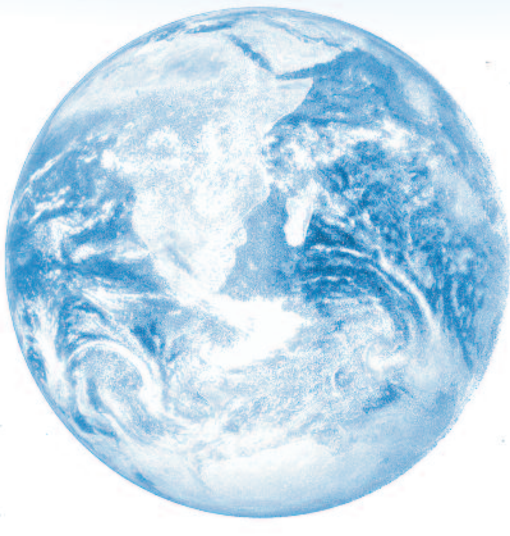
PEG-BOARD Research Discipline Use Cases



This document presents a number of user scenarios and use cases developed to support the wider community's use of palaeoclimate simulation and historical data, including the arts and humanities and the sciences.

The use cases presented are used as a source of functional requirements for PEG-BOARD metadata and API developments.

This information is currently being circulated for comments, and will be published following any requested amendments.

PEG-BOARD Research Discipline Use Cases

A report for the Joint Information Systems Committee-funded PEG-BOARD project

Document details

Authors:	Emma Tonkin and Greg Tourte
Date:	26/05/11
Version:	0.4
File Name:	
Level of QA Required (1-3):	
Level of QA Achieved (Y/N):	
Notes:	

Executive summary

This document presents a number of user scenarios and use cases developed to support the wider community's use of palaeoclimate simulation and historical data, including the arts and humanities and the sciences.

The use cases presented are used as a source of functional requirements for PEG-BOARD metadata and API developments.

This information is currently being circulated for comments, and will be published following any requested amendments.

Table of Contents

User scenarios

The user scenarios given here have been generated from a variety of sources:

- Interviews with BRIDGE staff
- Discussion with palaeoclimate researchers
- Interviews with researchers in other domains (evolutionary biology, archaeology, and earth science).
- Literature review

How to apply user scenarios and use cases

User scenarios are typically high-level/user-level descriptions of interaction with a computer system – a real-world narrative. They are intended to identify relevant tasks or events, present desired outcomes, and identify the task/process flow through which the interaction proceeds. A user scenario can provide evidence for user-level requirements, and give an idea of the functionality and data objects required to complete a task.

Because user scenarios are high-level and generally focus on the user-level perspective, we developed a number of user scenarios into use cases. These are longer, more complex and less readable, but attempt to describe both user-level and system-level interactions, and are a good basis for a formal specification.

Both are suitable for use as a resource for development of the PEG-BOARD revision requirements document.

Archaeology

Introduction

Archaeology is a broad and complex subject that draws upon many external sources. Its interdisciplinary nature reflects the fact that many aspects of the ancient world – culture, environment, biology, geographical and social context – are likely to have had an impact on the way in which practices and communities alter and develop.

The archaeologist is faced with a certain amount of evidence, often fragmentary and prone to complexities of interpretation and apparent contradiction, from which to puzzle out a vertiginous landscape of tens of thousands of years of human prehistory.

The role of technologies on the development of human societies and ways of life is profound and far-reaching. By 'technologies', the archaeologist refers to a large group of developments, for example:

- domesticated grain seed and livestock;
- effective methods of knapping flint;
- new architecture and construction mechanisms;
- administrative technologies allowing basic information on functional subjects to be written and read – in other words, writing.
- new developments, for example in metallurgy or textiles

Not all of these technologies are dependent on climate, but many are, and the ability to practice others is often dependent on the availability of a surplus enabling individual specialisation – specialist craftspeople can only exist when society has made it possible for that specialisation to occur. If the craftsperson is required to spend each day working for subsistence, then there is little chance of specialisation.

The climate may also act as a limiting factor in other ways. For example, consider the differential development paths of the Chatham Islanders (the Moriori) and the society from which they originated a few hundred years previously, the New Zealand Maori. The domesticated crops depended upon by the Maori were barely appropriate for the New Zealand climate, having arrived from Polynesia. They simply were not appropriate for the cooler, wetter climate of the Chatham Islands. As a consequence, the Moriori reverted to hunter-gatherer existence, and were eventually attached and routed by the larger and more warlike Maori population.

In the particular example of the Chatham Islanders, we know approximately what the prevailing climactic conditions were likely to have been during prehistory simply because the Western conception is that New Zealand's prehistoric period ended only in the 17th-18th centuries, and therefore it is somewhat unlikely to differ greatly from the currently measured values. In other cases, palæoclimate data is the only way to gain a good idea of the prevailing climatic conditions during the relevant period.

Furthermore, it has been suggested by many that the climate may be not only a limiting factor, but may also be causative – a catalyst for change. It is often suggested that the Younger Dryas may have been a causative factor for the development and adoption of agriculture by the Natufian people, formerly hunter-gatherers, in the Levant(Bar-Yosef & Belfer-Cohen, 1999).

To conclude, therefore, we see that the influence of the climate on the adoption of technologies – and indeed on a larger scale of the development of societies – is potentially very significant indeed. As a consequence, this information is of interest to a broad range of interdisciplinary teams and subject specialists on topics other than palæoclimatology. How best to support this type of work?

References

Bar-Yosef, O. and A. Belfer-Cohen (1999). "Facing environmental crisis. Societal and cultural changes at the transition from the Younger Dryas to the Holocene in the Levant." In: The Dawn of Farming in the Near East. Edited by R.T.J. Cappers and S. Bottema, pp. 55–66. Studies in Early Near Eastern Production, Subsistence and Environment 6.Berlin: Ex oriente.

Archaeology User Scenario 1: Historical climate information for a given area

Title	Historical climate information for a given area
Persona	Alice
Task	Retrieving climate information for a given area
User group	Students
Scenario	Alice is studying archaeology within her degree. She has been given an essay question that requires her to describe the likely factors behind the Viking movement in the 8th century AD and the fate of their expeditions to Vinland and elsewhere. She has read in various places that there is a strong likelihood that some of these changes have to do with climate – she is especially interested in the Medieval Warm Period. She knows roughly which areas are of relevance to her, and so she goes to the PEG-BOARD web site and picks out those areas on an embedded Google Map. On the options menu, she gives a time period covering the relevant times, and then clicks on ‘download report’. The report page she receives contains data tables for each area, providing her with several pieces of information about climate – rainfall, temperature, and so forth. It also provides a sequence of graphs showing the progression of climate change over the time period she has entered, along with a cautionary note explaining the expected accuracy of the data, its sources, and inviting her to contrast the data with additional sources such as ice cores. Alice reads the information carefully, and then writes her essay, citing the web site and using the relevant images as she builds up her argument.

Archaeology User Scenario 2: Mapping the Fertile Crescent

Title	Mapping the Fertile Crescent
Persona	Richard
Task	Developing learning materials for an introductory class
User group	Teaching staff
Scenario	Richard has been tasked with creating a series of learning resources to explain the 'cradle of civilisation' to a group of teenage students. He has decided that it would be best if he were to begin by siting the region according to present-day political geography, as he knows that many students would have difficulty identifying Mesopotamia on a map, and then to talk about the physical features of the Fertile Crescent, including the climate. He begins by creating his modern-day map of the region, using Google Earth. From this, he extracts a series of coordinates bounding the area under discussion. He then takes these coordinates to the PEG-BOARD site, where he requests a map of the area as it would have appeared at different times – from 11,000 to 6,000 BC. Richard would like to have a mechanism to improve the data presentation, as many palæoclimate models use a very low resolution grid compared to the sort of visualisation that students are used to seeing from the BBC. For example, he'd like to be able to superimpose important physical features like the Tigris and Euphrates rivers. However, in the meantime he is happy with the fact that the data is available via a scientific data creative commons licence, and that the images he is creating using the system can be used according to a creative commons attribution licence, making it very simple for him to repackage it into the school's learning materials. An additional advantage for Richard is that the kids like the idea of working with real data, as it makes for an interesting project rather than just going 'through the motions.'

Archaeology User Scenario 3: Exploring island ecologies

Title	Ocean currents and the 'express train to Polynesia'
Persona	Johann
Task	Assessing the impact of ocean currents on island ecologies
User group	Researchers
Scenario	Johann is a researcher looking at the impact of ocean currents on the lifestyles of island dwellers. He is aware of genetic studies and of linguistic evidence, written upon in some detail by Jared Diamond and others, to explain the colonisation of the Polynesian island grouping. He is looking at variation in the lifestyles and practices of various sites within the grouping, and exploring the impact of oceanic currents upon the subsistence methods used. This is a complex piece of work, and involves data taken from many sources. Contemporary information regarding ocean currents can be gleaned from satellite sources such as SEAWIFS, QuikSCAT and its successor, ASCAT. Trade winds, too, are well documented. However, historical data is not so easy to come by. Palaeoclimate models can only partially respond to this need, acting as they do on a lower-granularity simplified model of the world, but they do offer a starting point from which Johann can work. He accesses the site looking for significant changes in relevant factors ie. major ocean currents and similar, and for any historical observations or seed data that can be of use to him in scoping out the progression of his research project. His interest is broad and general and he is looking for a starting point rather than specific data, so he spends some time browsing the site before downloading some detailed data, which he will then use as one source among several in seeding his own model.

