The PEG-BOARD Project: a case study for BRIDGE

Gregory Tourte[1]*; Emma Tonkin*[2]*; Paul Valdes*[1]

[1] School of Geographical Sciences.
University of Bristol,
Bristol, United Kingdom
{g.j.l.tourte,p.j.valdes}@bristol.ac.uk;
[2] UKOLN,
University of Bath,
Bath, United Kingdom
e.tonkin@ukoln.ac.uk

Abstract

With increasing public interest in the area of historical climate change and in models of climate change in general, comes a corresponding increase in the importance of maintaining open, accessible and usable research data repositories. In this paper, we introduce an e-Science data repository containing extensive research data from palæoclimatology. Initially designed to support internal collaboration and organise data, the sharing of research outputs became an increasingly significant role for the service over several years of practical use. We report on a data preservation and interoperability assessment currently under way. Finally, we discuss the ongoing significance of open research data and capacity for analysis in the area of climate research, with palæoclimatology as a case study.

Keywords: palæoclimate modelling; data management; data curation.

1. Introduction

The BRIDGE research group, or Bristol Research Initiative for the Dynamic Global Environment, focuses on the emerging area of 'Earth System Science' exploring the complex interactions between the Earth's components: the oceans; atmosphere; ice sheets; biosphere; and the influence of human activity on global change. This approach requires the input of multidisciplinary teams drawn from across Bristol University Glaciology, Hydrology, Biogeochemical Cycles, Chemistry, Earth Sciences, Mathematics, Engineering, Biological Sciences, Archæology, Personal Finance Research) and beyond (Hadley Centre, British Antarctic Survey, UK Met Office, DEFRA, Environment Agency, Centre for Global Atmospheric Modelling, Oil Industry).

Climate, 'the synthesis of atmospheric conditions characteristic of a particular place in the long term', is 'expressed by means of averages of the various elements of weather'; climatology, then, is the scientific study of climate [1]. The main research effort of the group is to improve the understanding of the causes of climate change, by testing the computer climate models used to predict future climate change. Major themes include:

- quantifying environmental and climate change in the distant past through the combined use of data and models;
- evaluating climate models with accurate proxy climate records, especially during periods of rapid climate change;
- improving climate models by incorporating additional components of the Earth System and detailed analysis of these processes for past, present and future change;
- assessing the impact of future climate change on spatial and temporal scales relevant to society and including timescales from decadal to millennial.

Many of these activities require—and produce—many terabytes of data. Making this data widely available is therefore a complicated and non-trivial process.

Researchers worldwide in both the sciences and humanities reuse BRIDGE data in their work. The project developed and applies de-facto preservation and data compression policies. Since the types of information required by users from areas as diverse as evolutionary biology, archæology and earth science very greatly, the project also developed an in-house interface designed to support tailored information extraction from climate model information.

Despite the complications associated with open access to large scientific datasets, openness in procedure and output is a priority for BRIDGE, and has been for many years. The importance of open data in climatology research in general has recently been highlighted, due to the high profile of the research area in the media and politics.

1.1 Background

Sweet [2] divides climate modelling into theory, empirical work, and modelling, and notes that modelling attracts the most attention since this area most directly assesses impact and produces predictions. It is expensive; simulations can take up to three months to run on high-performance computers ('supercomputer' clusters) and can equate to up to a hundred thousand pounds worth of computer time, excluding the cost of storage. The existing archive of resulting data sets consists of over 2,000 simulations and represents several million pounds worth of CPU time. The cost of CPU time has reduced; however, the scale of models has increased as a result. In terms of data requirements, a single model simulation can produce up to 2 TiB of raw model output data. A smaller subset of 2 to 50 GiB per simulation is retained.

Adopting Sweet's approach, we view the area as containing three areas of endeavour: empirical work, including data collection and preservation, theory, and modelling. In practice, these areas are difficult to divide; Edwards [3] qualifies the model/data relationship in climate science as 'exceptionally complex'. The boundaries between a global climate model (GCM) and data are 'fuzzy', and the interaction between model and theory is supple and ongoing. A model inspired by theory may apply initial conditions taken from measured data points. Data generated via a GCM may be compared with observed data points to evaluate the *validity* of the model. This demonstrates that model results agree with observations and that no detectable flaws exist, rather than that the GCM is essentially correct, but is nonetheless a significant step in establishing realism.

e-Science has a strong tradition in climate science. In data collection, for example, Benford et al. [4] describe the use of a Grid-based networked device to enable remote monitoring of Antarctic freshwater lakes and explore the potential for distributed collaborative research based on the resulting dataset. Benford et al. [4] highlight Anderson and Lee's [5] four phases of software fault tolerance as key to ensuring confidence in the resulting data: error detection, damage confinement and assessment, error recovery and fault treatment. Data, then, is only part of the story; provenance and context are required to ensure confidence.

Climate modelling software, too, is increasingly designed in order to make use of e-Science concepts and facilities. The SciDAC-supported Earth System Grid Center for Enabling Technologies (ESG-CET), for example, enabled all of the simulation data from the IPCC's AR4 to be made available to climate scientists worldwide [6]. The GENIE—Grid ENabled Integrated Earth modelling system—also applies a Grid-enabled architecture, in this case designed with the intent to 'build simplified and faster-running models of the Earth's climate system, and make them easier to use and more widely available to other people who want & need to use them' [7] . GENIE is designed to facilitate cyclic improvement of models through comparison with available datasets; to improve traceable integration between various model types, and to integrate multiple representations of the natural Earth system. GENIE enables large ensemble studies on the Grid, supports complex workflows and provides Grid-based data handling and post-processing facilities [8]. In each of these applications, as is generally true with Grid-based approaches [9], rich and descriptive metadata, including extensive information about data provenance, is required to enable effective use of available data.

The political significance of climate modelling as a research area is currently such that openness is absolutely key. With publicly funded research, the 'citizen scientist' should be considered as a stakeholder, and ultimately this is dependent on working with the user community [10].

