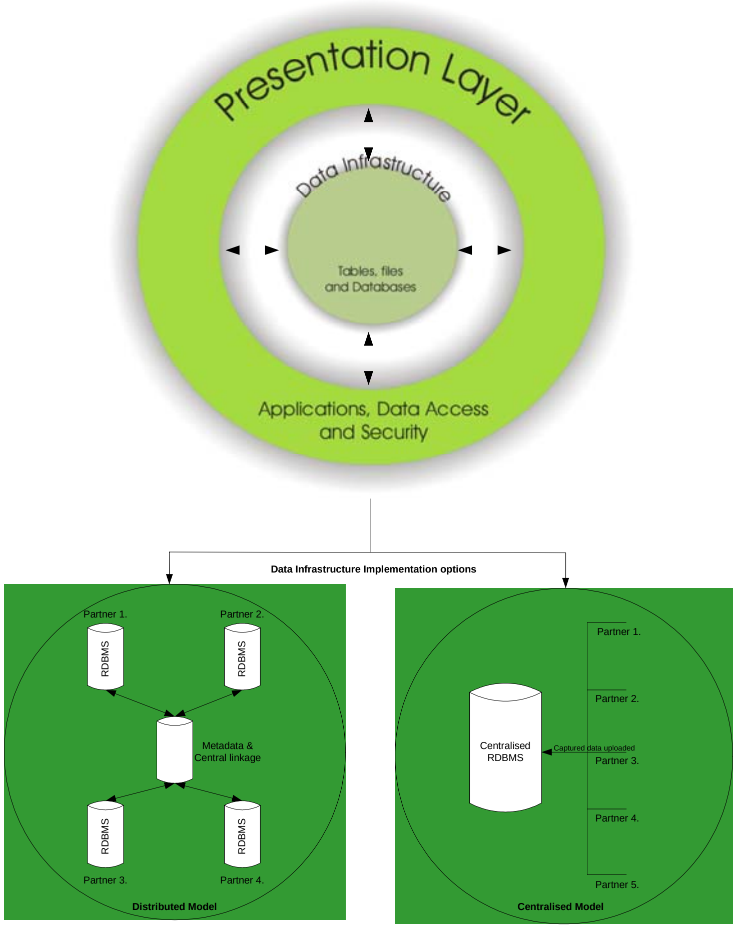
British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

A 'Data System' to support a Geosphere Characterisation Programme - An Awareness Document

Information Management and Environment & Hazards Programmes

Commissioned Report CR/05/042N



BRITISH GEOLOGICAL SURVEY

INFORMATION MANAGEMENT AND ENVIRONMENT & HAZARDS PROGRAMMES

A 'Data System' to support a Geosphere Characterisation Programme - An Awareness Document

G Baker

Contributor/editor

R Shaw, J Careless

Keywords

Report; Database; Geosphere; web-GIS; Portal; Data Model.

Front cover

Proposed Information System model.

Bibliographical reference

BAKER, G. 2005. Scoping Study to define requirements for the design and implementation of a database system for a Geosphere Characterisation Programme. *British Geological Survey Commissioned Report*, CR/05/042N. 32pp.

Keyworth, Nottingham British Geological Survey 2005

BRITISH GEOLOGICAL SURVEY

The full range of Survey publications is available from the BGS Sales Desks at Nottingham, Edinburgh and London; see contact details below or shop online at www.geologyshop.com

The London Information Office also maintains a reference collection of BGS publications including maps for consultation.

The Survey publishes an annual catalogue of its maps and other publications; this catalogue is available from any of the BGS Sales Desks.

The British Geological Survey carries out the geological survey of Great Britain and Northern Ireland (the latter as an agency service for the government of Northern Ireland), and of the surrounding continental shelf, as well as its basic research projects. It also undertakes programmes of British technical aid in geology in developing countries as arranged by the Department for International Development and other agencies.

The British Geological Survey is a component body of the Natural Environment Research Council.

British Geological Survey offices

Keyworth, Nottingham NG12 5GG
☎ 0115-936 3241 Fax 0115-936 3488
e-mail: sales@bgs.ac.uk
www.bgs.ac.uk
Shop online at: www.geologyshop.com

Murchison House, West Mains Road, Edinburgh EH9 3LA
☎ 0131-667 1000 Fax 0131-668 2683
e-mail: scotsales@bgs.ac.uk

London Information Office at the Natural History Museum (Earth Galleries), Exhibition Road, South Kensington, London SW7 2DE
☎ 020-7589 4090 Fax 020-7584 8270
☎ 020-7942 5344/45 email: bgslondon@bgs.ac.uk

Forde House, Park Five Business Centre, Harrier Way, Sowton, Exeter, Devon EX2 7HU
☎ 01392-445271 Fax 01392-445371

Geological Survey of Northern Ireland, Colby House, Stranmillis Court, Belfast, BT9 5BF
☎ 028-9038 8462 Fax 028-9038 8461

Maclean Building, Crowmarsh Gifford, Wallingford, Oxfordshire OX10 8BB
☎ 01491-838800 Fax 01491-692345

Sophia House, 28 Cathedral Road, Cardiff, CF11 9LJ
☎ 029–2066 0147 Fax 029–2066 0159

Parent Body

Natural Environment Research Council, Polaris House, North Star Avenue, Swindon, Wiltshire SN2 1EU
☎ 01793-411500 Fax 01793-411501
www.nerc.ac.uk

Foreword

This report is the published product of a study by the British Geological Survey (BGS) into the requirements necessary for design and implementation of a database system for a possible geosphere characterisation programme by UK Nirex Ltd.

The database system outlined within this report is designed to support UK Nirex Ltd aims to demonstrate, on the basis of designs and documents that a programme of geosphere characterisation could be implemented by Nirex, if requested to do so, and that if it is practicable to characterise a site for the development and implementation of a phased geological disposal facility in the UK.

The proposed solution outlines a 'Data System' that we feel will fulfil the requirement to correctly and professionally hold, query and manage the data from a geosphere characterisation programme. Where applicable details of any horizon IT developments have been included in this report to aid in decision making and highlight future development pathways of the proposed solution.

Acknowledgements

A large number of individuals in both the 'Information Management' and 'Environment and Hazards' programmes have contributed to this investigation and report. This assistance has been received at all stages of the study. In addition to the collection of data, many individuals have freely given their advice, and provided specialist knowledge important to this scoping study. Key staff have helped to review draft chapters of this report. Of the many individuals who have contributed to the project we would particularly like to thank the following:

Mr J R A Giles

Mr J Careless

Dr R P Shaw

The authors would like to thank the following named individuals for their specific technical contributions to the scoping study presented within this report.