Archaeology User Scenario 4: Broadcast-quality video output

Title	High-definition video simulation output
Persona	Rebecca
Task	Extracting broadcast-quality video demonstrating climate change through the Younger Dryas
User group	Broadcaster
Scenario	Rebecca is working with a production company who are developing a project entitled 'The Cold World.' The resulting television programme has a lengthy section discussing wildflowers such as *Dryas octopetala*, which gave its name to the time period that it helped to define. They have three choices – to use the University's visualisation, to develop their own visualisation in-house, or to get an external organisation to do the work. Because the documentary will be published in HD, they need to ensure that the result is of a good enough quality to reuse in HD. Some years ago, this would not have been possible due to the bottleneck imposed on simulation research – the high cost of storage. However, as the price of storage has come down and the models have improved, it has become possible to offer such a visualisation service at a competitive price. Because scientific accuracy is important and Rebecca is not a subject expert, she assembles a small team – an external expert and a member of the BRIDGE project – to set the parameters of the visualisation that she needs. When the visualisation is complete, it would ordinarily become available on the service site. However, in this instance Rebecca asks that it be embargoed until the programme has been broadcast.

Archaeology User Scenario 5: Discovering data sources

Title	Discovering data sources
Persona	Nashida
Task	Identifying sources of information about palaeoclimate data
User group	Researcher
Scenario	Nashida is a researcher looking to build up an index of data sources. She is aware of a variety of existing sources, such as: **NGRIP**, the North Greenland Ice Core Project the **NOAA** index of proxy data and of present-day data, the **GISP2** data, the **PANGAEA** geoscience data publishing network, and a number of sources of palaeoclimate reconstruction data: the **NOAA** reconstruction index, the GISS and GFDL reconstruction portals, but she knows that there are many groups working on palaeoclimate reconstruction. She would like to glean an overview of the various groups' motivations, areas of interest, working methods and findings. As she is not a domain expert in the area of palaeoclimatology, she is rather uncomfortable with navigating the area and is looking for a simple place to start, such as an index and an FAQ.

Palaeoclimatology and Earth Science

Introduction

Given the current focus on climate science in the media, palaeoclimatology is a familiar topic today. As a research topic it is able to contribute to our understanding of the factors influencing climate over time, both through collection of evidence from the environment, and through simulation.

Many useful sources for historical climate data exist. A famous example is dendroclimatology, the collection of evidence about tree growth by studying growth rings, useful on the order of several thousand years into the past. Ice cores, which may be analysed for trapped gas, stable isotope ratios, and trapped pollen (Bruckner, n/d), are used as a resource to collect data up to several hundred thousand years into the past. Sedimentary rock can provide evidence from much further into the past, including evidence such as microbiota and pollen (ibid.).

Such sources offer patchy and incomplete coverage, providing only limited information and accuracy. Even so, they are useful as a basis on which to develop and test theories about past and future climate change, and an important factor in understanding changes in plant and animal population; climate models, such as the Unified Model (a product of the UK Met. Office) are used for this purpose.

References

Bruckner, M. (n/d). Paleoclimatology: How Can We Infer Past Climates? Retrieved from http://serc.carleton.edu/microbelife/topics/proxies/paleoclimate.html

Palaeoclimatology User Scenario 1: Comparing Palaeoclimate Models

Title	Comparing palaeoclimate models
Persona	An international research group
Task	Comparing and developing palaeoclimate modelling software
User group	Researchers
Scenario	Various climate modelling software is in use worldwide. An international group of researchers meet at an annual conference, and agree to compare the output of various climate models by collecting the outputs from various runs and comparing against: a) the output of other modelling software/configurations, and b) real-world data taken from other sources such as ice cores, sediment, and so forth. In order to make this possible, the lead institution puts in place a large repository database. They invite their partner institutions to run the same predefined experiment on their various platforms/software. Each institution then uploads the outputs from a predefined series of model runs to the database, alongside relevant metadata. After a while, they have enough data to make an extensive comparison of the models' behaviour on the specified type of problem. Using relevant external sources, such as an online database of palaeoclimate proxy data, they are also able to identify relevant real-world data. This enables the researchers to find out which model gives the best results on that particular class of problem, which models work less effectively, and the statistical variation between model findings. It also provides a starting-point from which to identify problematic areas within existing climate models, so that the model's performance can potentially be further improved.

Palaeoclimatology User Scenario 2: Exploring a geochemical hypothesis

Title	Exploring a geochemical hypothesis
Persona	Richard, a geochemist, and Hannah, a palaeoclimate researcher
Task	Sharing and hypothesising around primary data sources
User group	Geochemistry researchers, palaeoclimate researchers
Scenario	A geochemist, Richard, is involved in fieldwork in western Siberia. While collecting sediment samples and analysing their isotope composition, he comes up with a hypothesis concerning the relationship between local climate and the mechanism by which these deposits were formed. He therefore decides to contact a colleague, Hannah, and discuss the implications of this hypothesis. In a lengthy discussion, they develop a variation on his initial hypothesis that can be evaluated by means of a palaeoclimate simulation. Hannah configures the UM appropriately via the UMUI and sets up a UM experiment, compiling it. After configuring it, she runs a script to compile it on the HPC machine to which she has access. She then uses the script to add it to the queue. She periodically checks the queue status, and eventually sees that it has disappeared from the queue. Checking her files, she sees that output from the full experimental run is available, suggesting that the run was successful. She then takes her files to a separate machine for processing. Here, she converts them from UM proprietary format into NetCDF and calculates the stastical information necessary to check the original hypothesis, which centered on precipitation and ocean currents; notably she extracts air temperature and ocean temperature at a range of altitudes/ depths. She creates a series of graphs to share with Richard, and they schedule an online meeting to discuss the findings. Richard will later use these findings to design a future research project.

Palaeoclimatology User Scenario 3: Exploring the Panama Isthmus

Title	Exploring the Panama Isthmus
Persona	Jonathon, a palaeoclimate researcher
Task	Exploring the impact of past events on present-day climate
User group	Palaeoclimate researchers
Scenario	The formation of the Panama Isthmus is considered an important, relatively recent, geological event [1] in that it acted to separate the Pacific from the Atlantic. The volcanic formation of islands, which eventually silted up entirely to close the gap between the two seas, resulted in major changes in ocean currents and, consequentially, climate. For example, the Gulf Stream is a result of the existence of the Panama Isthmus. There are other consequences, such as an increase in salinity in the Atlantic ocean and the growth of the Arctic ice cap [2]. A palaeoclimatologist, Jonathan, decides to examine the impact that the formation of the Panama Isthmus has had. In order to answer this question, he develops two UM experiments, one in which the isthmus does not develop, and another in which it does. He compiles them and adds both to the queue. Once the runs are complete, he converts the data into NetCDF and calculates relevant statistical information - average temperatures, precipitation and so forth. By comparing the findings of each, he is able to establish just how events in the past have altered the present day climate. [1] http://earthobservatory.nasa.gov/ IOTD/view.php?id=4073 [2] http://www.whoi.edu/ oceanus/viewArticle.do?id=2508

Ecology/Palaeoecology

Introduction

The Collins English Dictionary (5th edition) defines paleoecology as 'the study of fossil animals and plants in order to deduce their ecology and the environmental conditions in which they lived'. The Concise Oxford English dictionary, on the other hand, defines paleoecology as 'The study of human communities in their environment and especially the effects that people had on the physical environment and vice versa.'

A region's ecology may be reconstructed (or modelled) on the basis of the fossil record. As we have already seen, the fossil record is already used as a source to deduce environmental conditions, such as climate variables. Climate variability has been shown to have 'major impacts on terrestrial and coastal ecosystems' (Willard & Cronin, 2007), although the immediate effect of a changing climate on a given species depends on biological factors such as a species' ability to regulate its own body temperature. Variables such as vegetation are also a causative factor in determining global climate, however; as an ecosystem alters, so does the climate.

Climate modelling is of interest to researchers working in areas such as evolutionary biology, but there are a number of challenges in linking palaeoclimate modelling to evolutionary biology, not least the complexity of the various models involved.

References

Willard, D. A., Cronin, T. M. (2007). Paleoecology and ecosystem restoration: case studies from Chesapeake Bay and the Florida Everglades. Front Ecol Environ 2007; 5(9): 491–498, doi:10.1890/070015

Palaeoecology User Scenario 1: Exploring the effect of palaeoclimate on species variation

Title	Exploring the effect of palaeoclimate on species variation
Persona	Sarah, a postgraduate research student
Task	Modelling the effect of climate instability on species variation
User group	Evolutionary biology research
Scenario	Having recently worked on a literature review on the subject of species diversity and climate, a student studying evolutionary biology, Sarah, decides to focus on a similar subject in her own MSc research. She has read that species variation is associated with climate stability in reptiles and amphibians (Araujo et al, 2008), whilst in mammals and birds, present-day climatic conditions have a greater impact on species diversity (Hawkins et al, 2003). She has also read that the Australian context is particularly interesting, as the country by and large escaped major extremes in climate. With her supervisor, she develops a research plan which explores the appearance of monotremes, marsupials and amphibians. She uses a dataset of present-day species occurrence made available by the Australian government[1]. Having made sure that she has enough information to begin the study, she contacts a palaeoclimatologist with whom she discusses the study plan. They agree on her needs: she's looking for a realistic dataset that can describe Australia's climate over the last twenty thousand years, in one-month steps. Fortunately an appropriate dataset already exists, so there is no need to develop or run a new study initially. Her contact points her to the relevant page on the web interface so that she can download what she needs. As her research proceeds, she finds it necessary to generate a set of figures of appropriately high resolution to be sent to an academic journal. She emails her contact to ask for help solving this problem, who returns to the website and generates a higher-quality image using the service interface.