1.2 The case for open access to data

The importance of open data in climatology research in general has been highlighted in recent years, due to the high profile of the research area in the media and politics. Climate modelling, particularly in the area of climate prediction, is subject to a high level of scrutiny.

Consider for example a recent news article [11], discussing the open review of a recent report, the 4th Assessment Report or AR4, published by the Intergovernmental Panel on Climate Change (IPCC). The process described is a review conducted by 'climate "sceptics", [...] busy searching the rest of the panel's report for more mistakes'. One statement queried is described as 'basically correct but poorly written, and bizarrely referenced'; the process of establishing accuracy has highlighted issues regarding appropriate referencing and clarification of the distinction between 'grey', or non-peer-reviewed, literature, and peer-reviewed sources. Harrabin suggests 'a need for much greater transparency'. A further famous example are the international repercussions (both political and scientific) surrounding the recent 'leak' of emails from the Climatic Research Unit at the University of East Anglia, dubbed 'Climategate' by many.

Access to data and modelling resources is variable. For example, the UM (Unified Model), the popular suite of atmospheric and oceanic numerical modelling software developed and used at the UK's Met. Office has limited availability, being primarily available to UK academic researchers. Availability of the GENIE software is currently limited, as the software remains work-in-progress. A great deal of data is available, from sensor data released by the

British Antarctic Survey, the Australian Antarctic Division and others to the OpenGeoscience service offered for non-commercial use by the British Geological Survey; a great deal of open-access data may be discovered via the NERC Data Services initiative (http://ndg.nerc.ac.uk/) that gathers together the NERC data centres. Data centres typically hold collections of empirical data (e.g. observations and measurements).

Open procedure and open access are priorities for BRIDGE, and a software platform has been developed over many years to support this aim, allowing modellers to publish datasets along with relevant experimental metadata. Although the present iteration of the software predates recent best practice in the area, the service has been widely used for those requiring secondary data, to the mutual benefit of BRIDGE and external users of the data.

1.3 The PEG-BOARD project: Palæoclimate & Environment data Generation – Building Open Access to Research Data

In response to the community's need for openly accessible research data, we need to make sure that the data generated as part of our research remains accessible and preserved for a certain amount of time after its creation and original use.

However, preservation of digital information is a very complex subject. Su-Shing Chen in the Paradox of Digital Information [12] explains why it is difficult to come up with a simple definition of what 'to preserve digital information' means. He says that 'on the one hand we want to maintain digital information intact as it was created' (one facet of preservation) 'on the other, we want to access this information dynamically and with the most advanced tools' (preserving access to the data).

This is extremely relevant to our data as the models used to generate it as well as the hardware architecture on which the models are run evolve and change over time. A particular experiment run five years ago may not run on current hardware or if it runs, may not produce the same results. We have seen recently that the implications of publications and data may be seen and questioned decades later. However, from a more pragmatic point of view, the benefit of keeping old data can easily be questioned. The cost of storing large dataset is very high, despite the raw cost of storage going down dramatically with time, archival enterprise grade storage is still very expensive and the long-term maintenance cost of keeping a storage system working and up-to-date may well be higher than the cost of re-running the experiment, especially when computers speed is also increasing with time. Another point to consider is the fact that the science included in the models evolves as well.

With computers becoming more and more powerful, the complexity of the models have increased, adding CO_2, NO_2 and H_2O exchanges to atmospheric models as well as vegetation over the last 15 years [13]. This means that old experiments will be inaccurate compared to our current understanding of the earth system and therefore may as well be re-simulated to get a result more in line with the current science.

With that in mind, the PEG-BOARD project has several aims, targeting every aspect of our data and our user base :

- assist the work of modellers by facilitating data processing, manipulation and analysis by the modellers and scientists who generate data as part of their research;
- facilitate data reuse by modellers and by any consumers of the data by providing methods to search and browse through the data;
- discover and characterise modes and means of data reuse, and identify relevant user groups;
- identify current patterns of metadata use, the standards used and the extent to which they comply with relevant data types;
- describe current data retention policies and relevant standards;
- provide clear guidelines to research groups and researchers to help manage their data;
- ensuring proper data retention and curation policies based on both the research and the data life cycle;
- disseminate documents and software to wider community to provide better understanding and better accountability for the research communities to the wider public and stakeholders.

We are now in the requirements-analysis stage of a new project, PEG-BOARD, designed to support the curation of historical climate data within BRIDGE's large global consortium of palæoclimate researchers, and to ensure ongoing availability of this data for reuse within research, teaching and the media. This work is carried out in the context of the UK e-Science infrastructure [14]. The project focuses on providing the community with a better understanding of the data and the limits of its validity, and defining a clear policy structure for palæoclimate data. An improved data management infrastructure is expected to improve availability and accessibility of data, as well as providing a stabler structure for collaborative reuse. Open availability of well-structured and documented research data is key, enabling open and easy creation of malleable prototypes, adaptable to relevant research or interest communities.