Mr B Cannell

Mr M Nayembil

Contents

1 Introduction

This report is aimed at informing programme and project managers at UK Nirex Ltd regarding a proposed method for an integrated database system to centrally store the data holdings gathered from a possible “Geosphere Characterisation Programme”. The report deals with the wide variety of IT options available while clearly identifying a ‘Data System’ model (Data Infrastructure Core + Presentation Layer) that BGS believes is the baseline IT system needed for this task. The data system and associated ‘presentation layer’ options are presented for consideration in a non-technical fashion.

Vision:

The integrated geoscience database suggested here will support a Geosphere Characterisation Programme by providing an easy to manipulate and use database that will act as the primary repository for all geoscience information collected during the course of the programme. The conceptual information system described in the report falls into two broad parts; the ‘Data Infrastructure Core’ and a range of options that we feel would be best incorporated within a presentation layer (applications and customer facing functionality).

The system will provide use with a single, reliable information resource for the whole of the programme. It will be the primary repository from which all other information products will be sourced. Standards established before the start of the programme and implemented within the kernel of the database architecture and its management procedures will contribute to the programme quality system.

The authorised user will be able to access the information from any Internet enabled computer using a standard web-browser, such as Microsoft Internet Explorer. The use of virtual private networks (VPN) will ensure security whilst not imposing undue limitations on use.

The user will see a typical web page interface providing them with a range of options to access the databases, manipulate the data and produce output that can be viewed or downloaded for manipulation desktop applications or included in reports. The integration of Internet Mapping applications will permit the visualisation of the spatial component of the information in the familiar map paradigm.

The added-value components will cover extended information system functionality such as:

- A portal
- Complex visualisation using GRID enabled technologies
- Automated capture of field data
- Electronic Records and Document Management System

The proposal builds on the experience gained by BGS in the development of the NDGD during the 1990s and of the development of the PADAMOT Database and web site during the early 2000s. This is augmented by the experience gained by enhancing the BGS database and consulting and advising on geoscience information systems of the geological survey organisations of Ireland, Mozambique, Afghanistan, Saudi Arabia, and Papua New Guinea.

BGS Capability:

BGS is a world leader in applying information technology to the geosciences with over 30 years experience, centred upon geoscience database design, development, interrogation and integration. We have been working upon our corporate geoscience database for over 20 years. It is currently based within an enterprise-level object-relational database product, Oracle®. We have over 2500 geoscience data and dictionary tables that are viewable to all BGS staff, with a subset of these being viewable by the general public via web-enabled applications within ColdFusion®, and spatially through an ESRI ArcIMS® interface. BGS have also been pro-active in the release of some BGS dictionaries, such as the Lexicon and Rock Classification Scheme, either through joint ventures with agencies like the AGS or publicly downloadable from our website (www.bgs.ac.uk). Our aims in releasing these schemes are to promote them as standards within the geoscience industry in the UK (and the world where applicable), and to subsequently allow for easier integration of data through their adoption.

Through internal BGS information projects such as ‘BGS-geoIDS’ resources have been devoted to generating database design, naming, application standards and best practice, for use in BGS IT developments. This has enabled new BGS database development to fit into our corporate data holdings (when project datasets become corporate data) with much less subsequent modification or alteration. We specifically targeted the generation of a corporate dictionary standard (design, names, constraints, audit functions) for use in the generation of new or updated dictionaries (that encapsulate a geoscience classification) and the regulation of corporate data to reference the approved corporate dictionaries. It is essential in these days of geoscience investigation and increasingly complex geoscience product generation to be able to relate geoscience datasets to each other to allow the modern geoscientist to have access to the entirety of data available from within his or her organisation when producing new work or interpretations. This can only be achieved successfully and efficiently through the interrelation of data using common data, dictionaries and design standards. A centralised corporate database repository also helps with elements of the system such as; data backup, query speed and standardised data management.

Recent growth within the IT department in BGS has been in the areas of web and web-enabled application skill sets. We have invested a great deal of development time over the last two years within the ‘BGS Corporate Digital Data Framework’ project in the creation of a single corporate web-enabled application to query our geoscience databases. This application entitled the ‘Intranet Data Access’ or IDA has helped our geoscientists to quickly gain access to the data held in our digital data holdings, through a standard web form interface. The goal of the IDA was to be simple in format, rich in functionality and allow those not having great database skills (Structured Query Language) access to BGS RDBMS data. The IDA is currently being extended to link or incorporate further, more complex viewing applications (previewers) thereby allowing the geoscientists to visualise the data not only in a table or query format but also as graphical logs in formats such as Scalable Vector Graphics (SVG) or a GIS layer within the ESRI ArcIMS® application iGDI. We hope in the near future to also include simple block models / enhanced visualisation techniques utilising Java3D and other web technologies. The IDA will then truly have become a BGS data portal.

BGS has demonstrated all of these developments and standards within its own digital data holdings and in joint partnership with other survey’s, such as the Geological Survey of Ireland or Knowledge and Research projects (DFID/KaR) such as ‘Strategies for Maximising Geoscience data’. The creation of an integrated geoscience database and the associated corporate web-enabled spatial applications has been a protracted task and BGS has gained significant experience during the work.

History: Nirex Digital Geoscience Database - Lessons Learnt

In 1995 the BGS was awarded a contract to design, implement and manage (with a dedicated help desk) the Nirex Digital Geoscience Database (NDGD). The NDGD was then managed by BGS, with many contract extensions, until it was mothballed after the final Public Inquiry and a government decision to not proceed with an underground test characterisation facility in 1997.

Many key lessons where learnt from the design, implementation and management of the NDGD that are especially relevant to this proposed information system.

1. The integrated geoscience database with a suite of access interfaces and development tools to address the differing questions being asked of the data (be that business, geoscience specialisms or general data queries), was a fundamental success of the project and ensured that UK Nirex Ltd could produce the variety of reports and deliverables required by the Public Inquiry.

2. It was clearly demonstrated that the integrated geoscience database and, crucially, a ‘data acquisition plan’ were needed prior to the main data-gathering phase of the work commencing. This was not possible for the NDGD but is a key recommendation for any future developments.

3. Prior to the design stage of the NDGD, or a future development, it would be crucial to have already agreed the programme scientific dictionaries, these will encapsulate the geoscience classifications that are to be used by any project teams. This would ensure that all data gathered was constrained against the same dictionary classifications. This was amply demonstrated in the NDGD with respect to the changing nomenclature of fractures in differing boreholes drilled and logged at differing times. This proved expensive to back interpret and subsequently relate to each other within a single cohesive fracture dataset. The entire cost could have been negated if common programme-wide dictionaries and standards had been implemented prior to the start of the project data gathering phases.
4. The help desk proved invaluable in order to meet the needs of the end user community accessing the NDGD. It helped bridge the IT/RDBMS knowledge gap between the staff requesting complex data queries of the data holdings, and the integrated geoscience database.

5. Data management was carried out by BGS for UK Nirex Ltd upon the NDGD. BGS acted as data custodians, making updates and changes as appropriate or instructed by Nirex authorised project members. It was recognised that assigning full active data custodian roles within Nirex would have been a preferable method of ensuring higher levels of data integrity and hands-on data management.