	Araujo, MB; Nogues-Bravo, D; Diniz-Filho, JAF; Haywood, AM; Valdes, PJ; Rahbek, C (2008) Quaternary climate changes explain diversity among reptiles and amphibians, Ecography, 31, pp.8-15. doi:10.1111/j.2007.0906-7590.05318.x Hawkins, B. A. and Porter, E. E. (2003). Relative influences of current and historical factors on mammal and bird diversity patterns in deglaciated North America. Global Ecol. Biogeogr. 12: 475-481. Young, J. (2010). Australian amphibians among the oldest in the world. Australian geographic, July 2010. http://www.australiangeographic.com.au/journal/australian-amphibians-among-oldest-in-the-world.htm [1] http://www.environment.gov.au/biodiversity/threatened/publications/action/marsupials/10.html

Palaeoecology User Scenario 2: Exploring the interactions between palaeoclimate and vegetation

Title	Exploring the interactions between palaeoclimate and vegetation
Persona	Sam, a palaeoclimate researcher
Task	Understanding impact of terrestrial vegetation
User group	Palaeoclimate researchers, archaeologists, evolutionary biologists
Scenario	Terrestrial vegetation has a number of effects on the climate, including the energy cycle, the water cycle, and the carbon cycle (Wang et al, 2005). Sam decides to review the results presented by Cowling et al (2008), notably the suggestion of 'a considerably greater presence of broadleaf trees than assumed previously' in continental Africa during the last glacial maximum. Sam discusses the problem with his colleagues, and identifies an available model, such as HadGEM, that permits him to experiment with the results of assuming a variety of models of vegetation growth and dieback. He designs a series of experiments that adequately reflect the different possibilities identified in Cowling et al, and then configures and compiles the model for each one, adding each one to the HPC queue. When they are complete, he takes the resulting data and transforms it into NetCDF. He then generates a series of relevant data tables that allow him to compare the results of his experiment with the results described by Cowling et al. Cowling, S. A., Cox, P. M., Jones, C. D., Maslin, M. A., Peros, M. and Spall, S. A. (2008), Simulated glacial and interglacial vegetation across Africa: implications for species phylogenies and trans-African migration of plants and animals. Global Change Biology, 14: 827–840. doi: 10.1111/j.1365-2486.2007.01524.x Wang, Y, Mysak, LA, Wang, Z. and Brovkin, V (2005). "The greening of the McGill Paleoclimate Model. Part I: Improved land surface scheme with vegetation dynamics." Climate Dynamics, 24, 469-480, doi: 10.1007/s00382-004-0515-9,

	2005

Palaeoecology User Scenario 3: Modelling evolution and distribution of bats

Title	Modelling evolution and distribution of bats
Persona	Jane, an evolutionary biologist
Task	Understanding impact of changing climate on the migration and hence speciation of bats
User group	Palaeoclimate researchers, evolutionary biologists
Scenario	There is relatively little evidence about the evolution of many species, as insufficient information is available. As such, modelling is one approach that may permit theories to be built about species development and migration. In order to approach this problem, several different aspects of the environment must be modelled. Climate modelling permits a rough understanding of the likely climatic conditions (although on a very approximate level, eg. 1 degree resolution). From that information and further knowledge about the local environment it is possible to model relevant biological systems, such as vegetation. Because bats are homeotherms, they thermoregulate, and therefore do not need to migrate to follow optimal climate conditions. However, whilst they do not closely track the climate, the are likely to track the foodstuffs to which they are best adapted (variable depending on species). To complete this complex mapping problem, Jane uses data from a realistic model run taken from the BRIDGE project, covering the relevant timescale (approx 8 million years). She uses a transform function to extract relevant information in the WORLDCLIM BIOCLIM dataset standard format, giving her relevant bioclimate variables such as peak temperatures and extremes of precipitation layers. When she has completed her model, she generates images from the palaeoclimate data, layering her own findings on top. She then runs the data through an intermediate software designed to improve the display resolution of the

	data (adding a 'health warning' about the actual precision of the data to the image's key).

Use cases

These use cases are chosen as representative examples developed from the scenarios above. They have been created collaboratively on the basis of the scenarios.

Use case 1: Mapping the fertile crescent

Use case brief: Selecting datasets

Use Case Name	Selecting datasets
Actors	Data consumer Data repository Web interface
Use case description	The user defines their requirements with reference to their task. The user identifies: • required geographical and temporal coverage • dataset characteristics ('realism' and resolution) • output data variables The user will provide this information to the system. The system will respond with: one or more matches; one or more near matches; or a failure message. The failure message will provide an alternative mechanism such as browsing. The failure message will also recommend support/help documentation.

Use case: Selecting datasets

Use Case Name	Selecting datasets
Actors	Data consumer Data repository Database Web interface
Triggers	The user sets parameters for dataset selection.

Preconditions	The user may be logged in OR The user may have anonymous access The user has set parameters for search
Postconditions	The system will evaluate user permissions The user will receive a set of candidate dataset descriptions tailored to user's permissions The user will be given a set of possible actions (ie. datasets to choose from).
Normal flow	1. The user will submit their search parameters to the web service 2. The web service will evaluate user permissions/login status 3. The web service will hand the request across to the database 4. The database will identify the subset of records to which the user class is granted access 5. The database will test each record for compliance with the user's search parameters 6. IF the database identifies a nonzero number of possible matches, the database returns that set of possible matches 7. The web service retrieves that set of possible matches, formats it appropriately with actionable links, and returns it to the user.
Alternate flows	6B. IF the system does not identify a nonzero number of possible matches, the database tests each record for near compliance with the user's search parameters. 6C. IF the database identifies a nonzero number of possible matches, the database returns that set of possible matches. OR 6C. IF the database does not identify a nonzero number of possible near matches, the database returns no matches. 7. The web service identifies the failure of the database to retrieve possible matches and returns an

	appropriately formatted message inviting the user to either: 8. Change some search parameters OR 8. Widen search parameters OR 8. IF the user is logged in, there may be a permission flag in their profile that enables them to request a dataset to be developed.

Use case 2: Mapping the fertile crescent

Use case brief: Retrieving dataset

Use Case Name	Retrieving dataset (including for m2m reuse)
Actors	Data consumer Data repository Web interface
Use case description	The user has chosen the required dataset from the search results. The user clicks on the required dataset, and is presented with a specific page for that dataset, providing more information on the dataset. The user can choose between precreated outputs , be it graphics or data in specific formats, or the possibility of creating plots from that dataset that have not been created previously (ie. custom plots). If the user is logged in, and creates a new plot, that new configuration may be added to stored system presets on the user's request.

Use case: Retrieving datasets

Use Case Name	Retrieving dataset (including for m2m reuse)
Actors	Data consumer Data repository Database Web interface
Triggers	The user selects a dataset from the list presented by the system.
Preconditions	The user may be logged in OR The user may have anonymous access The user has selected a dataset from the list presented by the system
Postconditions	The system will evaluate user permissions The user will be shown detailed information about the dataset.

	The user will be given a set of possible actions (ie. generation of graphs showing the dataset, dataset download in raw format, NetCDF, etc).
Normal flow	1. The user will submit their choice of dataset to the service 2. The web service will evaluate user permissions/login status 3. The web service will hand the request across to the database 4. The database will identify the subset of records to which the user class is granted access 5. The database will return this information to the web service 6. The web service will, IF the user is permitted access, hand the request across to the data repository 7. The data repository will retrieve metadata describing what is available for that particular experiment, and return it to the web service 8. The web service will transform the metadata returned according to appropriate formatting rules (ie. XML OR HTML OR JSON) 9. IF HTML is chosen, the user is given a page offering a number of choices: between precreated outputs LISTING graphics or data in specific formats INCLUDING PostScript, PDF, PNG, JS GIF, NetCDF , CSV, XLS, or custom outputs such as previously uncreated plots from that dataset. 10. IF the user chooses to create a custom plot, and IF the user is logged in, the web service offers an option to STORE that configuration in stored system presets. 11. IF the user chooses that option, the web server STORES the configuration in user-level storage, AND adds it to the administrator’s inbox to review for general access. 12. IF groups are used within authentication then group-level access to custom plots is an additional option

Alternate flows	6A. IF the user is not permitted access then the web server returns an ERROR message. OR 9A. IF HTML is not chosen then data will be returned in the chosen format, without formatting information. OR 10A. IF the user is not logged in then no storage mechanism for custom plots is available. 11A. The user is presented with the results of the custom plot process.