2. Methodology

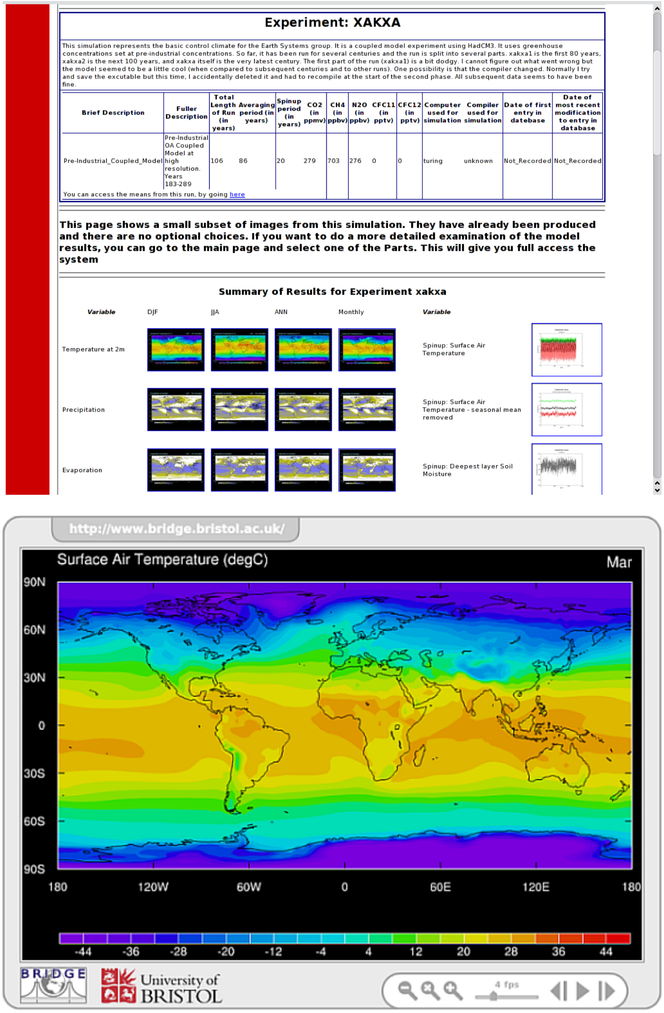
Due to the strong user-analysis component of these aims, we chose to begin with a phase of user analysis of the present system. Various mechanisms exist for exploring user requirements; indeed, the field of requirements engineering has over time attracted a large and very active research community. Requirements engineering is described by Laplante [15] as 'a subdiscipline of systems engineering and software engineering that is concerned with determining the goals, functions, and constraints of hardware and software systems'. Nuseibeh & Easterbrook [16] describe requirements engineering as follows:

'The primary measure of success of a software system is the degree to which it meets the purpose for which it was intended. Broadly speaking, software systems requirements engineering (RE) is the process of discovering that purpose, by identifying stakeholders and their needs, and documenting these in a form that is amenable to analysis, communication, and subsequent implementation.'

RE is not a single operation but a sequence of operations. Stakeholder analysis is a necessary precursor, a part of the process that in our case has been explored for previous developments in the BRIDGE area, but which due to the nature of the problem area is necessarily an ever-shifting target. Nuseibeh & Easterbrook describe the core areas of RE as: *eliciting* requirements, *modelling and analysing* requirements, *communicating* requirements, *agreeing* requirements and *evolving* requirements. The mechanisms used in the PEG-BOARD project thus far can be fitted into this overall model of the process of requirements engineering, although some aspects were explored prior to the beginning of the project (stakeholder identification in particular).

The processes of eliciting, modelling and communicating requirements are all touched on in this paper. Requirements are elicited initially by the exploration of existing systems in use as part of the task decomposition process – via interface surveys (see Section 2.2), and then via the use of structured interviews with selected users. This is completed in two areas; with users internal to the BRIDGE project, and with a case-study of an external consumption of BRIDGE data. 'Data Sharing Verbs' are used as part of the modelling and communication of requirements.

We chose to begin with a series of interviews, exploring a number of 'characteristic' individual users' perceptions of their interactions with the BRIDGE services. The results of this process form part of the background material for the Results section of this paper (Section 3).



Experiment: XAKXA

This simulation represents the basic control climate for the Earth Systems group. It is a coupled model experiment using HadCM3. It uses greenhouse concentrations set at pre-industrial concentrations. So far, it has been run for several centuries and the run is split into several parts. xakxa1 is the first 80 years, xakxa2 is the next 100 years, and xakxa itself is the very latest century. The first part of the run (xakxa1) is a bit dodgy. I cannot figure out what went wrong but the model seemed to be a little cool (when compared to subsequent centuries and to other runs). One possibility is that the compiler changed. Normally I try and save the excutable but this time, I accidentally deleted it and had to recompile at the start of the second phase. All subsequent data seems to have been fine.

Brief Description	Fuller Description	Total Length of Run (in years)	Averaging period (in years)	Spinup period (in years)	CO2 (in ppmv)	CH4 (in ppbv)	N2O (in ppbv)	CFC11 (in pptv)	CFC12 (in pptv)	Computer used for simulation	Compiler used for simulation	Date of first entry in datebase	Date of most recent modification to entry in database
Pre-Industrial_Coupled_Model	Pre-Industrial OA Coupled Model at high resolution. Years 183-289	106	86	20	279	703	276	0	0	turing	unknown	Not_Recorded	Not_Recorded

You can access the means from this run, by going here

This page shows a small subset of images from this simulation. They have already been produced and there are no optional choices. If you want to do a more detailed examination of the model results, you can go to the main page and select one of the Parts. This will give you full access the system

Summary of Results for Experiment xakxa

Variable	DJF	JJA	ANN	Monthly	***Variable***	
Temperature at 2m					Spinup: Surface Air Temperature	
Precipitation					Spinup: Surface Air Temperature - seasonal mean removed	
Evaporation					Spinup: Deepest layer Soil Moisture	

Figure 1: The BRIDGE Data Access Portal

2.1 Exploring existing software development

We continued by exploring the current software system put in place to manage palæoclimate research data as this system has been and continues to be extremely dynamic, in order to follow the science involved and the needs of the scientists who use it. This is therefore an extremely valuable source of information on user requirements, technological requirements and preliminary insight into the de-facto research and data lifecycle evolution.

However, the system is currently very much designed to simplify the work of the climate modeller in that the interface really helps a scientist to work on his/her own experiments: the metadata describing the experiments usually references other experiments on the system which were used to create it, as well as parameters used in the first place by the central UM interface. Within each experiment, the variables shown on the web interface are taken verbatim from annotations stored within the file itself, each of which follow the CF metadata standard.