6. Due to the constraints of the database technology available in 1995 it was not possible to hold all the variety of data types from the Sellafield investigations. This did lead to other repositories of data, such as models or documents, that could now be better held as part of a single system under a new future system.

7. Modelling using the data extracted from the integrated geoscience database (NDGD) was carried out in many locations across the UK making collaboration a difficult task after many iterations of SQL scripts had been developed to extract data in formats that suited the modelling packages in use at that time. It was expressed that a more streamlined and efficient pathway to modelling and collaborative work in this area would be of particular benefit in any future implementation of such a system.

8. Snapshots of the data holdings to match stages within the investigation, for example just prior to any Public Inquiry, were carried out during the lifecycle of the NDGD project. These snapshots were often offline tape archives and not an easily accessible form for subsequent comparative data queries.

9. Clear stakeholder positions are required within the project held by partner geoscience companies and consultancies. This is to ensure that they “buy in” to the standards proposed for data gathering and use of the information system.

10. The database is the definitive source of all derived information, reports are an output of the database not a master copy of any dataset contained within.

2 Assumptions

The following is a list of assumptions that are necessary to constrain what could easily be a "blue-skies" discussion document.

1. A Geosphere Characterisation Project would involve holding data for multiple sites around the UK.

2. Each site will generate considerably larger data volumes than the previous NDGD 'Sellafield' investigations.

3. The range of data types being presented for inclusion in the RDBMS will be similar to the previous Sellafield investigation with an emphasis on including time series data plus data types such as images, interpretations, maps or profiles that are now more easily integrated into a single system. *This is clearly laid out in the initial UK Nirex Ltd 'Site Investigation Flow Diagram'.*

4. To ensure that the proposal does not become overly complex with many possible combinations. A conceptual model for 'Data System' is presented within this report. This is composed of a 'Data Infrastructure Core' and a 'Presentation Layer', which BGS suggest should be of the professional standard required for implementation when undertaking a programme of this magnitude.

5. Two broad implementations of the 'Data Infrastructure Core' ('centralised' and 'distributed') along with discussion and comparison are included within this report.

6. A range of options for inclusion within the 'Data System - Presentation Layer' has also been included where we feel these are relevant to the aims of the overall system.

7. Costs (staffing, hardware and software) have not been included within this 'awareness document'. Additional decisions from UK Nirex Ltd and further logistical clarification of the "Geosphere Characterisation Programme" would be required prior to the creation of scoping study or costed project proposal.

8. Developments within IT are fast paced and often significant developments can be released that may supersede some of the suggested options within this report. We suggest that it is advisable to undertake an updated review of physical implementation models encapsulated through key software manufactures if this report is still being considered more than 18 months after its original creation. The proposed 'Data System' will still be valid but the implementation models and software (with associated built in functionality) may well have developed.

9. This report is based upon the lessons learnt from the previous investigation database, the Nirex Digital Geoscience Database (NDGD) and subsequent BGS information projects both internal and commercial.

3 Data System – Data Infrastructure Core

Definition

The 'Data Infrastructure Core' is defined within this solution as the RDBMS configuration to house the data holdings from the proposed Geosphere Characterisation Programme.

The 'Data Infrastructure Core' - Design

The following tasks are considered to be the core elements necessary for the implementation of a high quality system to hold the variety of geoscience data in an array of differing data types.

1. Database design to hold geoscience data gathered from multiple sites within the UK. *Centralised single RDBMS or distributed and linked RDBMS physical implementation options both viable models.*

2. Approved scientific programme dictionaries and contents against which the entire data holdings will be constrained. Classification schemes to be used by all project teams no matter which site they are working upon or where the data is initially stored.

3. Data acquisitions plan to specify the data types and volumes to be gathered by the multi-site project teams.

4. Extensive Metadata: 'Discovery' metadata regarding datasets held by the system and 'Technical' and 'Application' metadata regarding the RDBMS objects implemented. Additional metadata to deal effectively with datasets, interpretations, profiles, maps and models their complex interrelationships and composition will also be an essential "Data Infrastructure Core" requirement.

5. The spatial element of the geoscience data gathered from site investigations to be held within a spatially enabled database.

6. Time series data from characterisation sites to be held within the RDBMS.

7. Documentation of the proposed system and its components through reports and project standards (data model, implementation scripts, scientific dictionaries, integrity rules).

8. 99.5% availability of RDBMS instance(s) with backup and live hot spare database facilities in the case of any failure.

9. Implementation within an enterprise-level RDBMS product using kernel rules to ensure high levels of data integrity and core management.

10. Robust server architecture model to ensure a capable system with facilities to seamlessly maintain service in the case of component or connection failure.

11. Detailed permissions model that truly reflects the working relationships of the differing project teams that will be investigating the multiple sites within the UK under the Geosphere Characterisation Programme. A detailed permissions model will address key issues such as entitlements and complex project working teams on multiple sites through the use of additional interrelated roles. The data management and audit functions would also be contained within specific roles set up within the RDBMS and web-portal applications.

4 Benefits and Drawbacks

Benefits:

An integrated geoscience database holding the data from all geosphere characterisation sites constrained to known geoscience classification standards (programme approved geoscience dictionaries) allows complex query, retrieval and interrogation of the data holdings. **(Strength)**

A single centralised RDBMS model holding all geosphere charactisation data to ensure resources and support can be focussed upon the main central database instance. The strengths of this model are that hardware/software/development costs are kept to a minimum and easily identified, maintained high levels of data integrity, centralised system support and minimum system complexity. **(Strength)**

The parallel model to this is a distributed RDBMS solution. Through the use of the latest RDBMS products and/or the use of linkage technologies it's become possible to join distributed RDBMS instances to form a cohesive 'logical data holding' for a programme such as the geosphere characterisation. This does add significant infrastructure costs and complexity to the system developed as linkages can become corrupted or broken therefore tainting the logical data holdings. It also becomes critical to ensure complete adoption of programme standards to ensure linkage of the disparate system remains possible. The distributed model is used primarily in the merging of existing long-standing data sources for industry significant data holdings but does crucially allow greater levels of system resilience while also allowing data gathering partner organisation to hold the data at their locations managed to a common set of rules, therefore encouraging programme buy-in. **(Strength)**

The consistent database can now hold larger objects and new data types allowing inclusion of data previously held on separate systems therefore allowing common data management and quality assurance procedures to be maintained for all data held. **(Strength)**

Online accessible metadata regarding the datasets held within the system that can be disseminated to the public via a web site. Technical metadata regarding the objects within the RDBMS allowing all project staff to easily identify the database objects held (tables, views or dictionaries). *(Definition: Technical metadata – metadata on objects contained within the RDBMS such as tables, constraints and views)*. **(Opportunity)**