Use case 3:

Use case brief: Generating datasets

Use Case Name	Generating datasets
Actors	Data consumer Data repository Web interface PUMA (UMUI) HPC facility HPC queueing system,
Use case description	The user begins the process of running an experiment by choosing a model, a resolution, a set of initial conditions and forcings. The user logs in to the PUMA facility and runs the UMUI to set up the experiment. They have the choice of creating a completely new experiment or duplicating an existing experiment from any users and modifying it to suits their needs. Once the experiment is configured, the user runs a script, which sends the experiment to a supercomputer of his choice. The experiment is then compiled on the HPC facility and added to the queueing system if it compiled properly. Once the experiment arrives at the head of the queue, the queueing system will send it to its assigned nodes where it will be run. Back at Bristol, the user has to initialise the experiment on the BRIDGE system, which includes adding a description and metadata, which are then stored in a database. The BRIDGE processing nodes will then automatically download all possible data from initialised experiments off the clusters the user has access to, including running ones (to save usually limited storage space on HPC facilities) and delete all non required data off it. This process include converting all files to NetCDF Once the experiment has successfully run, and all data has been transferred to the local servers, the user accesses the web interface and selects a set of processing steps, which must be applied to the data. These steps are added to a queueing system to manage load on the processing nodes. These processing step will usually apply averages and conversions to the data as well as create a set of pre-defined default plots and graphs. All these processed outputs are store on the

	data repository and the dataset is then added to the web interface as a finished product. If the experiment is valid and the data is as expected and useful, the user can request the raw data be archived to tape and removed from the processing nodes.

Use case: Comparing datasets

The detailed use case is not suitable for this reporting format, and will be covered in an accompanying document.

Use case 4:

Use case brief: Comparing datasets

Use Case Name	Comparing datasets
Actors	Data consumer Data repository Web interface
Use case description	The user chooses a variable (type of data, such as 'climate'). The web server then presents a series of choices allowing users to choose two experiments to compare on the basis of the variable chosen in the previous step. The web server then provides an interface allowing the user to compare results either by comparing two graphs, or calculating the difference between the two datasets.

Use case: Comparing datasets

Use Case Name	Comparing datasets
Actors	Data consumer Data repository Database Web interface
Triggers	The user selects a dataset from the list presented by the system.
Preconditions	The user must choose a data type to examine The user may be logged in
Postconditions	The system will evaluate user permissions The user will be shown comparison data between the two datasets. The user may manipulate the graphs in various ways to explore the differences between them.
Normal flow	1. The user will submit their choice of data type to the web server

	2. The web service will evaluate user permissions/login status 3. The web server then presents the datasets that the individual is able to access (and which are amenable to the type of analysis in question). 4. The user must choose two datasets for comparison. 5. The user must choose a variable on which to compare them. 6. The user must choose a type of plot/output format. 7. The user must choose a projection (longitude/latitude – coverage). 8. The user chooses the size of the resulting file. 9. The user submits his/her choices. 10. The user waits for the server to retrieve and process relevant data. 11. The web service retrieves relevant data from the data repository. 12. The web service processes the data via a processing node. 13. Once the data processing is complete, the data is made available on the website. 14. The web server returns a base page to the end user containing relevant information (including the newly generated graph).
Alternate flows	9A The user should be able to submit a job and leave it to pick up later. 9B The web server should return to the user and announce progress, once the job is complete. 9C The user should be able to check all of his jobs on the web interface, and see relevant progress. (See 14).

Use case 5: Modelling evolution and distribution of species

Use case brief: Reprocessing raw data

Use Case Name	Reprocessing raw data
Actors	Data consumer Data repository Web interface Data archive Processing Node
Use case description	The user identifies the experiment that holds the data required and realises that the processing applied to the data has trim the relevant information for his analysis. However the raw data is still available in archives. The user asks for the required experiment to be recovered from the archive and its data places on one of the BRIDGE processing nodes. The user then goes back to the processing part of the website (as used to create a dataset) and uses the custom processing facility to generate the required data. This is done in the form of scripts that will be added to the processing node's queueing system and eventually run to manipulate the data.

Use case: Reprocessing raw data

Use Case Name	Reprocessing Raw Data
Actors	Data consumer Data repository Database Web interface Data Archive Processing node
Triggers	The user selects a dataset from the list presented by the system.
Preconditions	The user must have selected the experiment required The user already knows the already available processed data is not sufficient The user must be logged in

	The user must be a BRIDGE user with access to BRIDGE processing nodes The required dataset must still be on the processing nodes or still in archive
Postconditions	The system will evaluate user permissions The user will be shown requested dataset. The user may manipulate the graphs in various ways to explore the differences between them. The user will be give access to retrieved archive data
Normal flow	1. The user will access the processing section of the website as per use case 3 (generating dataset) 2. The web service will check if the raw data is available on the processing node. 3. The user will go to the custom processing menu 4. The web service will present the variables available in this dataset along with associated parsed CF metadata. 5. The user will chose the first data to manipulate 6. The web service will present a set of alternative processing functions that can be applied to the selected variable 7. The user will select his required function. 8. The web service will add this variable/function combination to a todo list and offer the user to either process current list or add a new variable/function process. 9. Step 5,6,7 and 8 will be repeated until the user chooses to process the list. 10. The web service will generate processing script for the todo list and add it to the processing node's queueing system 11. The processing node will execute the script on the raw data and generate relevant data files, graphs and plots and add them to the experiment's processed data. 12. The web service will update the metadata

	database entry for this experiment to describe the newly created data using metadata associated with the processing scripts used. 13. The web service will present the newly created data, graphs and plots along with the existing one for the experiment 14. The user will choose the file formats and size (for raster plots and graphs) of his files to view /export/download.
Alternate flows	2a. The web service identifies that the data is no longer available on the processing nodes. 2b. The web service checks the archive database to see if the raw data is available in the archive. 2c. If the dataset is still in archive, the web service extracts the variables available in this experiment from the archive database. 2d. The web service will add an archive extraction script to the todo list. Or 2c. If the dataset is no longer in the archive, the web service will display a message to the user which details the reason why the dataset is no longer available (eg. Dataset is too old and has been remove following data policy).

Appendix A: Interview script

Internal to BRIDGE

1. Time/availability
2. What is your role?
3. How would you describe your research area or areas?
4. What sorts of data do you use, and where is it sourced from?
5. If the data weren't available from Bristol, where else might you get it?
6. What sort of data search mechanism might be useful to you?
7. What sorts of data do you generate?
8. Do you store it on the BRIDGE service, and/or an external data centre?
9. Do you have any particular requirements regarding data formats, metadata format, plots/presentation, preferred units, etc?
10. If you are making use of the BRIDGE website software on the server-side: are the technologies used appropriate, or would you prefer other APIs/languages to be used?
11. Do you have any further comments, complaints, suggestions, or feature requests? What are the best and worst things about the existing service?

External to BRIDGE

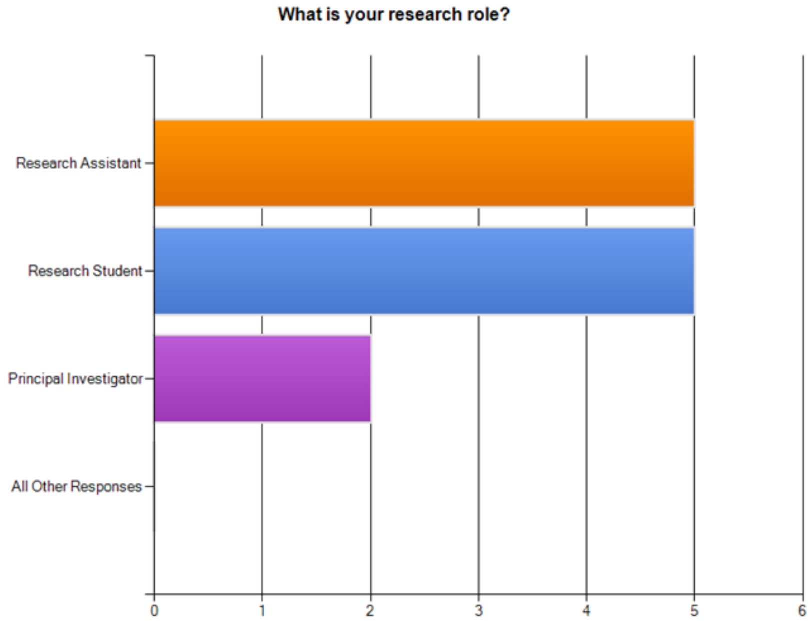
1. Time/availability
2. What is your role?
3. What is your research area or areas?
4. What sorts of data do you use, and where is it sourced from?
5. If the data weren't available from Bristol, where else might you get it?
6. What sort of data search mechanism might be useful to you?
7. What sorts of data do you generate?
8. Do you have any particular requirements regarding data formats, metadata format, plots/presentation, preferred units, etc?
9. Could you talk me through an example of a research problem that uses this sort of data – for example, coming up with a research idea, identifying relevant data, and analysing it?
10. Do you have any further comments, complaints, suggestions, or feature requests? What are the best and worst things about the existing service?