There is currently no requirement for the modellers to describe their experiments in a way an external, non-modeller, user could understand, or for that matter a way a computer could interpret. The use of CF metadata is a very good start but it is embedded within the file and only describes that specific file in which it is embedded with no references to the experiment to which it belongs. There is therefore a need to work on an experiment-level metadata schema that would describe the experiment as a whole and enable proper indexing on values that all users of the system could understand and not only the original modeller who created the data.

We have started looking at several metadata formats, such as the DIF (Directory Interchange Format) schema created by NASA [17] and the currently on-going work on the Scientific Data Application Profile [18].

2.2 Describing the Research Lifecycle

The process of creating, disseminating, storing and reusing research data is part of the overall research lifecycle. In order to come to an understanding of how this works, therefore, it is useful to characterise the research lifecycle that underlies it. There are considerable potential benefits to this process; if the process as it is today is well understood then it becomes possible to support the process as it stands, and potentially to find social, process-oriented and technical means to improve the speed, ease, and cost-effectiveness of that process further.

There are a number of models, mechanisms and proposed methods designed to support this process, a few of which we will briefly discuss here. Swann, for example, designed a model that was used for some time by the UKRDS (UK Research Data Service). This focused on separation of individuals involved in the research lifecycle into a set of possible types, notably data creator, user and viewer [19]. This was useful as a method of decomposition, but focused on categorizing people into one of a number of types. It was later suggested that individuals might more usefully be seen as involved in a number of different activities, and hence a later model focussed on individuals' roles at given times within a give research workflow.

'Data Sharing Verbs' represent one such model, a mechanism described by the ANDS as a 'structuring device', to support discussion about the technology and process of the data sharing aspect of the research lifecycle. The key insight underlying this is the assertion that thinking about the 'what' rather than the technical details of the system is useful — that user experience can be described through a description of what is being done from the user perspective. This mechanism is described by Burton & Treloar as 'Data Sharing Verbs' [20]; the candidate terms offered include Create, Store, Identify, Describe, Register, Discover, Access and Exploit, although additional verbs are likely to be required for specific use cases and as time passes.

This approach can be effectively compared to relatively traditional methods drawn from human-computer interaction and design methodologies, such as task analysis and decomposition. According to Kieras [21] task analysis is the process of understanding the user's task thoroughly enough to help design a system that will effectively support that user in doing the task. Task analysis aims to systematically analyse a task based on the knowledge and goals of the user, system, information and functionality (that is, social, organisational, technical factors). The 'Data Sharing Verb' idea could be described as a user-focused subset of this overall set of aims, specifically characterising an accessible researcher-level viewpoint on that overall area of endeavour. The fundamental aim of Data Sharing Verbs is as a structuring device, high-level architectural approach and descriptive mechanism [20]; they are described as 'one way of thinking about the things that need to take place', and it is noted that they 'encourage a focus on the functionality [and] result'. They can therefore be seen as an approach to collaborative representation and design. However, little information is provided regarding the mechanisms by which they are assigned to a novel usage context, so that is an area of interest for our ongoing work.

The work reported here was achieved using methods derived primarily from classical task analysis, with modifications designed to take in the useful idea of accessible data sharing verbs. There are many formalized methods in existence for the purposes of requirements gathering and task analysis in particular, but these do not in general provide a novel mechanism of analysing or understanding a task. In fact, much like the Data Sharing Verbs representation described above, most formal methods are ways to represent the results of a task analysis [21].

According to Kieras [21] the process of task analysis itself is usually based around some or all of the following methods:

- observation of user behaviour – a thorough, systematic and documented overview of observations with the aim of understanding the user's task. This may use a think-aloud protocol (ie. the user is invited to vocalise his/her observations about a task while working through it).

- Review of *critical incidents* and *major episodes* – rather than discussing the full span of user experience, a subset of particularly informative case studies are discussed.
- Questionnaires: these often suffer from difficulties with accuracy limitations, but are economical to use and can collect some types of user and task data.
- Structured interviews: talking to users or domain experts about a task is a good way of gaining some idea of the basics, and a more structured interview series at a later time can be an effective means of systematically exploring the area.
- Interface surveys: exploring existing interfaces, scripts, and so on, can provide useful information about interface characteristics, explanations, interface issues as perceived and annotated by users, and so forth.

Due to the inevitable time constraints of a relatively short-term project we chose to limit the use of observational/ethnographic methods to the latter phases of exploration of our system. Instead we looked towards the use of, initially, unstructured interviews, supplemented by an intensive interface survey series of the various visual and script-oriented interfaces that have been developed to serve the day-to-day needs of BRIDGE users of various types over the fifteen years of its operation. We then used this information to build a questionnaire, the results of which will be used to develop our initial findings as presented here into a second iteration.

We do, however, feel that ethnographic methods and think/talk-aloud workthroughs are likely to be of importance, particularly when exploring the cost-value propositions underlying our interface and those of other data providers/data centres in which the data is deposited. For example, it is often the case that users perceive deposit processes in particular as excessively lengthy and something of a waste of time, and in some cases there are very different ways to present that task to alter the value proposition as presented to the user.

3. Results

We begin by describing what has been elicited so far regarding data generation, storage, administrative and descriptive metadata, and reuse. We then present a candidate research data lifecycle model. Because the findings demonstrated emphatically that data consumption and reuse was a very significant part of the lifecycle, and indeed proved to represent the proximate cause of a great deal of the effort historically applied to this data collection, we found the need to place a far greater emphasis on it than was originally predicted.

BRIDGE data is generated via global climate models simulations (GCM), run on several national and international high performance computing (HCP) facilities. Our main tool is the Met Office Unified Model (UM) which runs a number of standard models such the Hadley Centre HADCM3 and HADGEM, or more recently FAMOUS, but we also use the European oceanography model NEMO or GENIE (Grid ENabled Integrated Earth). The majority of our output comes from the UM which uses a proprietary output file format. However the industry standard for such large data sets is NetCDF.

