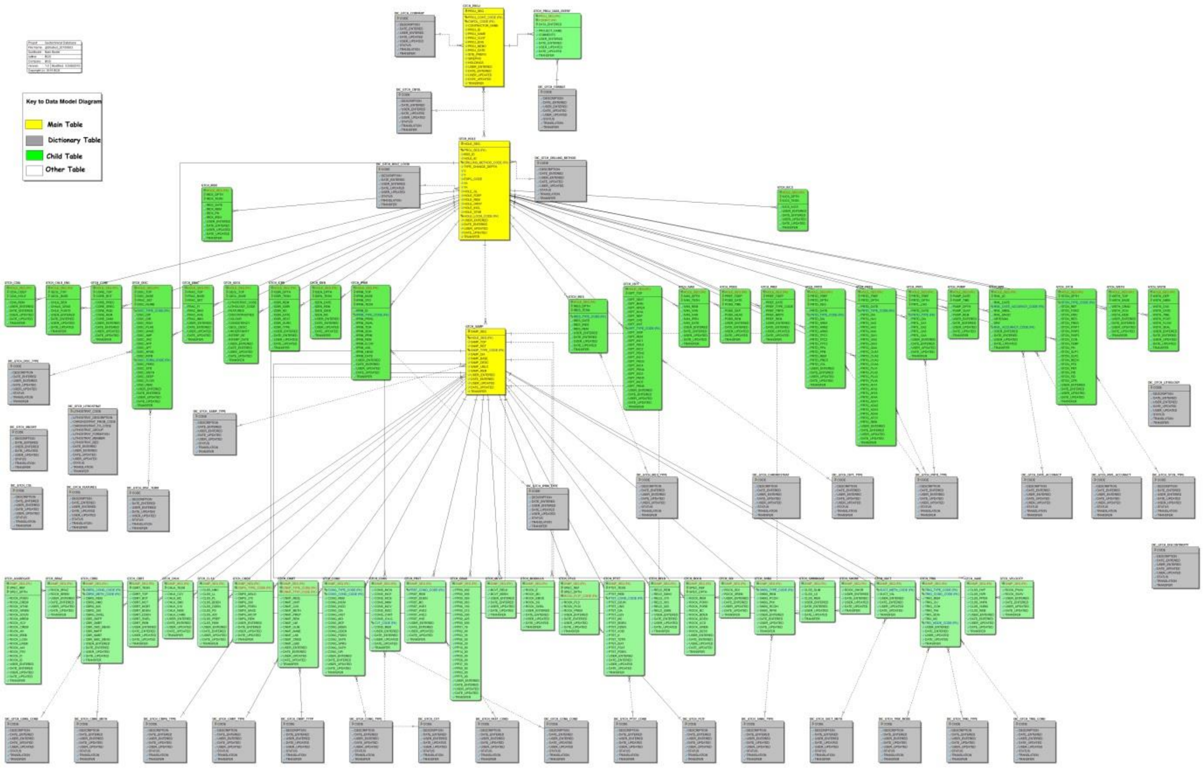


Figure 2: Data model of BGS's Geotechnical database (Engineering Properties - AGS) another example of a normalised database including main and child tables, dictionary tables, and relationships.