Appendix B: Interview responses

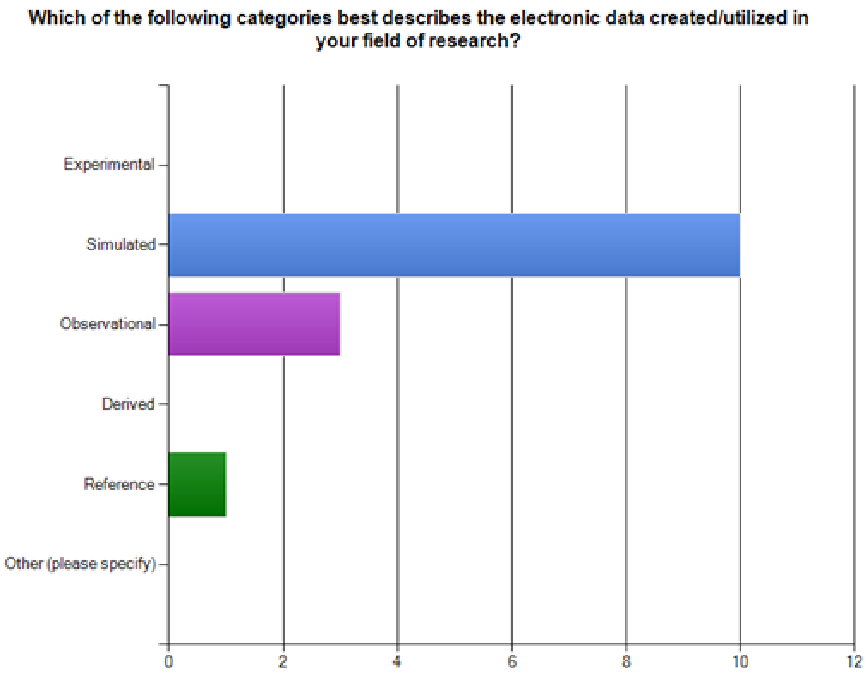
Interview responses have not been shared outside the project due to privacy concerns.

Appendix C: Questionnaire results and findings

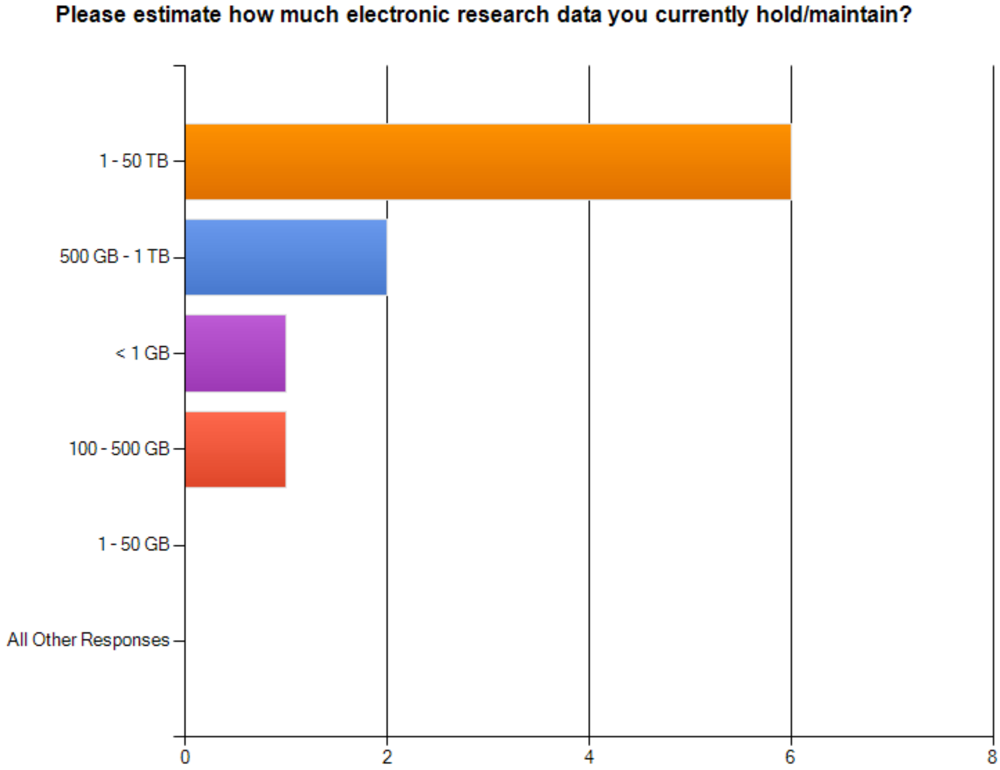
Research roles of respondents



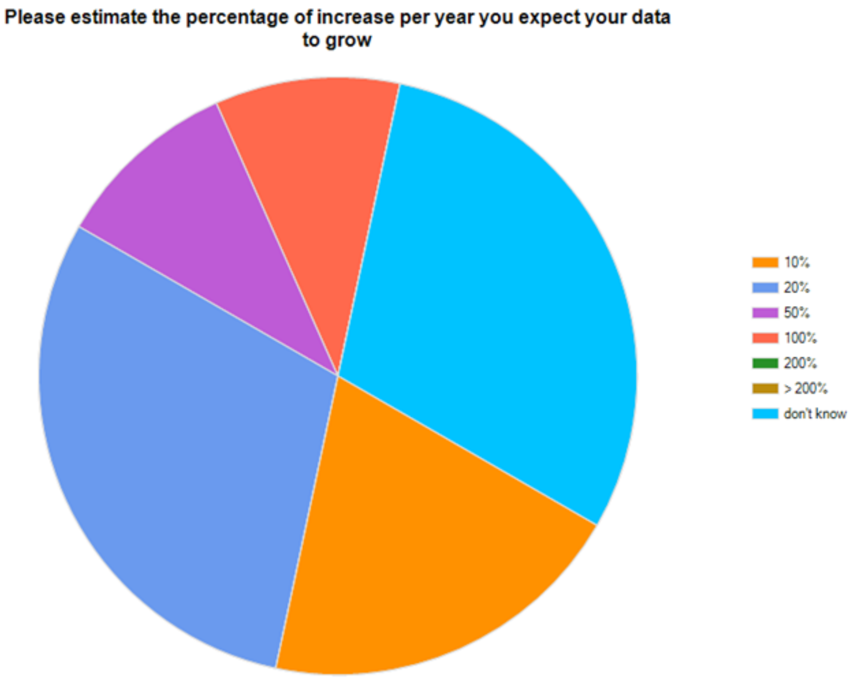
Data used in research



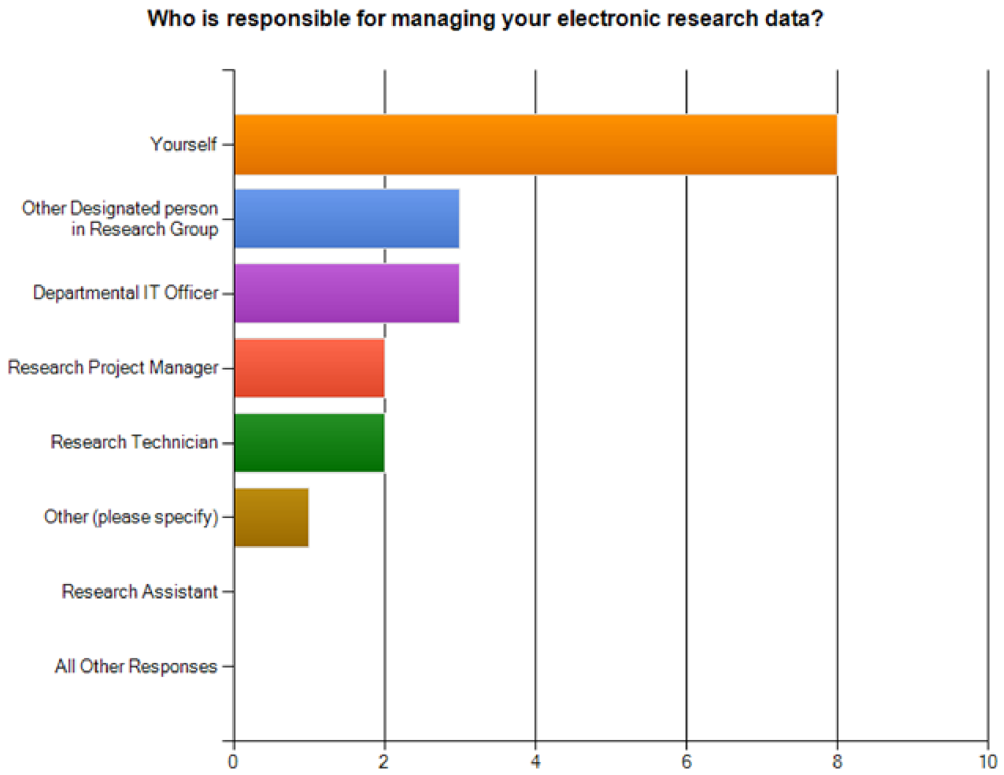
Quantity of electronic research data held/maintained