NetCDF, currently maintained by University Corporation for Atmospheric Research (UCAR), is a widely used open standard. It is an extremely flexible format optimised to store large multidimensional arrays of numerical data such as those describing high resolution planet-wide data.

When the data is created, it is moved and converted to NetCDF to a storage and processing farm of server where the data is processed. Climatology involves running weather simulation and then averaging the output to obtain the climate information. There is a number of default processes that are always running on the data to produce defaults sets of plots. It is then up to the modeller to add the specific output required for a specific project.

Due to the large amount of data created (around 2TB per day of raw output), it is not possible to store and keep everything, so raw output (from the UM) is discarded after conversion to NetCDF and calculation of intermediary averages generated from the converted files. Only the directly converted NetCDF files and the final averages and plots are kept. No expiration date is currently mandated for the data.

3.1 BRIDGE Service Design

The BRIDGE project at present has over 100 research groups spread over approximately 10 countries—see stakeholder analysis, figure 2 and figure 3. The multidisciplinary reach of palæoclimatology data presents some unique challenges in data dissemination. Historically, this diversity in user communities has meant that direct interaction with expert users of the BRIDGE environment is a necessary component in enabling access to, and reuse of, research data. However, as the number and diversity of background of stakeholders has continued to widen, these manual processes have become increasingly unfeasible. Enabling computer-supported scientific collaboration is at the intersection of Computer Supported Cooperative Work (CSCW) and e-Science [22], and the specific problem of data curation is a recent addition to the area.

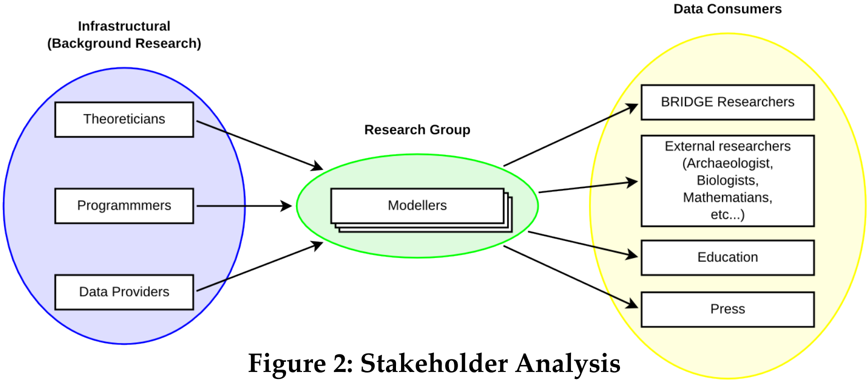


Figure 2: Stakeholder Analysis

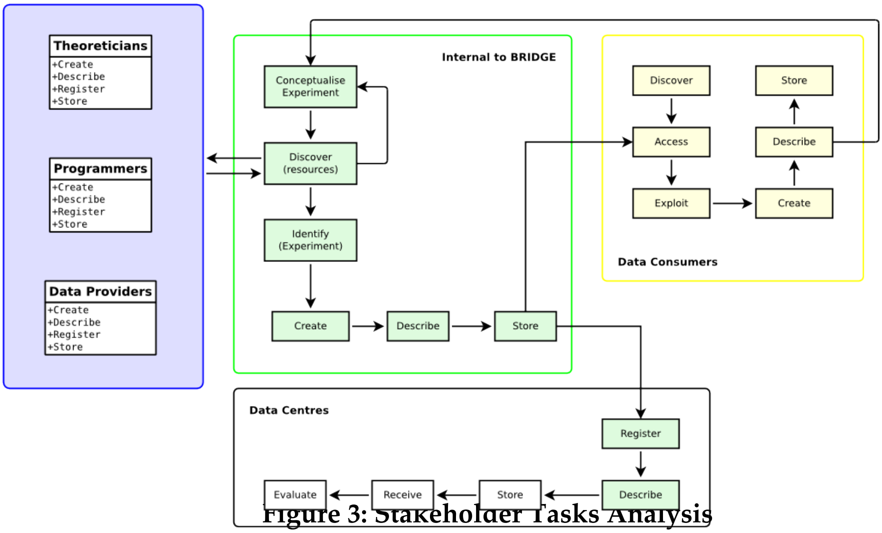


Figure 3: Stakeholder Tasks Analysis

The first challenge for those working in interdisciplinary research is to locate relevant data repositories and databases [23]. The second is to get 'up to speed' with the nature of the data and with its practical uses, metadata and it's provenance.

3.2 BRIDGE Systems Architecture

The current BRIDGE infrastructure only supports UM data which constitutes 99% of the data utilised. Compatibility with non-UM data is under consideration.

The current BRIDGE facilities provide services for the groups of stakeholders described here as the research group and the data consumers. Data providers are accessed by the modellers independently as the sources providing boundary conditions are rarely computer readable and usually come in the form of results published in scientific papers. These have to be 'translated' by the modeller before being added to the models.