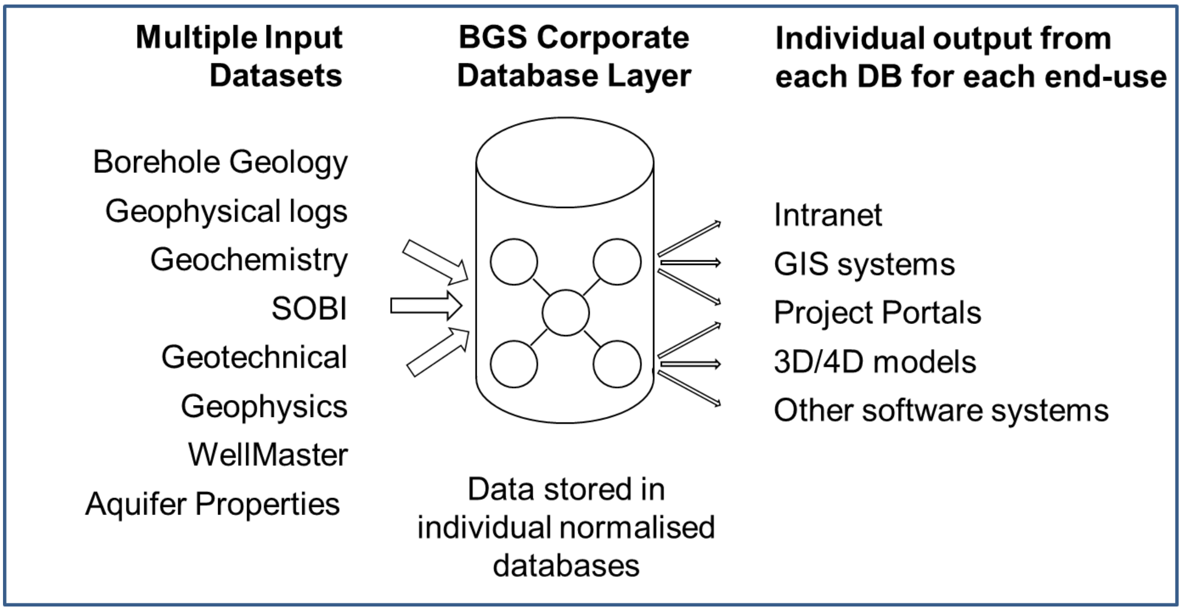


Figure 3: Diagrammatic representation of the BGS corporate data structures in a full normalised data models with multiple databases stored independently and queried individually by multiple bespoke systems