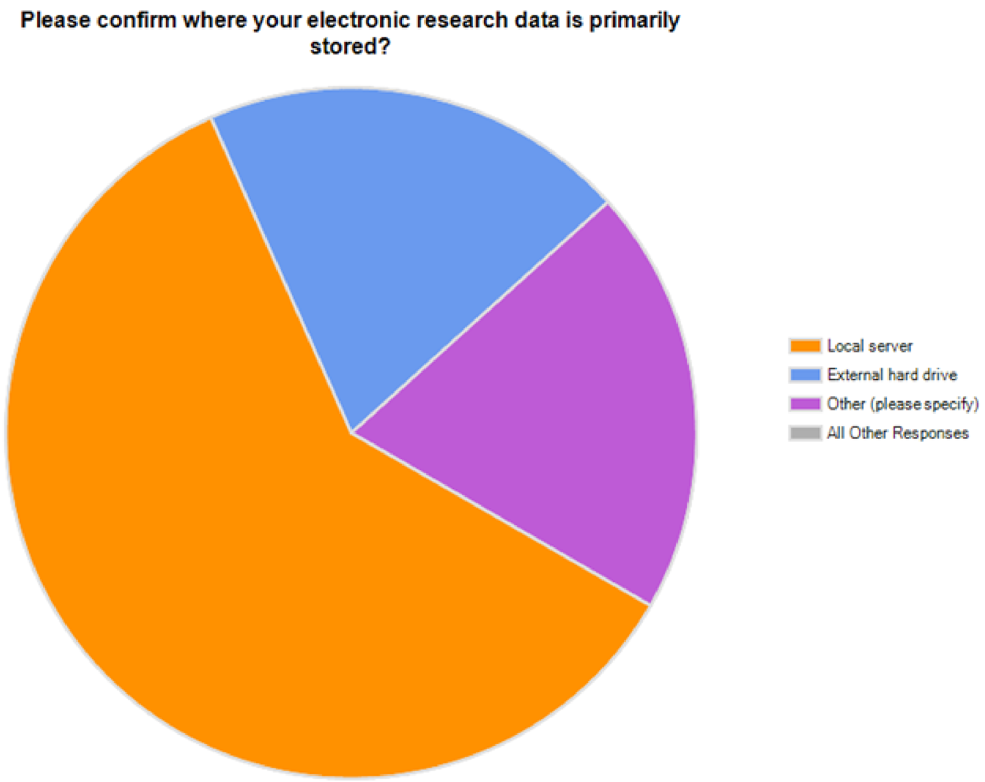
Increase in data storage requirements



Who manages your research data?



Research data primary storage

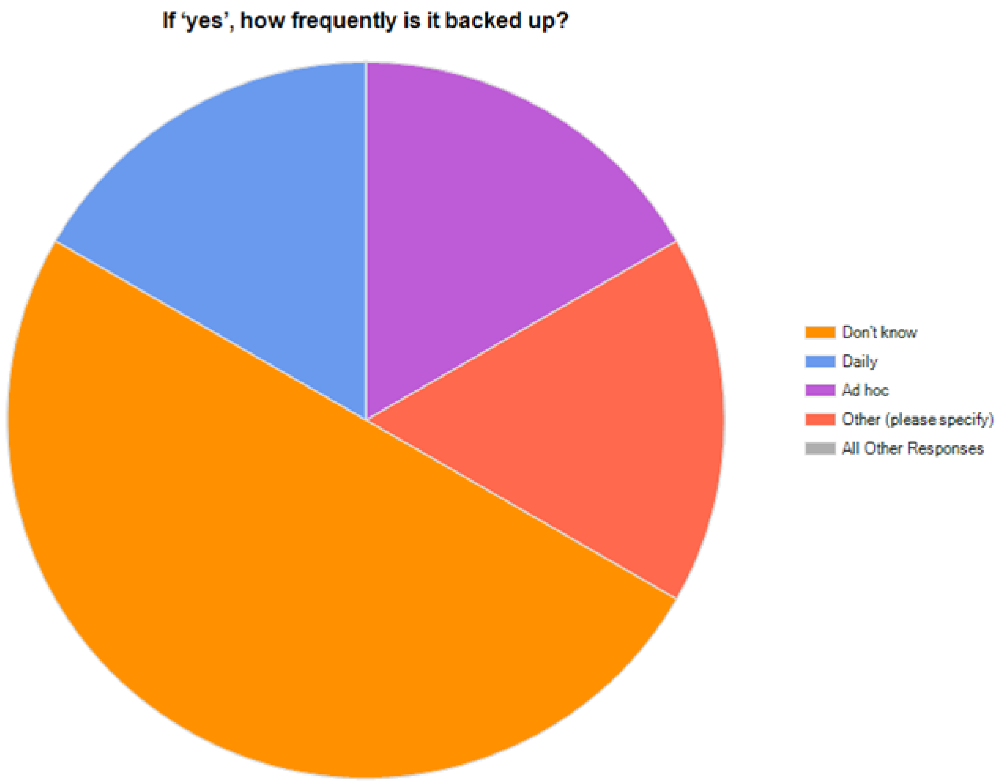


Others: 100% referred to non-local BRIDGE installation as primary storage.

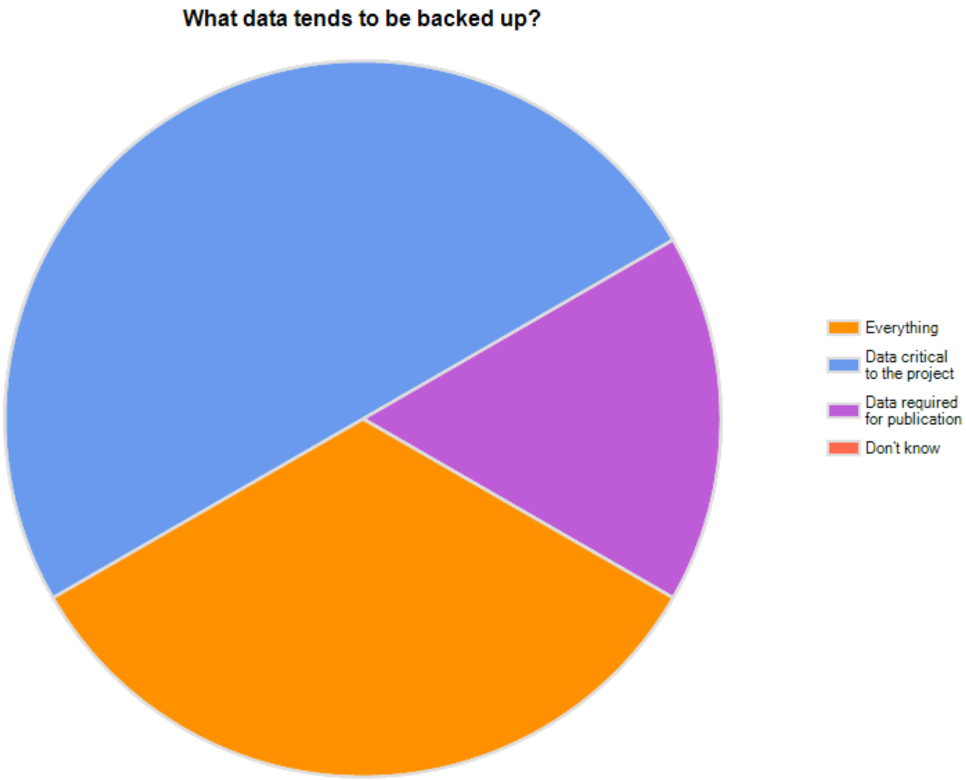
Regular backups

60% of participants stated that their data was backed up regularly.