A publicly accessible web site for dissemination of information about the 'Geosphere Characterisation Programme' will prove most useful during any programme of work that may be carried out. A lesson learnt from previous investigations into Sellafield was that the work being carried out to meet the aims of the programme needed to be widely disseminated to ensure that speculation and uninformed opinion did not cloud perception of the work being carried out. Metadata in its various forms would prove a useful online addition to this proposed site. **(Opportunity)**

Implementation of the integrated data model within a RDBMS product, will ensure that the geoscience data is structured in such a way that it conforms to RDBMS principles (when the design is carried out by trained 'Systems Analysts') and so can be easily related to each other for complex querying. **(Strength)**

High levels of server availability are essential to this system, a figure of 99.5% availability plus would be considered a requirement for the disparate projects teams. It is crucial that server failure and bandwidth contention does not impede project tasks. **(Risk)**

Full documentation of the data model, and the interfaces developed will ensure that this system can be maintained by a larger number of qualified IT professionals, therefore removing the reliance upon a small number of original IT developers. Documentation also helps to ensure that the developed system is fully transparent in its functions and assumptions during any query, interpretation or presentation of result sets. **(Strength)**

Cost effective but capable server architecture, to ensure non-contention between the integrated geoscience databases when dealing with queries from the 'Presentation Layer' (web-enabled applications or web-GIS components). A separate suite of development servers will be required to ensure that any on-going or new developments do not impact upon the 'live' system in any way. **(Opportunity)**

An extensive permissions structure to ensure programme 'roles' can be implemented upon the database and therefore accessible through the presentation layer. This will allow project teams and staff to be assigned to the correct role(s) depending upon the data they are querying or updating. **(Strength)**

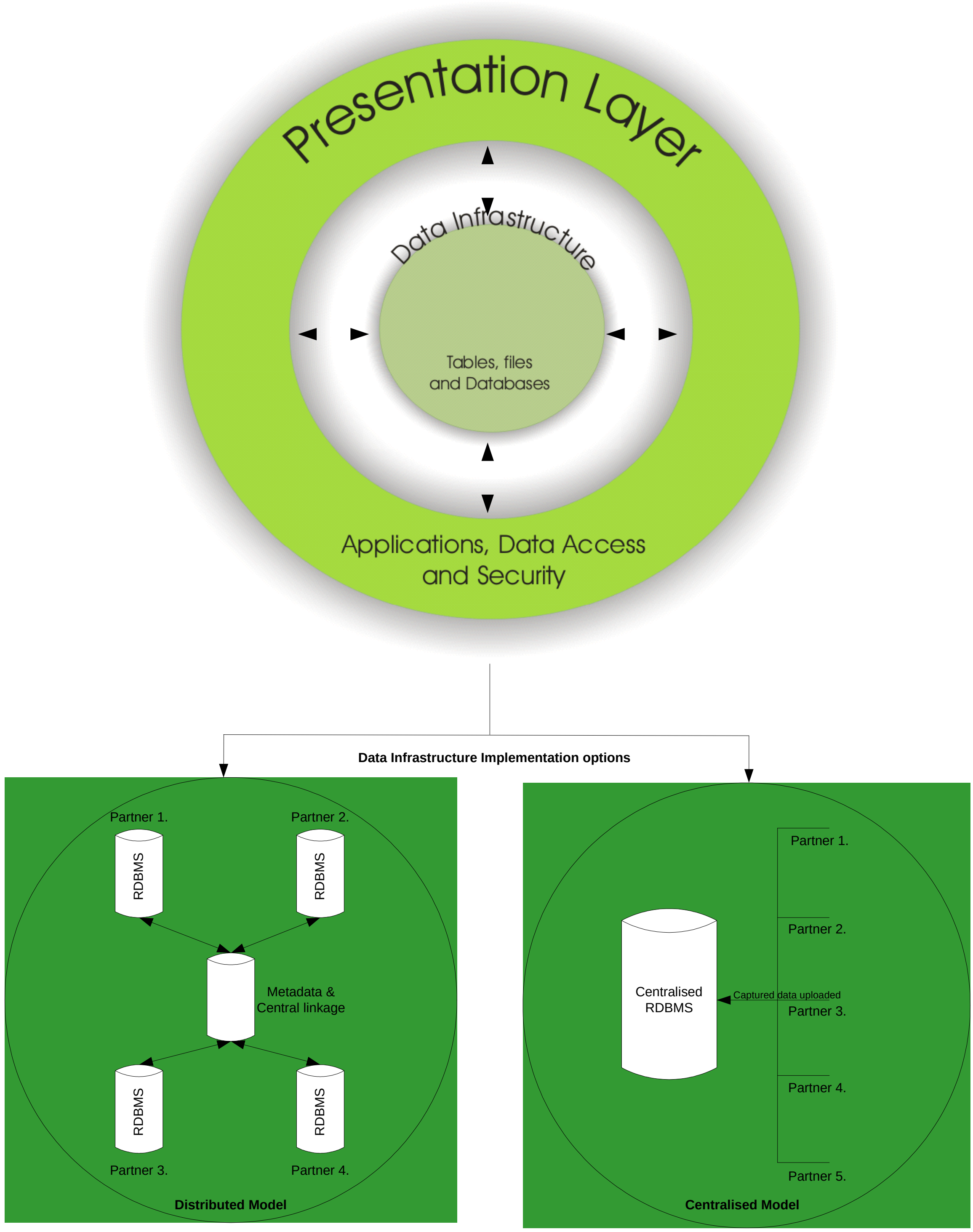
Drawbacks:

High server availability and bandwidth, nominally set at 99.5% is considered to be a suitable requirement of the proposed 'Data system' but does incur a high cost for server hosting and maintenance by a suitable selected Internet Service Providers (ISP). To ensure that such an investment is also rewarded with a robust and highly available system a RDBMS server located upon a differing ISP / Server farm as a hot spare would also be required. These requirements add significantly to the cost. **(Weakness – cost)**

Enterprise business scale RDBMS products will be suitable for the project data holdings, either within a centralised or distributed physical implementation model. An example, of such a high-end product is Oracle®, this does attract fairly significant licensing costs but has significant proven industry stability and a feature set more aligned to larger data holdings and multiple data types. **(Weakness – cost, but mitigated by the proven industry stability such a product offers)**

Extensive permissions model may ultimately fail to be flexible enough to cope with the diverse range of project teams and staff that need differing permission to access and/or update data set on the 'Data System' unless suitably investigated and modelled **(Risk – can be mitigated by ensuring comprehensive modelling of permissions mapped to project task and programme structure)**

5 Diagram of proposed 'Data System'



6 Discussion of benefits and enhancements over previous 'NDGD' system

1. Designs for the 'data acquisition plan', 'programme geoscience dictionaries' and 'information system' in place prior to data capture. This will ensure that programme standards compliance is an essential component to the work carried out by characterisation project teams. This promotes greater interrelation of data holdings, easier data loading and easier management within the RDBMS.