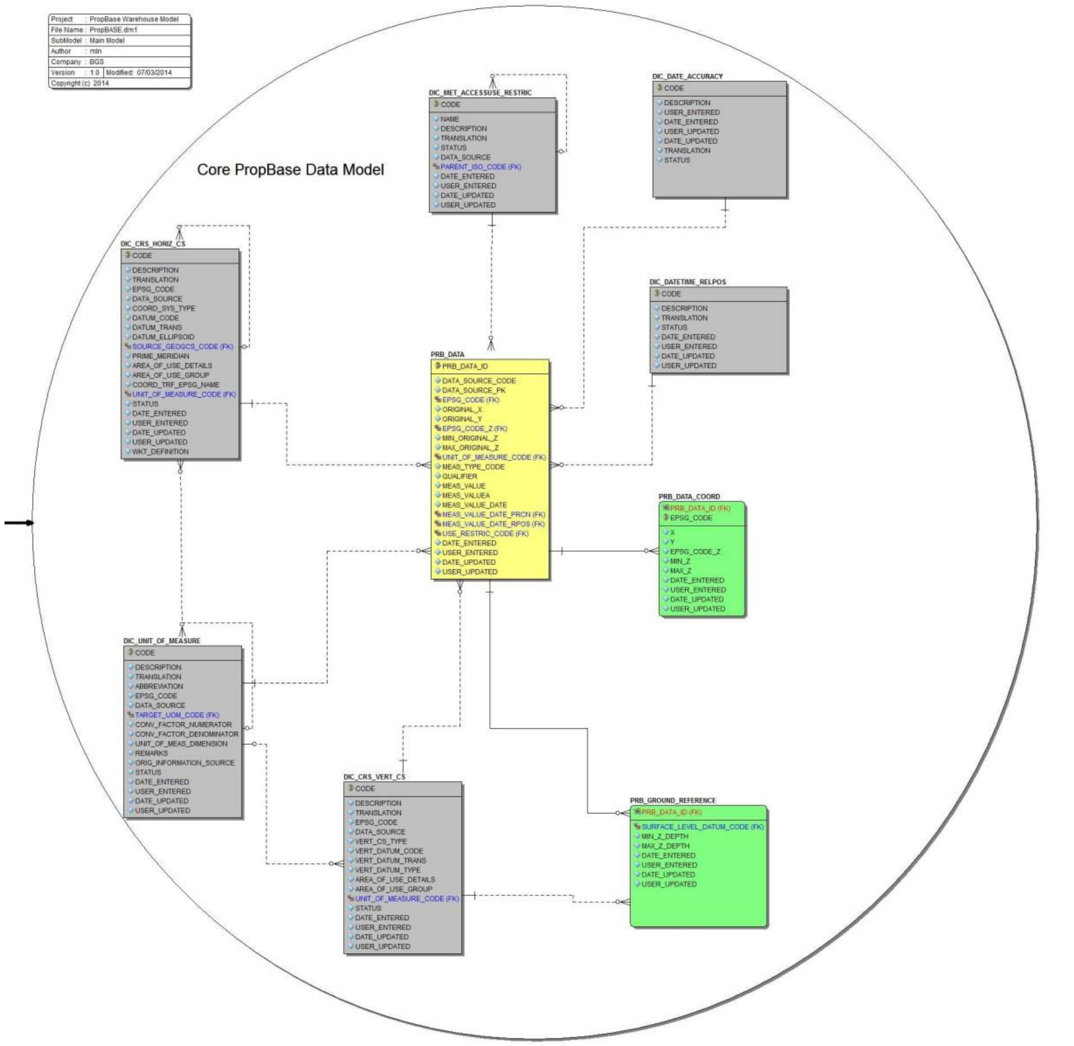


Figure 4: PropBase conceptual core data structure showing linkage between main data tables, subsidiary tables and dictionaries