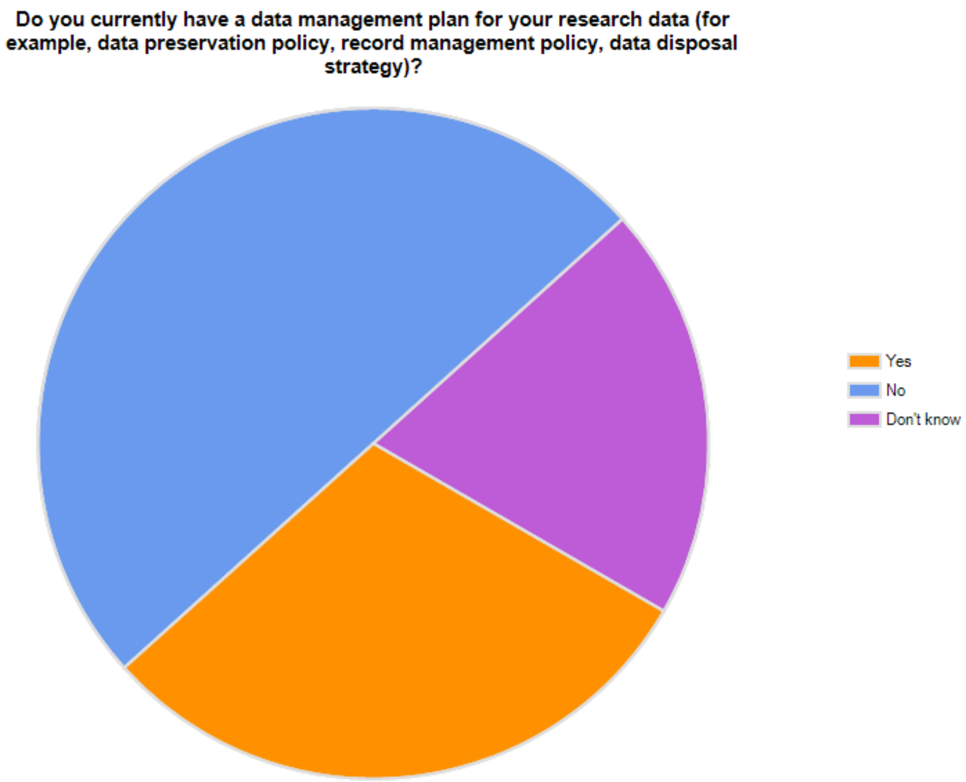
Frequency of backup



Data backup practices

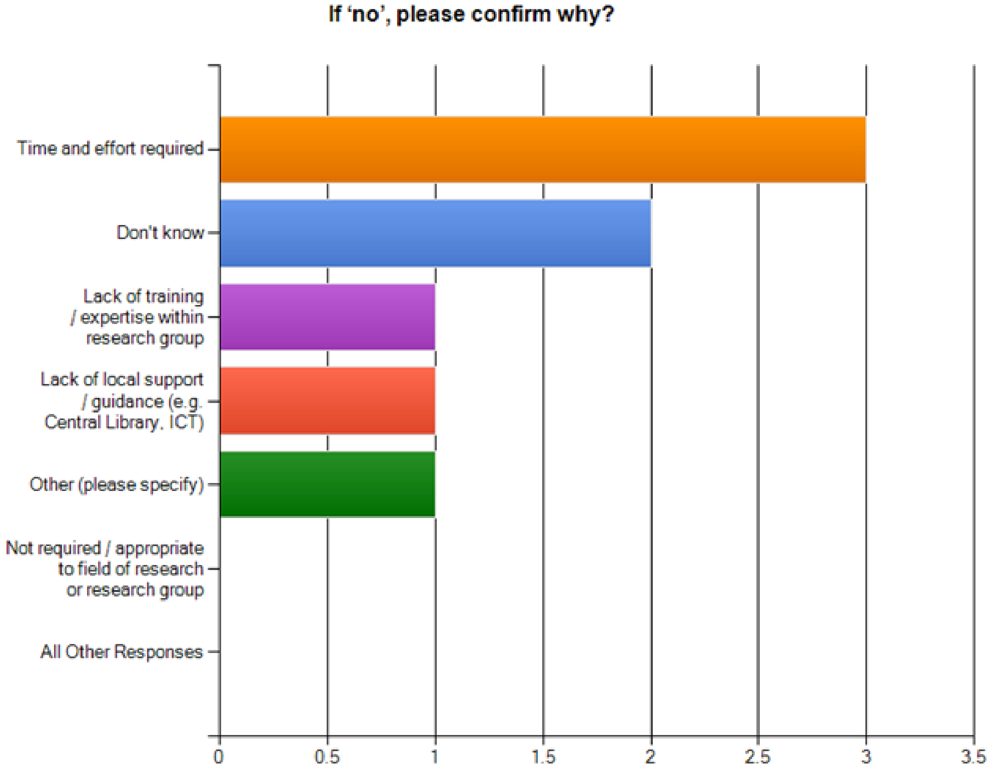


Data management plan



Of those with a data management plan, most gave their motivation for developing the plan as 'the volume of data', followed by 'the size of the project team'.

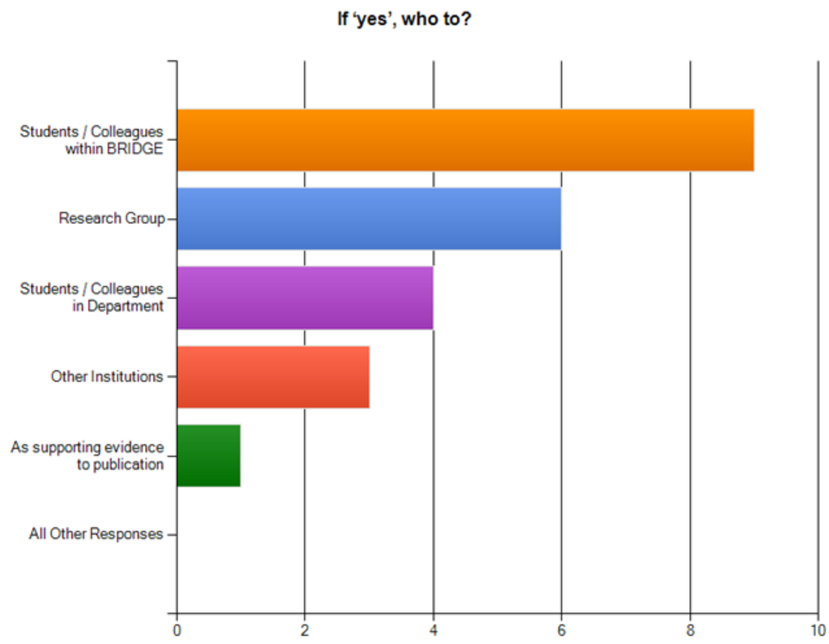
Reasons for not using a data management policy



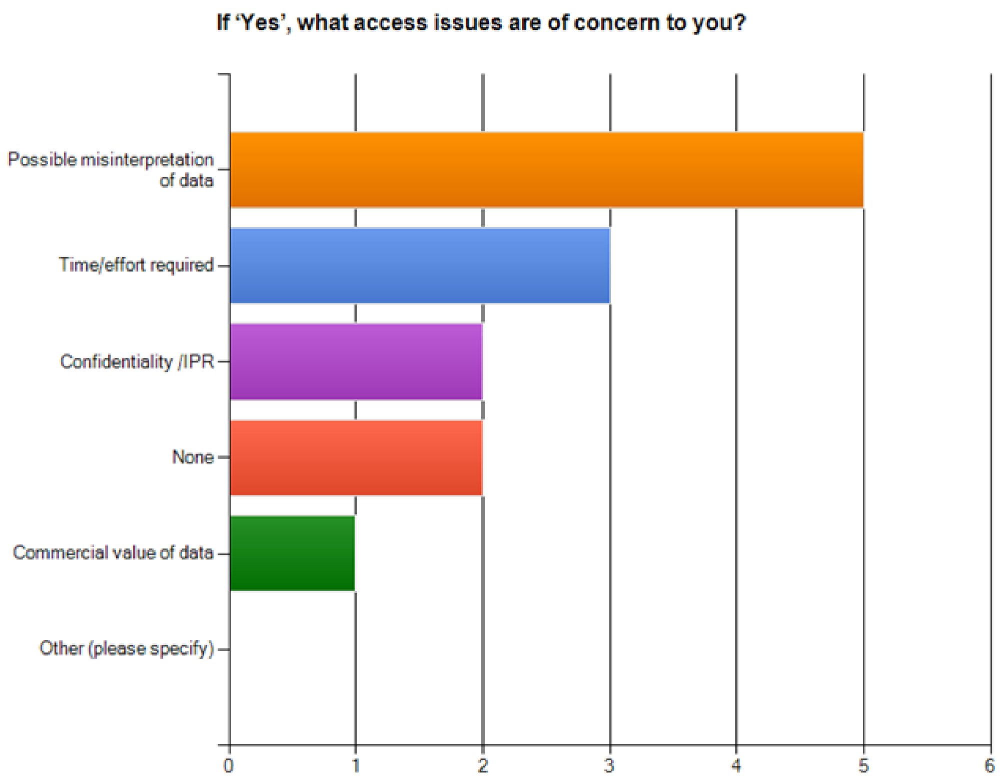
Open access

90% of users allowed at least some other individuals access to their datasets.

In practice, almost all of these users' data are open to anybody using the BRIDGE website, including the general public. The data below therefore reflects users' perceptions of access rights to the data rather than the practical availability of the data.



Access issues of concern



One user had blocked all access rights to generated data due to concerns about confidentiality/IPR.

Open access policy from funder

20% of users had been asked to make research data openly available outside of formal publication, 70% of users had not, 10% were unsure.

Externally sourced data

80% used at least one source of data other than the BRIDGE website, including simulation data, data from scientific instruments, data from prior publications, labnotes, and externally sourced Excel sheets.

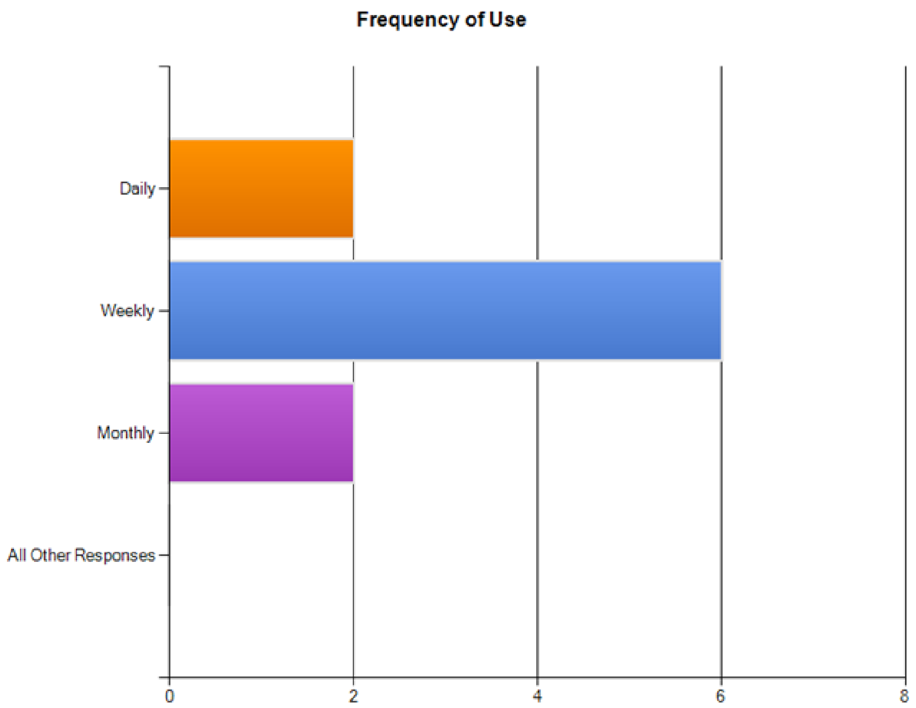
Time required to generate data

70% of respondents stated that their data takes on the order of months to generate.