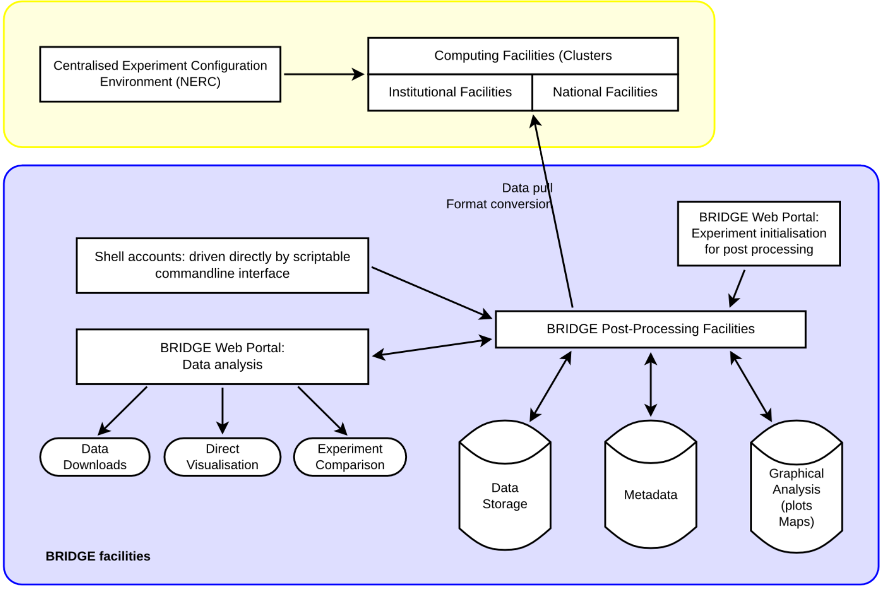

Centralised Experiment Configuration Environment (NERC)
Computing Facilities (Clusters
Institutional Facilities
National Facilities
Data pull
Format conversion
BRIDGE Web Portal: Experiment initialisation for post processing
Shell accounts: driven directly by scriptable commandline interface
BRIDGE Post-Processing Facilities
BRIDGE Web Portal: Data analysis
Data Downloads
Direct Visualisation
Experiment Comparison
Data Storage
Metadata
Graphical Analysis (plots Maps)
BRIDGE facilities

Figure 4: Architecture Diagram of the BRIDGE facilities

In figure 4, we show the overall architecture of the services provided by the BRIDGE portal. Experiments are configured on a centralised national facility provided by NCAS and run on national and institutional HPC facilities. In parallel, BRIDGE modellers need to initialise their experiments on the

BRIDGE facility by inputting details and metadata of the experiments. Once this is done, the modellers just have to let things run; the system is fully automated (unless the experiment fails in some way).

In order to avoid straining the limited storage capacity of HPC facilities, generated data is pulled regularly by the post-processing servers which then check and convert it to NetCDF from the proprietary UM format. The original UM files are then discarded. This process runs during the entire time taken by the experiment to complete, which can be up to several months. On completion, the modeller is given the choice to apply predefined averaging algorithms to the data, or define his own, in order to create an initial set of plots, maps and animations for a preliminary analysis of the data. The predefined algorithms are updated regularly by the research group to suit its evolving needs.

Once the experiment is processed, most of the data is archived and only the post processed data, enough to generate most graphics is kept available and shared on the portal. From this point, the experiment is available to 'data consumers'. The portal allows users to to view pre-generated graphics as well as creating new ones from the data, either to change the output format, to use different variables, or to combine multiple variables. The option is also given to compare the results of two experiments. This is made possible by the fact that extensive work has been done to make all graphics of the same type use the same colour scaling (visual conventions). All generated graphics outside the default predefined ones are cached for a limited time period so will not have to be re-generated at every access (a very time consuming and resource intensive process). Users also have access to post-processed data, either in NetCDF or converted into other formats such as CSV.

3.3 Case study: BRIDGE in Archæology

Archæology researchers at the University of Southampton make use of the BRIDGE software as part of their research. In this case study, their interest is in data regarding the climate in which a group of early Neanderthals lived. The specific information that can be provided as a result of BRIDGE palæoclimate simulations includes wind speed, temperature, and rainfall. Palæoenvironmental information can help archæologists understand likely patterns of migration as well as providing contextual information surrounding artefacts, etc. In particular, palæoclimatology may be key to our understanding of the extinction of the Neanderthals [24].

Originally, very few NetCDF viewing applications existed for non-UNIX environments. Therefore, the use of BRIDGE resources was required. Even now, the level of computer literacy required to analyse NetCDF data is very high. Our data uses meteorological units (temperature in Kelvin, precipitation in kg/m^2/s and wind in m/s) whereas what is usually required is the more every day units (temperature in Celsius, precipitation in mm/day and wind in mph or kph.) Doing a single numerical unit conversion may not be a complex process, however, the overall process of extracting thousands of values from a number of files and then performing type-appropriate batch conversion is relatively challenging and time consuming. It was therefore decided to add data conversion and merging services to the BRIDGE service.

Another issue regarding the interdisciplinary use of climatology raw datasets is the terminology used to describe the data variables contained in each files. This is even an issue for a glaciologist trying to use palæoclimate datasets. The netcdf files are all CF compliant (Climate and Forecast Metadata convention, as required for data generated as part of NERC funded projects) which includes over 30 variables describing some sort of air temperature— eg. *air_temperature, air_temperature_at_cloud_top, surface_temperature, surface_ temperature_where_land, surface_temperature_where_open_sea, …*— as well as over 15 names describing types of air pressure (*air_pressure, air_pressure_at_ cloud_base, air_pressure_at_cloud_top, air_pressure_at_convective_cloud_base, air_pressure_at_sea_level, …*). This multiplicity of terms which for some disciplines would be described as air temperature and air pressure makes exchange and reuse of data particularly difficult without very close collaboration with a scientist. An individual acting as a 'gateway' between disciplines would ordinarily be from the same field as the original data creator but who also understands the requirements of the scientist who is trying to use the data.

Issues brought up during this work included the difficulty of discovering appropriate datasets—finding experiments that contained relevant data. This was solved by requesting that appropriate experiments were recommended by BRIDGE team members. This, coupled with the need to automate common tasks, meant that the collaboration had a significant cost in terms of time. Hence, changes made to the service at the time included a concept of 'typed' data—for example, precipitation—to which a number of standard conversions may be applied. The need for appropriate metadata is also very clear, but with a legacy of over a decade of datasets (over 2000 simulations), the problem of introducing an improved standard includes the need to deal with a large amount of legacy content. Metadata applied to the data should also enable the cross-disciplinary browsing, discovery and use of the data, by the use of some sort of description table or translation table to either provide this translation automatically or provide the user with a plain english description of the term to allow him or her to choose the right one.