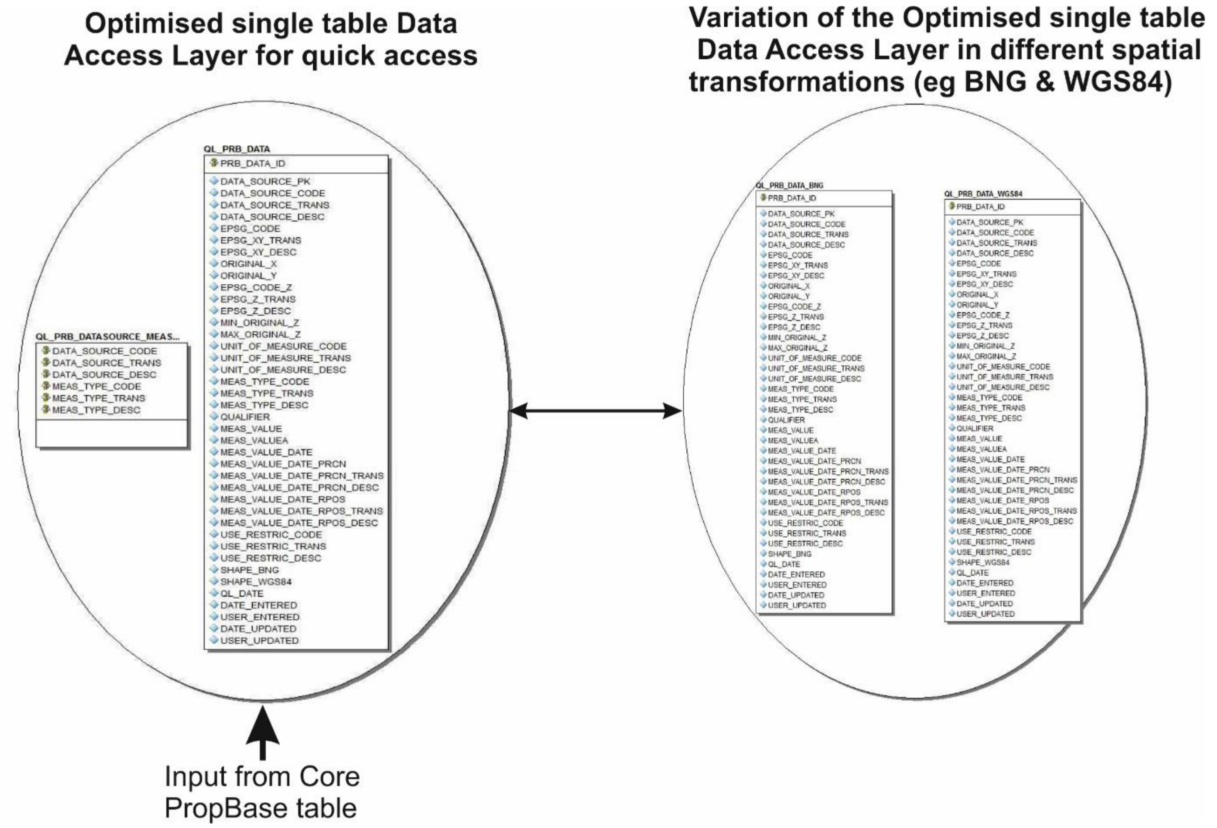


Figure 5: Conceptual Data Model diagram of optimised single table data warehouse layer for query access

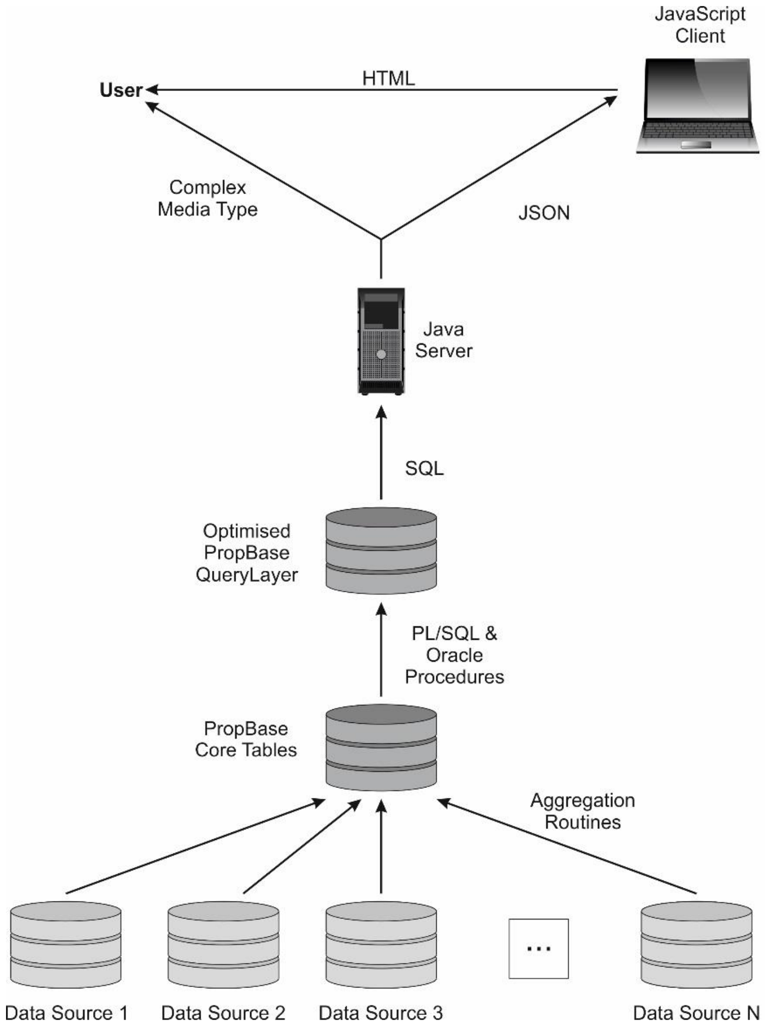


Figure 6: Diagrammatic representation of the entire PropBase data structure from multiple input data sources to human interface