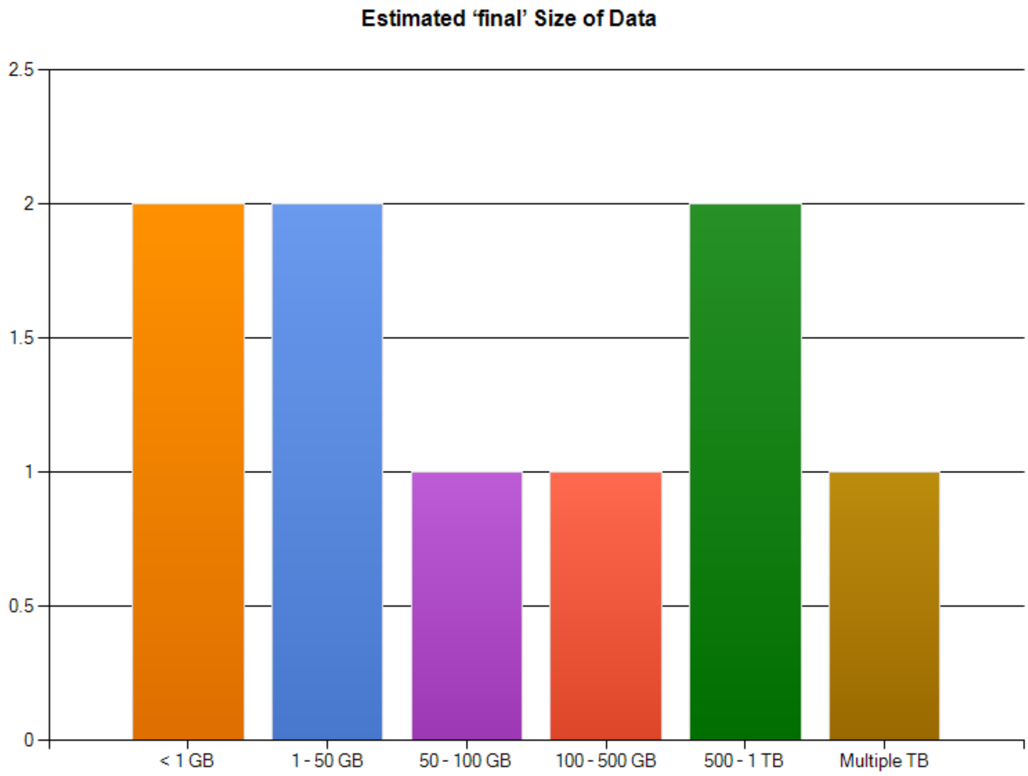
Data retention period

Almost half stated that they would retain data for 1-2 years; 20% expected to keep the data for up to ten years; over 20% expected to retain data 'indefinitely'.

Frequency of use



Estimated size of dataset overall



Proprietary format

Half of respondents stated that they used no proprietary formats for storing critical research data. Others stored data in the proprietary Met Office format.

Mission-critical nature of BRIDGE

70% of respondents stated that they could not complete their work without the BRIDGE service, 10% that they didn't know; 20% could complete their work without using the service.

Of the 20% that did not require the BRIDGE service, 80% did not know whether they could source the necessary data for their research outside the BRIDGE research group, whilst 20% stated that they would not be able to do so.

80% used the BRIDGE service to process their datasets, generated by the UM, whilst 20% did not.

60% used standard plots generated out of processed data, 30% did not, and 10% did not know. 50% used the BRIDGE service to generate customised plots out of processed data, whilst 50% did not; 50% used BRIDGE to generate presentations of their data for other researchers, whilst 50% did not.

90% of respondents reusing generated data used the NetCDF files directly, whilst 60% used GIF/PNG plots and 50% used animated GIFs and postscript/PDF plots.

50% of respondents needed to modify the data produced by BRIDGE. For half of respondents, this process took less than a day; for half, the process took one or more days.

Types of modification

Free-text responses included:

- change spatial/temporal averages, change graphics layout
- mainly processing to find e.g. the value at a certain point, or customised plots, or variables not generated by the webpages
- climatologies of fields not included in standard processing, time series analysis, e.g. PCA, spectral analyses etc.
- replotting using Ferret or [R] for specific comparisons

Experiment comparison feature

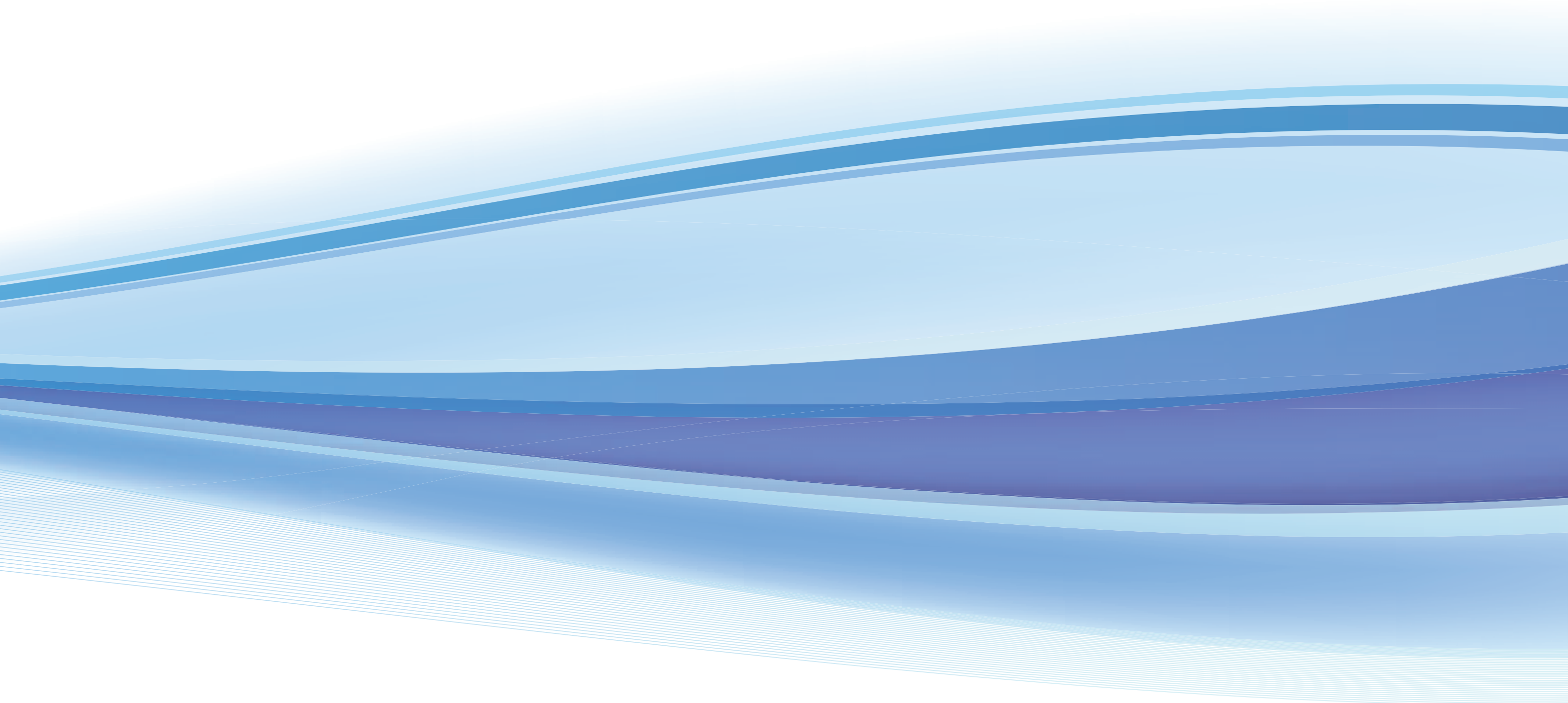
80% of respondents used the experiment comparison feature of the site.

Further requested modifications to the site

Almost 80% of respondents felt that modifications could usefully be made to the site. Notable suggestions included:

- Regridding; calculating score against observational data (such as a root mean square error); location and name of the main model variables in the netcdf files; atmospheric stream function data. Possiblity of adding new features.
- Incorporation of the Glimmer ice sheet model similar to BIOME/TRIFFID would be very beneficial The ability to produce publication quality plots from the website
- there are many diagnostics for climate data, but it is not obvious what the output means or what has been calculated. Also, it can be hard to get the diagnostic data out into a file required to make a publication quality plot. It would be good to get publication quality plots from the web pages. It is also hard to use with limited area model data. As much of the system is set up with global models in mind
- better graphic control larger choice of variables / periods

- ability to make plots of more complicated arithmetic functions (e.g. filed 1 minus filed 2 plus filed 3 times filed 4)
- more flexible time series plots statistical significance maps
- make code used to produce plots easily visible so that it can be copied and modified to make minor tweaks to individual graphs.



UKOLN
University of Bath
Bath BA2 7AY
Tel: 01225 38 6580
Fax: 01225 38 6838
www.ukoln.ac.uk