4. Discussion

The task analysis/preservation hybrid approach, making use of the 'data sharing verbs' to support discussion, has fitted well into our environment. Furthermore, it offers a strong theoretical basis in both preservation and HCI.

So far, we have successfully completed an investigation into the research lifecycle of research data from the BRIDGE project. We have built up an understanding of the existing software and hardware infrastructure that has been built up to support this lifecycle, and explored the rules associated with data creation and reuse, both external and internal in nature. We have also explored a case study of the reuse of palæoclimate data, in which archæology researchers at the University of Southampton make use of the BRIDGE software to access relevant datasets for the purpose of exploring patterns of migration. From this case study, we note a need for clear and consistent metadata, as well as for metadata to be applied to existing and older datasets – and we note that such collaborations often have a significant cost in terms of time, which can be reduced by enabling the development of software that supports ongoing collaboration by accessing consistent and well-defined data-access services or APIs.

The next stage for us is to ground our existing work with further detailed analysis of:

- the path(s) to completion of common tasks; for example, the time taken to complete a task, technical and knowledge-organisational issues and dependencies.
- technical infrastructure/system
- related infrastructural dependencies, such as the requirement to deposit information in data centres
- patterns of reuse of the data; impact, review and overall benefit to the community
- the costs and benefits of each aspect of the system.

4.1. Updating BRIDGE

Initially, we chose to focus on data management requirements analysis, exploring requirements for named stakeholders. Following the work described here, we have greatly improved our understanding of the broad technical and social processes that take part around the BRIDGE data. Now, however, we will need to identify appropriate methodologies for developing an improved understanding of the practical implications of the system as it is described here. For example, the time taken to complete any given process is very relevant to the question of the total cost of that process. For example, the time taken to develop an archival copy of a dataset (depending on the definition of the term 'archival'; this necessarily depends on the choice of archiving method, so that the costs of putting data into a data repository and that of storing it locally are very different) may be measured.

We will also continue to explore the practical issues and opportunities surrounding the reuse of BRIDGE data both in local formats and in the data-centres' preferred representations and formats.

4.2 Requirements analysis: Preservation, accessibility and metadata extraction

We intend to continue by consolidating our work with further questionnaires, observational studies and interviews. Tor this purpose, we have identified relevant components of the JISC Data Audit Framework (DAF) [25], DRAMBORA [26], the Planets project-preservation planning workflow [27], and similar tools to help identify and develop a formal data management strategy for palæoclimate model data, taking into account the requirement that consistency with the NERC Data Grid is a critical factor.

This should enable us to analyse the workflow described above in more details. In particular, we are looking to gain further information about users' (data creators and consumers alike) viewpoints and experiences with the data, its administration, access issues and potential enhancements. To this end, we have developed a questionnaire adapted from the Data Audit Framework, which is generally expected to be used primarily within an organisational context. The use of the DAF to explore data reuse externally constitutes a change from the usual way in which the framework is used, so it will also be an opportunity to explore and evaluate this approach.

In the following phase we expect to apply these components to the problem area described in this paper.

4.3 Requirements analysis: Automated metadata extraction.

We expect to explore the use of relevant metadata standards—PREMIS [28], STFC, etc.—to enhance the structures currently in use, as well as exploring the use of metadata extraction in order to supplement the file-format specific metadata currently used as the primary data management tool. The scale of data generated in palæoclimate research means that, wherever possible, metadata will need to be automatically generated.

Automated metadata extraction is the process of mechanically extracting metadata from a source document [29]. A completely automated process is unlikely to give perfect results; however, augmenting a manual metadata extraction process with an automated mechanism, even one that has an error rate of perhaps 10% to 20% of cases can nonetheless increase the speed and, potentially, the consistency of a metadata generation process. It can also increase user satisfaction with the interface; that the system has tried to support the user, even if it has not totally succeeded, can lead to a less frustrating user experience than a totally manual system.

In this instance, automated metadata extraction may explore the datasets and their associated files and format metadata as sources. Additionally, one may use the paper-based outputs of the research process as a source of information about the simulations that took place. One particularly relevant point to this process is the problem of data citation; what should a data citation look like, and what does it resemble at present? Informal exploration of the problem area has suggested that a co-author relationship is often used as an alternative to dataset citation, acknowledging the contributor of the research data in an implicit manner.

4.4 Supporting user reuse of data: Accessibility and visualisation

Exploration of the extent and diversity of the user community surrounding the BRIDGE dataset has demonstrated that data reuse is widespread and diverse. Much of this is data reuse is formally uncharted, which is to say that although it appears in individual researchers' records, often as citations, it is not always acknowledged as such. The nature of the data makes many different representations possible; as geographical data it can be directly explored using software such as NASA's WorldWind [30]. However, radically different representations may be appropriate for different user groups – so the collection of end user requirements is key to scoping out relevant activities such as developing appropriate recommendations for APIs, services or policies relating to preferred data storage formats.

A few specific cases that we expect to explore in the near future include the requirements for development of clear, high-quality visualisations, suited for high definition broadcasting in the media, and the requirements for the development of simulations that support haptic rendering, which is to say, that augment visual representations with tactile feedback. Provision of an application programming interface that can support this work is expected to facilitate this sort of development in future, as it should reduce the cost, complexity and learning curve associated with making use of the dataset. Because ongoing reuse of the data is an important part of this research data lifecycle, making it as easy as possible for developers to work with the information is likely to be an effective way of increasing the impact of research data publication in the area.

4.5 Expectations

The key assertion underlying this project states that adoption of appropriate data management strategies, appropriate to partner institutions across the various research disciplines involved, will have several benefits. The most visible initially is expected to be improved accessibility for potential users of the dataset. We additionally assert that the sustainability of a research data curation programme is dependent on the existence of data management strategies with a robust approach to appraisal. Finally, a strong data management strategy should improve traceability, reducing the difficulty of answering questions such as data origin and confidence levels.