2. Greater range of data types (including time series datasets) can now be held within the RDBMS due to recent developments within these software technologies and parallel developments in increased storage capacity.

3. Spatial data can be stored with the attribute data in the RDBMS via the use of add-on components, current examples are Oracle Spatial or ESRI SDE rather than separately within the proprietary GIS files, as in previous investigations.

4. Data Management and QA procedures for the entire data holdings can be applied and enforced.

5. Web access to the entire data holdings is now possible using web-enabled applications and previewers (Presentation Layer). The range of information is highlighted below;

   - Work programme / project team
   - General site location and overview
   - Site geoscience data including spatial
   - Programme documents and reports
   - Large objects such as modelled surfaces, interpretations, profiles or maps

6. Central storage of the entire programmes' data holdings allows project team sharing in real time of new information, interpretations and any lessons learnt.

7. Central management and control of the data holdings can lead to a cost saving over disparate data sets held on many differing sites in many differing formats collated to differing standards.

7 Data Management – Issues for Consideration

The sections below are a discussion of data management issues that need to be addressed to enable this system to succeed. These should not be underestimated and are essential to the success of the proposed information system.

Scientific Languages / Dictionaries

A fundamental task for the Geosphere Characterisation Programme is the definition of a comprehensive set of terms that will be applied across all project teams. These will be resolved into a series of codes for use within the database by “Look Up tables” (dictionaries) and are essential to the ability to centrally combine and relate data from various geoscience disciplines and, crucially, in this case differing sites. This task can set the tone for the success or failure of the entire project, centralised database projects (and for that matter distributed database projects with portal linkages) cannot begin to work without common languages to relate data.

Data Custodians

Data Custodians are essential within the “centralised RDBMS with web-enabled access to project teams” model being proposed. The data tables and, crucially, the constraining dictionaries need to be owned by specific individuals within the Geosphere Characterisation Programme. This allows an overarching data management function to control the data held in accordance with the aims of the programme and not any specific short-term project goals.

Data Custodians are often themselves attached to projects and may be responsible for deliverables within the programme framework but must be specialists in the dictionary areas that they are assigned and capable of stepping back from any local project tasks, in order to manage the tables they have been assigned corporately (from the programmes perspective). Data Custodians work closely with the Database Administrator to ensure rules compliance for the tables they manage and to ensure that the design, database kernel rules and content are held according the standards outlined by the project data management plan.

Metadata

Metadata will prove a crucial requirement within any new ‘Data System’. The wide range of data and complex relationships between the initially outlined geoscience and logistical datasets, interpreted products, processes and final models will need to be easily identifiable to ensure that full and open audit tracking is present within the proposed system. Many levels of metadata (‘data about data’) will be required within the proposed system; all should be created to accepted ISO metadata standards and be easily searchable by all programme partners and staff.

Audit trails – traceability

In a large centrally held database it is important to be able to track any changes made to the data held. This ensures the programme can have confidence in the data management from a variety of data custodians on differing project teams. This is carried out by the use of comprehensive data audit tables and database functions. These tables log the changes (insert, update or delete functions) made to any dataset within the RDBMS. Development staff or the Database Administrator can then investigate if problems occur and rollback to previous data versions if errors have been introduced to the 'live' data set.

The currently developed audit system could be expanded to log reasons for alteration if that would be of significant benefit upon a sensitive dataset where open and honest reporting will be essential.

Problems with the transfer of data between systems

A key lesson learnt from the NDGD was that consideration should be given to the transformations carried out during re-formatting and loading of field data into the central RDBMS. The data sets that will be gathered for the characterisation programme are often highly detailed, captured to great levels of precision and accuracy. These precise and expensive datasets can easily be undermined without careful and considered loading through acknowledged procedures. ODBC drivers that link differing software systems do by their nature manipulate data when transferring between systems.

It should be clearly stated that the quality of the database system implemented is not only based upon a sound design, it also crucially depends upon the data placed within it and the subsequent management carried out upon these data holdings.

8 'Data System' – Extension Options

Definition

Extensions are defined as 'additions' to the proposed 'Data System' that will enhance its features and lead to a more robust and secure information system being implemented. These additional features come at an additional cost and consideration of the benefits versus the added cost needs to be undertaken by the commissioning organisation to ensure that value for money is achieved.

Option 1 – Enhanced Security

8.1.1 Virtual Private Networks (private and public web sites)

Use of a Virtual Private Network (VPN) to authenticate and secure any transactions between a client and the central RDBMS/applications. The VPN would be utilised whether it was a simple data query or a more sensitive upload of some new site-specific geoscience data. The 'public good' web site would fall outside the VPN structure to ensure its availability to all web browsers.

8.1.2 Option 2 – Enhanced Server Availability / Bandwidth

8.1.2.1 Enhanced availability via hosting on a CR server farm

Enhanced access of the 'live' servers upon the WWW through the use of an ISP hosted server farm. This involves the server being hosted at a server farm and located upon the Internet using an Internet Service Providers (ISP) existing web linkage.

ISP's provide a range of options within the service level agreement purchased dependant upon;

- the software and technologies you wish to use
- the level of 'availability' you wish to maintain upon these services (e.g. 99%, 99.5% etc..)
- the bandwidth that you feel your main servers will require during a week/month to allow your business to be carried out upon the servers.

It should be noted that costs increase significantly for additional 'availability' or bandwidth requirements of the business. Costs within this environment are particularly variable depending upon many IT market factors and the specific requirements of the service level agreement being negotiated.

There is an option to purchase the necessary hardware, both servers and network equipment, commercial connection to the Internet with suitable bandwidth and some skilled management staff to enable UK Nirex Ltd to host their own servers for the 'Data System'. This would however be by far the most expensive option, Internet Service Provision (ISP) is a dynamic and extremely competitive market and cheaper server hosting with an SLA would be possible via outsourcing this programme requirement.

8.1.2.2 Additional servers, clustered resources (database, web)

Server availability, both with respect to high levels of demand from user community and also from failure of components within a server (including software services failing), is often addressed by the use of clustered servers. Clustering servers allows requests for work tasks to be shared out to a greater number of server CPU's, therefore allowing a faster response rate and also has the added benefit of being able to cope with the failure of a machine (or more dependant upon the number of the clustered servers and their functions). Rebuilding servers to a functional state along with the downtime (from failure and including re-build) can be costly to an organisation or work programme, and seriously undermine deliverables and data management if the system is non-functional. Clustering adds resilience to an IT system. A similar approach can also be taken by the use of a "hot-spare" server upon which a copy of the current data and/or applications is also held in order to be used if any failure in the main 'live' system occurs. The use of clustering and hot-spare servers often goes with the requirement for high levels of server availability and bandwidth as the costs to implement these solutions are aligned to an expensive, highly accessible web based system.