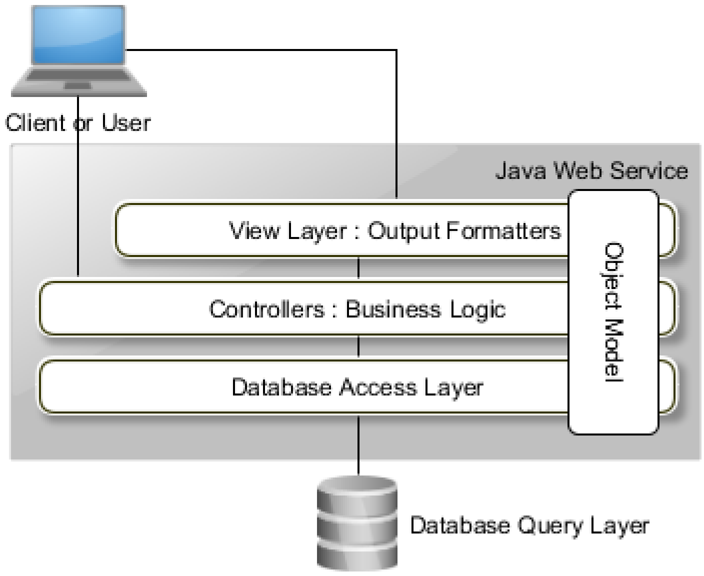


Figure 7: Diagrammatic representation of the PropBase Model View Controller Architecture

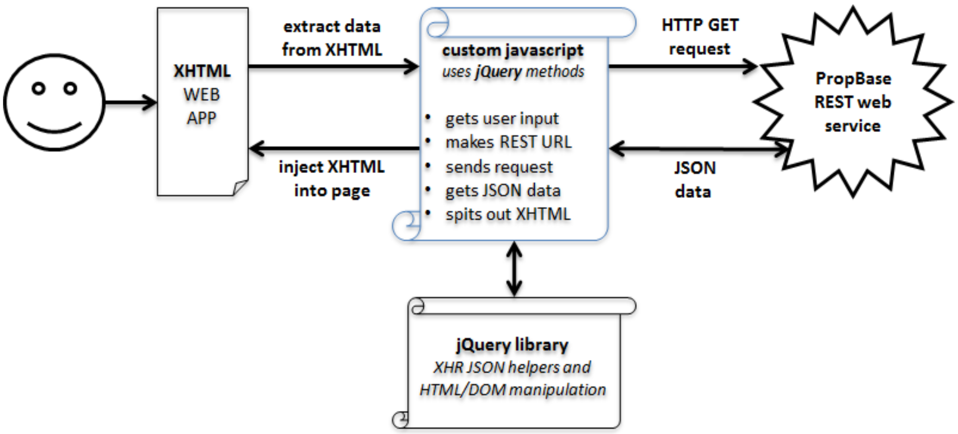
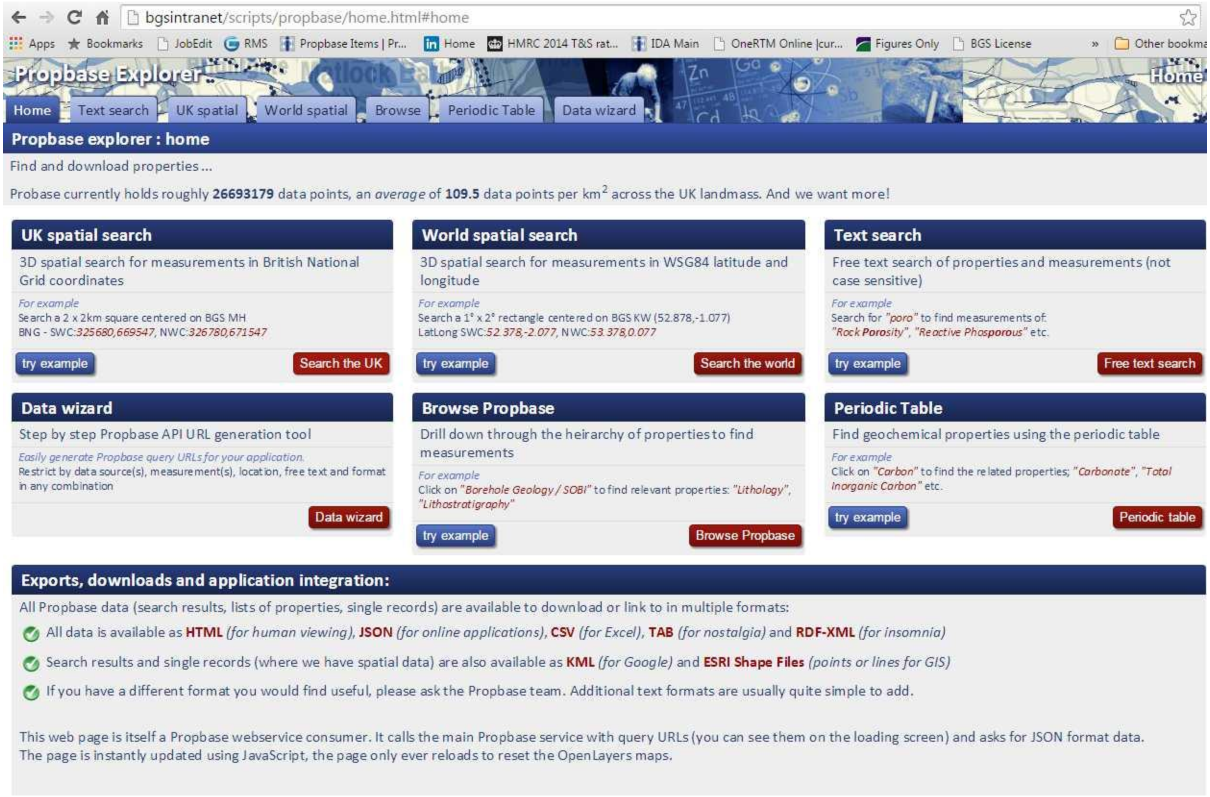