5. Conclusion

BRIDGE software is already being used to support a wide range of reuse patterns, including those described above. From exploring practical usage patterns, we have developed a number of updated requirements. The first is the need to provide high-quality metadata, enabling us to develop means for searching or browsing—exploring—the data, in an appropriate manner for specific end-user groups, be they archæologists, statisticians or biologists. It requires a viewpoint on metadata that is not excessively prescriptive or restrictive in terms of form or interface, but that enables the base dataset to be presented to many different user groups, in their own terms. The current BRIDGE software review will take into account these disparate user requirements in designing a flexible architecture that can support the generation of a wide variety of data representations.

Secondly, preservation is a key issue, along with provenance and the ability to precisely cite a given data set. Climate science is not a subject in which the 'fire and forget' philosophy can be adopted. However, it is also an area of e-Science that generates very large quantities of data. Data curation and preservation in this area is reliant upon the development of appropriate data retention policies; as part of the PEG-BOARD project we will explore data management requirements and develop appropriate policies along with any infrastructural dependencies.

Finally, better-quality visualisations and tools able to support accessible exploration of data are very important enablers for data reuse and widening the impact of completed research. This is a rich and open field for further research and development, particularly but not exclusively for educational purposes; high-quality visualisations are also sought after in many other fields, including audiovisual broadcast.

Notes and References

[1] LINACRE, E. *Climate data and resources : a reference and guide*. Routledge, 1992.

[2] SWEET, B. Three Cultures of Climate Science. *IEEE SPECTRUM*, 2010.

[3] EDWARDS, PN. *Global Climate Science, Uncertainty And Politics: Data-Laden Models, Model-Filtered Data. Science as Culture,* 8:437–472, 1999.

[4] BENFORD, S. et al. *e-Science from the Antarctic to the GRID*. In 2nd UK e-Science All Hands Meeting, 2003.

[5] LEE, PA; ANDERSON, T. *Fault Tolerance: Principles and Practice (Second Revised Edition).* Springer-Verlag, 1990.

[6] ANANTHAKRISHNAN, R. et al. *Building a global federation system for climate change research: the earth system grid center for enabling technologies (ESG-CET)*. Journal of Physics: Conference Series, 78, 2007.

[7] *Geniefy: Creating a grid enabled integrated earth system modelling framework for the community.* http://www.genie.ac.uk/GENIEfy/, 2009.

[8] LENTON, TM; et al. *Using GENIE to study a tipping point in the climate system.* Phil. Trans. R. Soc. A, 367(1890):871–884, 2009.

[9] SIMMHAN, YL; B. PLALE, B; GANNON, D. *A survey of data provenance in e-Science.* SIGMOD Rec., 34(3):31–36, 2005.

[10] DE ROURE, D. *e-Science and the Web. IEEE Computer*, August 2009.

[11] HARRABIN, R. *Harrabin's notes: IPCC under scrutiny.* http://news.bbc.co.uk/1/hi/sci/tech/8488395.stm, 2010. Retrieved Jan 30, 2010.

[12] SU-SHING, C. *The Paradox of Digital Preservation,* IEEE Computer, 34, p. 24–28, 2001.

[13] MCGUFFIE, K; HENDERSON-SELLERS, A. *A Climate Modelling Primer, Third Edition*, p. 47–79, 2005

[14] LYON, L. *Dealing with data: Roles, rights, responsibilities and relationships.* Technical report, UKOLN, Bath, UK, June 2007.

[15] LAPLANTE, P. *Requirements Engineering for Software and Systems (1st ed.),* CRC Press, 2009.

[16] NUSEIBEH, B; EATERBROOK, S. *Requirements engineering: a roadmap,* Proceedings of the Conference on The Future of Software Engineering, p.35-46, June 04-11, 2000, Limerick, Ireland

[17] OLSEN, L. *A Short History of the Directory Interchange Format (DIF),* http://gcmd.nasa.gov/User/difguide/whatisadif.html.

[18] BALL, A. *Scientific Data Application Profile Scoping Study Report,* http://www.ukoln.ac.uk/projects/sdapss/, June 2009

[19] SWAN, A; BROWN, S. *The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs,* 2008.

[20] BURTON, A; TRELOAR, A. *Designing for Discovery and Re-Use: the ? ANDS Data Sharing Verbs? Approach to Service Decomposition.* International Journal of Digital Curation, 4(3), 2009.

[21] KIERAS, D. *Task Analysis and the Design of Functionality*, In CRC Handbook of Computer Science and Engineering, p. 1401–1423, CRC Press, 1996.

[22] KARASTI, H; BAKER, KS; HALKOLA, E. *Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER).* Network. Comput. Supported Coop. Work, *15(4):321–358, 2006.*

[23] LUCE, R. *Learning from e-databases in an e-data world.* EDUCAUSE Review, 43(1):12–13, 2008.

[24] VAN ANDEL, TH; DAVIES, W; edts. *Neanderthals and modern humans in the European landscape during the last glaciation.* Cambridge: McDonald Institute for Archæological Research, 2004.

[25] JONES, S; BALL, A; EKMEKCIOGLU, Ç. *The Data Audit Framework: A First Step in the Data Management Challenge.* International Journal of Digital Curation, 3(2), 2008.

[26] MCHUGH, A; ROSS, S; RUUSALEEP, R; HOFMAN, H. *The Digital Repository Audit Method Based on Risk Assessment (DRAMBORA).* HATII, 2007.

[27] FARQUHAR, A; HOCKX-YU, H. *Planets: Integrated Services for Digital Preservation. International Journal of Digital Curation, 2(2), 2008.*

[28] *PREservation Metadata: Implementation Strategies Working Group.* PREMIS Data Dictionary, 2005.

[29] TONKIN, E; MULLER, H. *Semi automated metadata extraction for preprints archives,* JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries, p. 157–166, 2008.

[30] NASA World Wind Java SDK, http://worldwind.arc.nasa.gov/java/