Any of the clustering or hot-spare options outlined above can be further strengthened by the use of a tape or disk backup solution. This would be used to recover a server to its previous state in the case of any drive failures. Back-ups can range from simple tape devices run by an existing server, through dedicated servers and robotic tape libraries, to 'hot spare' or disk-based backup within an offsite server. The faster that a backup might need to be in use replacing the failed server or the higher the speed to be able to extract backed up data, the higher the purchase cost (hardware, software and maintenance).

Indicative cost: Second server(s) with added software licences (operating system, database, application and backup) There is a need for additional servers that are mirrors for existing systems and as such need identical software configurations with their own software licences.

9 ‘Data System’ – Presentation Layer Options

Definition

The options outlined below are developments that are essential within a presentation layer built upon the ‘Data Infrastructure Core’. These options enhance the usability and functions of the data holdings and form the components of the ‘Presentation Layer’ of the proposed ‘Data System’.

Web Portal (Web Site and Data presentation through combined viewers)

A public and project information web site regarding the programme including a subset of publicly available data, if appropriate. To improve and assist both the scientific interpretation and general access to the geoscience datasets the addition of a graphical visualisation system would be of value to a standard public web site. Such Data Portals allow the rapid visualisation of spatial extents plus time series datasets as well as more intelligent previewing of structured data such as geophysical well logs and geological downhole logs. All of the functionality required within such a preview and graphical interrogation system can be provided through a 'thin-client' model where the accessing client requires no more than a working Internet connection and a standard web browser.

A Data Portal would allow differing geoscience datasets to be quickly cross-referenced and allow multi-disciplinary teams to select and appraise datasets that are directly relevant to their investigations. Additionally, a Data Portal can provide download functionality to allow the centrally held datasets to be transformed into alternative formats suitable for differing software applications - the scope of this functionality is dependant upon an agreed set of output formats suitable for the range of scientific products in use around the differing geoscience project teams.

Costs for this option will vary depending upon the number of datasets that need to be visualised and the number of previewers required. The baseline implementation will build upon the web-GIS aspects of the core model with additional application infrastructure to support complex data transformations and visualisation.

Complex Visualisation (GRID enabled)

Traditionally, visualising complex 3D models, has required the dedicated use of specialised desktop modelling software. Developments within the UK eScience Programme and GRID initiatives have highlighted the use of GRID systems to provide cluster-based computation and the ability to present fully rendered 3D models to 'low power' clients. In this model it is feasible to provide a single location for the 3D model or simulation to be held and through computational steering, allow remote clients to both visualise (render) the resultant views and control parts of the simulation on laptops and standard desktop machines. In the case of a 3D model each remote client could have full control over the rotation and interrogation of the model, whereas within a running simulation, direct manipulation of the parameters is possible from each remote collaborator.

In order for this distributed visualisation to occur, a GRID (Globus middleware) infrastructure would have to be provided to allow suitable security controls (X.509 digital certificates) over participation within collaborative visualisations. The computational engine required to deliver the models and simulations could be in the form of a single multi-processor machine or a 'dedicated' cluster of machines. The advantages of cluster-based systems are seen through increased scalability and resilience due to the supply of 'redundant' nodes whilst still providing a computational infrastructure.

The costs for this option are defined by:

- Computational hardware infrastructure,
- Distributed visualisation software,
- Software customisation to suit specific needs.

Automatic Upload of Field Based Data

Field based capture of digital data has been an area that BGS has investigated over the last few years exploring a number of possible options including laptops, tablet pc’s and recently HP Ipaq pocket pc’s. This is a complex and difficult area that is split into two main issues to be addressed. The first is the data capture device and designed applications in order to maintain the integrity of the data, and the second is the method of returning this data to the central integrated geoscience database.

Crucial to the success of digital field capture systems is a clear understanding of the elements of fieldwork that are best encapsulated in a digital system. Most observational and automatically logged data types can be captured in modern field-based systems. Drilling, observational core logging, time series data from in-situ testing and especially geophysical data can be recorded using either a propriety system or its not beyond capability to built solutions for flied based capture around laptops or pocket pc technology. This area is particularly in depth area that would need to be discussed with any partner organisations brought on-board for field data capture at any characterisation site. Further discussion of this area falls outside the scope of this ‘Awareness Document’.

9.1.1 Selection of Field based equipment

The type of field device is dictated by the requirements of the field data capture. The more complicated the data needing to be captured the greater the application development task. Well-constrained forms with multiple dictionaries plus field GIS operations will require ruggedised laptops or tablet PC’s, more simplistic point observation and form entry through constrained forms can be carried out by pocket PC's especially useful in positions where a highly mobile solution is required.

Options for consideration are;

- Laptop (possibly ruggedised) with modem or GPRS/3G card
- Tablet pc with modem or GPRS/3G card
- Handheld pocket pc with built-in GPRS/3G

There are numerous other digital devices that can be used to capture field data, these devices are in essence small robust laptop or pocket pc's, normally hard wired to a field device such as a drill rig, geophysical unit or ground monitor to report and store data. These devices often deal in propriety format for their local data holdings but there is no foreseeable problem in capturing this data either in its own transfer format or porting this to a programme authorised transfer format for delivery to the 'Data System'.

Most modern solutions highlighted above can be linked or attached to a GPS for precise location information to be gathered.

9.1.2 Software interface development and customisation

The success of field based digital data capture relies upon suitable interfaces being produced for the gathering of data. Ideally these should be web-based and as such managed upon a thin-client model. This allows central control in a single location of these web forms that will be used to gather remote data. Wherever possible the data should be constrained by the use of the same dictionaries used in the main RDBMS.

GIS data capture or manipulation in the field is better carried out by powerful laptops using software such as ESRI ArcView. Clear procedures on the update of this field data held upon the laptop to the proposed system upon return from the field need to be in place. The GIS industry has yet to fully embrace the distributed working model with a central repository of programme GIS data. ESRI are working towards this goal (along with added integration with RDBMS products) in their ArcServer product but as yet this has not reach the marketplace.

9.1.3 Remote linkage of field based equipment to RDBMS

The range of options for remote connection to a central information system has grown significantly in recent years. The lowest cost option would be dial-up modem usage from laptop, tablet or pocket pc's to upload data after a days field data gathering. Through the use of laptop PC-CARD and pocket PC's which can have built in GPRS/3G wireless abilities, it is possible to live link field based systems to the central information system therefore allowing instant upload of field data to the core RDBMS. The adoption of a more flexible costing model by the mobile networks using charges per MB's of data transferred (upload or download) rather than a permanent open costed line has led to remotely link field staff being a more effective option. There are advantages to this more costly secondary linkage option via GPRS/3G as the field based staff will always be working upon the latest data and will have access to any additional and supporting data from the integrated geoscience database / information system, therefore ensuring that they are not basing new field data observations or work tasks on out-dated core information. These links are now often secured using VPN technology, even from small handheld pocket PC devices.

It needs to be stressed that although the wireless network infrastructure within the UK has been seriously increased in recent years and operators have built GPRS/3G technologies to exploit this, areas can often be found where the connection is weak and so downgraded to smaller line connection speeds, and in the more remote places sometime no connection at all can be achieved.