Figure 8: Diagrammatic view of how the PropBase Explorer's Javascript client interacts with the Webservice



Propbase explorer : home

Find and download properties ...

Probase currently holds roughly **26693179** data points, an *average* of **109.5** data points per km^2 across the UK landmass. And we want more!

UK spatial search

3D spatial search for measurements in British National Grid coordinates

For example
Search a 2 x 2km square centered on BGS MH
BNG - SWC:*325680,669547*, NWC:*326780,671547*

World spatial search

3D spatial search for measurements in WSG84 latitude and longitude

For example
Search a 1° x 2° rectangle centered on BGS KW (52.878,-1.077)
LatLong SWC:*52.378,-2.077*, NWC:*53.378,0.077*

Text search

Free text search of properties and measurements (not case sensitive)

For example
Search for *"poro"* to find measurements of:
*"Rock **Poro**sity"*, *"Reactive Phos**poro**us"* etc.

Data wizard

Step by step Propbase API URL generation tool

Easily generate Propbase query URLs for your application.
Restrict by data source(s), measurement(s), location, free text and format in any combination

Browse Propbase

Drill down through the heirarchy of properties to find measurements

For example
Click on *"Borehole Geology / SOBI"* to find relevant properties: *"Lithology"*, *"Lithostratigraphy"*

Periodic Table

Find geochemical properties using the periodic table

For example
Click on *"Carbon"* to find the related properties; *"Carbonate"*, *"Total Inorganic Carbon"* etc.

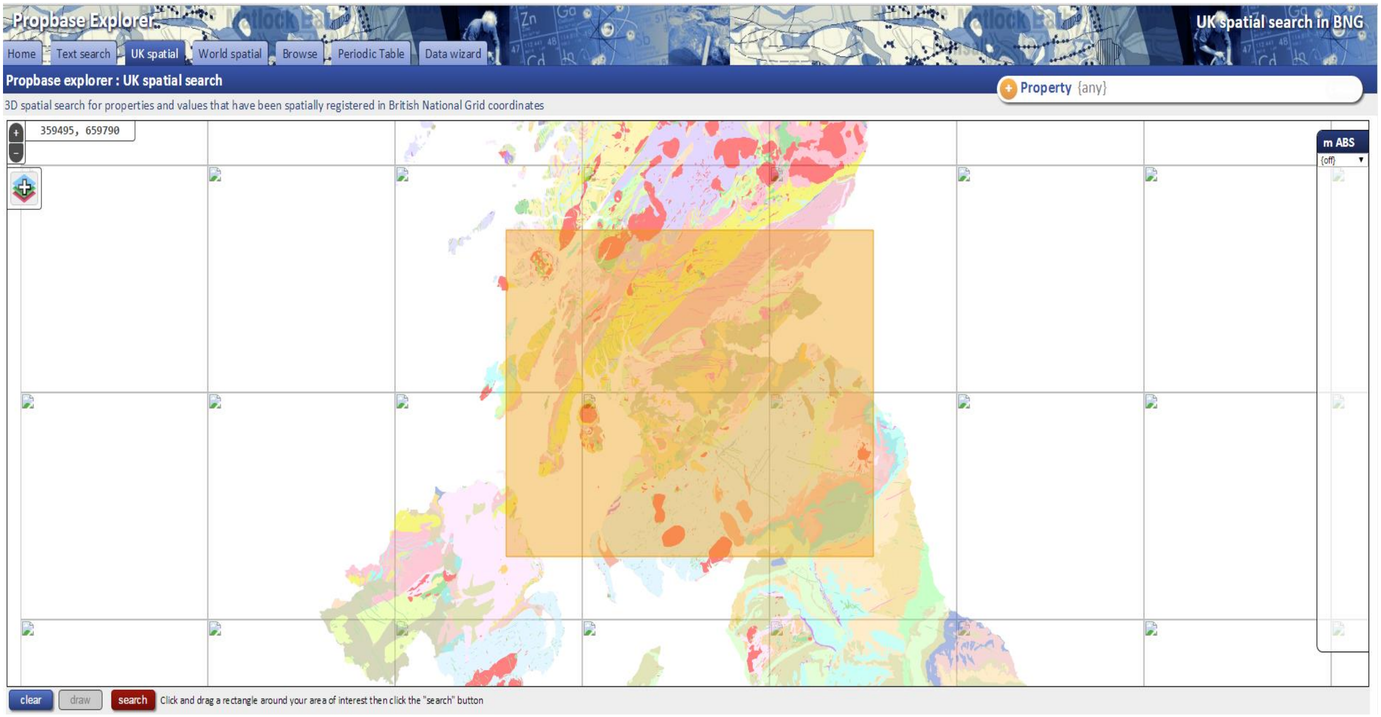
Exports, downloads and application integration:

All Propbase data (search results, lists of properties, single records) are available to download or link to in multiple formats:

- ✓ All data is available as **HTML** *(for human viewing)*, **JSON** *(for online applications)*, **CSV** *(for Excel)*, **TAB** *(for nostalgia)* and **RDF-XML** *(for insomnia)*
- ✓ Search results and single records (where we have spatial data) are also available as **KML** *(for Google)* and **ESRI Shape Files** *(points or lines for GIS)*
- ✓ If you have a different format you would find useful, please ask the Propbase team. Additional text formats are usually quite simple to add.

This web page is itself a Propbase webservice consumer. It calls the main Propbase service with query URLs (you can see them on the loading screen) and asks for JSON format data. The page is instantly updated using JavaScript, the page only ever reloads to reset the OpenLayers maps.

Figure 9: Main landing page for the PropBase Data Explorer tool showing multiple searching criteria for identifying physical property datasets



Propbase explorer : UK spatial search

3D spatial search for properties and values that have been spatially registered in British National Grid coordinates

Click and drag a rectangle around your area of interest then click the "search" button

Figure 10 PropBase Explorer tool showing a spatial search for physical property data within the for an area of interest defined by a mapped rectangle