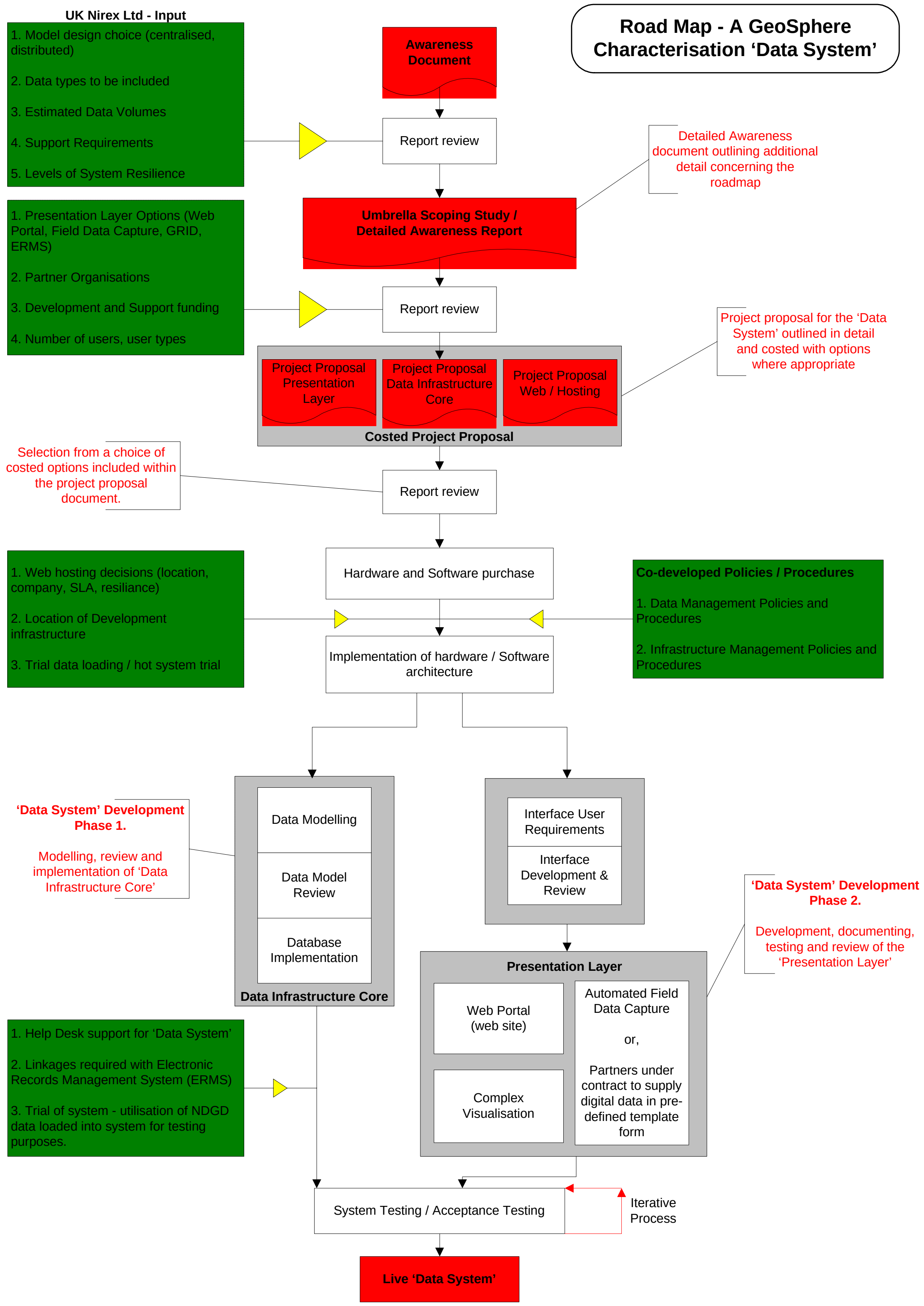
A viable alternative would be to pre-define templates or transfer formats aligned to the 'Data System' RDBMS data model for partner organisations and stipulate digital supply of data under any geoscience data capture contracts. The upload could be bulk update from site or continuous stream via technologies identified above.

Document Management for Characterisation Programme

The integrated geoscience database can, if requested, hold all relevant site geoscience data required by the project teams, ideally uploaded directly from the file locations. It would be a worthy addition to locate the necessary project and document data in the same location accessible through the secure web site or portal. Electronic records management systems are becoming wide spread within companies in order to best manage their documents. An ERMS linked to the information system will allow collaborative document production and crucially secure storage with necessary metadata to allow rapid retrieval and query if necessary.

10 Roadmap for future developments and decisions

The diagram below outlines the suggested path forward from this present ‘Awareness document’ towards an implementation of a ‘Data System’ to support a Geosphere Characterisation programme.


Road Map - A GeoSphere Characterisation ‘Data System’
UK Nirex Ltd - Input
1. Model design choice (centralised, distributed)
2. Data types to be included
3. Estimated Data Volumes
4. Support Requirements
5. Levels of System Resilience
Awareness Document
Report review
Detailed Awareness document outlining additional detail concerning the roadmap
1. Presentation Layer Options (Web Portal, Field Data Capture, GRID, ERMS)
2. Partner Organisations
3. Development and Support funding
4. Number of users, user types
Umbrella Scoping Study / Detailed Awareness Report
Report review
Project proposal for the ‘Data System’ outlined in detail and costed with options where appropriate
Project Proposal Presentation Layer
Project Proposal Data Infrastructure Core
Project Proposal Web / Hosting
Costed Project Proposal
Selection from a choice of costed options included within the project proposal document.
Report review
1. Web hosting decisions (location, company, SLA, resiliance)
2. Location of Development infrastructure
3. Trial data loading / hot system trial
Hardware and Software purchase
Implementation of hardware / Software architecture
Co-developed Policies / Procedures
1. Data Management Policies and Procedures
2. Infrastructure Management Policies and Procedures
‘Data System’ Development Phase 1.
Modelling, review and implementation of ‘Data Infrastructure Core’
Data Modelling
Data Model Review
Database Implementation
Data Infrastructure Core
Interface User Requirements
Interface Development & Review
‘Data System’ Development Phase 2.
Development, documenting, testing and review of the ‘Presentation Layer’
Presentation Layer
Web Portal (web site)
Automated Field Data Capture
or,
Partners under contract to supply digital data in pre-defined template form
Complex Visualisation
1. Help Desk support for ‘Data System’
2. Linkages required with Electronic Records Management System (ERMS)
3. Trial of system - utilisation of NDGD data loaded into system for testing purposes.
System Testing / Acceptance Testing
Iterative Process
Live ‘Data System’

Key:	**Green boxes**	- Issues / Questions for UK Nirex Ltd to address.
	Red Documents	- Documents produced during the workflow to a live system
	Grey boxes	- System development
	Red box	- Live ‘Data System’

The roadmap identifies key stages in the specification and development of a ‘Data System’. Documents are highlighted with a red background and input from UK Nirex Ltd is shown as green boxes, specifying the decisions or issues that needs to be discussed at that stage.

This initial high-level road map is subject to change in light of future decisions and input made by UK Nirex Ltd. The development stages of the roadmap (highlighted in grey) are particularly generalised at this ‘Awareness’ stage and will be outlined in greater detail during the ‘scoping study’ and ‘project proposal’ stages.

Explanation of Issues / Question for UK Nirex Ltd

Awareness Document and Review:

1. Consideration of model approach to be taken by ‘Data System’ – centralised or distributed.
2. Final data types to be stored by the ‘Data system’ after ‘Geosphere Characterisation’ programme are more fully defined.
3. Estimated data volumes from multiple data types to be stored by the ‘Data system’.
4. Levels of support required by UK Nirex Ltd and its geoscience partner organisations when utilising the ‘Data system’. How quickly does the system need to be repaired? What level of support cover is required by end-users?
5. What levels of resilience should the ‘Data System’ have built-in? Should failure of any part (or specific parts) of the system appear transparent to the end-user community? Is there a level of system downtime that can be tolerated in order to manage system development cost with respect to hot-spare and back-up systems?

Umbrella Scoping Study and Review:

1. What are the elements of a ‘presentation layer’ will be required by the end-user community and UK Nirex Ltd are prepared to pay for the development of?
2. What will the policy be with regard to the supply of digital geoscience data from partner organisations;
    - Use bespoke field systems supplied by programme, automatic data uploading
    - Supply data in approved transfer format for inclusion in ‘Data System’,
    - Supply digital data in proprietary formats for inclusion in ‘Data System’

3. Numbers of users, partner organisations and outline ‘user types’ for which the system would be expected to be able to cope (for example; Administrators, Senior Data Managers, Data Managers, Site Managers, Geoscience Staff, Developers, Public)
4. Funding levels available for development and support of the ‘Data System’ (to act as a guide in producing a number of predefined options for consideration at the ‘Costed Project Proposal’ stage).

Costed Project Proposal and Review:

1. Selection of preferred option from ‘Costed Project Proposal’ document that fits best UK Nirex Ltd’s vision of the ‘Data System’.
2. Timescale from ‘costed proposal’ to implementation in accordance with development and testing schedules to match general ‘Geosphere Characterisation’ programme plan.
3. Selection of ‘Development Organisation’ to implement the preferred option.
4. Location of development infrastructure.
5. Selection of hardware and software required and selection of a preferred infrastructure supplier.
6. Purchase of suitable hardware and software.
7. Selection of web hosting (ISP) organisation.
8. Service Level Agreement with preferred ISP to house ‘live’ and ‘Backup’ servers for the ‘Data System’.
9. Production of data management policies and procedures for end-user community.
10. Production of management policies and procedures for the infrastructure purchased (for live, backup and development systems).

System Development:

1. Review data model and collated interface development requirements.
2. Review Presentation Layer elements as production is carried out.
3. Suggest linkages that are required by UK Nirex Ltd to other systems, administration or Electronic Records Management Systems.
4. Testing / Trial of the developed ‘Data system’ (possibly using an existing NDGD dataset).
5. Will a ‘help desk’ solution be required to support the developed ‘Data System’?

11 Summary

The BGS would suggest from our experience implementing similar systems both internally and for commercial clients (UK and international), that the following would be the strongest initial 'information system' to implement for a Geosphere Characterisation Programme.

Suggested Components:

- 'Data System' (Data Infrastructure Core + Presentation Layer)
    i. Complex geoscience and logistical RDBMS
    ii. Web Portal
    iii. Complex Visualisation linkage
    iv. Robust and resilient architecture
- Live servers hosted by commercial Internet farm (99.5% plus availability)
- Enhanced security by the use of a Virtual Private Network

12 Recommendations

In addition to the proposed information system outlined in this report there are a number of additional recommendations.

1. This document is an awareness report to outline the decisions and technologies that would be necessary for such a 'data system' to be progressed. It should be a precursor to an in-depth scoping study report that would build upon the 'Geosphere Characterisation' programme decisions that will have been taken by that point in the current planning cycle.
2. UK Nirex Ltd to review 'Awareness Document' and study 'Data system' workflow. This information should be combined with other 'Geosphere Characterisation' programme study to address the questions and issues raised prior to commissioning the next stage 'scoping study' report.
3. A report is compiled recommending storage options and management of analogue and sample collections.

Glossary

Term	Definition
3G/ GPRS	Wireless technologies used by mobile phone companies to enable rapid data transfer to mobile phones, PDA's or laptops in remote locations that are still covered by a provider's cellular grid network.
Data Custodians	Data Custodians are corporately assigned staff whose role is the management of the data held within a dataset to the aims of the organisation or programme and not project specific.
Database Schema	A logical section of the database where tables are held that can be constrained to each other.
Discovery Metadata	Information about data that allows someone to discover its attributes, such as title, spatial extent, abstract, coverage, density, usage, format, usage restrictions etc.
Enterprise-level	"High level, scalable, robust and secure" when applied to software or hardware architectures statements.
ERMS – Electronic Records Management System	Software to implement an ERMS within an organisation
GRID	Grid computing is a form of distributed computing that involves coordinating and sharing computing, application, data, storage, or network resources across dynamic and geographically dispersed organizations. Grid technologies promise to change the way organizations tackle complex computational problems. However, the vision of large scale resource sharing is not yet a reality in many areas - Grid computing is an evolving area of computing, where standards and technology are still being developed to enable this new paradigm.
Horizon IT developments	Future IT developments that are already being discussed or planned by the industry.
ODBC – Object Database Connectivity	A technology utilised by many programmes to link to each other for the purpose of sharing data.
Portal	A web gateway application that allows visualisation of the data holdings along with download and management.
RDBMS	Relational database management system, the two most prominent examples are Oracle ® and Microsoft MS-SQL Server®.
Spatial Data Engine / Oracle spatial	A spatially enable component within the relational database system used to hold spatial data.
Technical Metadata	Technical metadata – metadata, (or 'data about data') on objects contained within the RDBMS such as tables, constraints or views.
Virtual Private Network (VPN)	Access technology to allow only authorised people access to a server or suite of servers. Link is encrypted and secure based upon roles and permissions.

References

The following are sources of reference used in the creation of the proposed information system model outlined in this report.

Web Sites:

DFID/KaR – R7199 Strategies for Maximising Geoscience Data Value. Link: http://www.bgs.ac.uk/dfid-kar-geoscience/r7199/home.html

PADAMOT - 'Palaeohydrogeological Data Analysis and Model Testing', database design and web site. Link: http://www.bgs.ac.uk/padamot/home.html