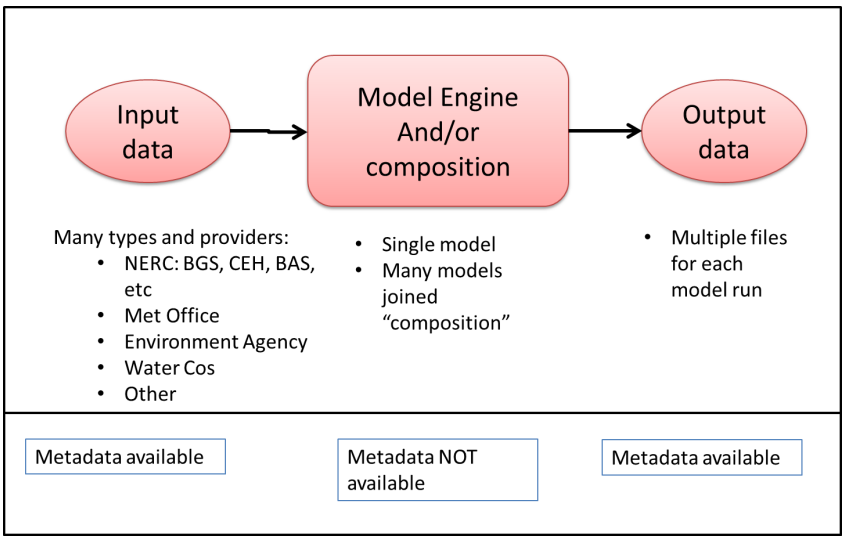
British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

Meta-model: Ensuring the widespread access to metadata and data for environmental models - Scoping Report

Environmental Modelling Programme

External Report OR/13/042



BRITISH GEOLOGICAL SURVEY

Environmental Modelling PROGRAMME

EXTERNAL REPORT OR/13/042

Meta-model: Ensuring the widespread access to metadata and data for environmental models - Scoping Report

Hughes A.G., Harpham, Q.K., Riddick, A.T, Royse, K.R. and Singh, A.

Bibliographical reference

HUGHES, A; HARPHAM Q, RIDDICK A, Royse K, Singh A. 2013. Meta-model: Ensuring the widespread access to metadata and data for environmental models - Scoping Report. *British Geological Survey External Report*, OR/13/042. 47pp.

Keyworth, Nottingham British Geological Survey 2013

BRITISH GEOLOGICAL SURVEY

The full range of our publications is available from BGS shops at Nottingham, Edinburgh, London and Cardiff (Welsh publications only) see contact details below or shop online at www.geologyshop.com

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

We publish an annual catalogue of our maps and other publications; this catalogue is available online or from any of the BGS shops.

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

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

British Geological Survey offices

BGS Central Enquiries Desk

Tel 0115 936 3143 Fax 0115 936 3276

email enquiries@bgs.ac.uk

Environmental Science Centre, Keyworth, Nottingham NG12 5GG

Tel 0115 936 3241 Fax 0115 936 3488

email sales@bgs.ac.uk

Murchison House, West Mains Road, Edinburgh EH9 3LA

Tel 0131 667 1000 Fax 0131 668 2683

email scotsales@bgs.ac.uk

Natural History Museum, Cromwell Road, London SW7 5BD

Tel 020 7589 4090 Fax 020 7584 8270

Tel 020 7942 5344/45 email bgslondon@bgs.ac.uk

Columbus House, Greenmeadow Springs, Tongwynlais, Cardiff CF15 7NE

Tel 029 2052 1962 Fax 029 2052 1963

Maclean Building, Crowmarsh Gifford, Wallingford OX10 8BB

Tel 01491 838800 Fax 01491 692345

Geological Survey of Northern Ireland, Colby House, Stranmillis Court, Belfast BT9 5BF

Tel 028 9038 8462 Fax 028 9038 8461

www.bgs.ac.uk/gsni/

Parent Body

Natural Environment Research Council, Polaris House, North Star Avenue, Swindon SN2 1EU

Tel 01793 411500 Fax 01793 411501

www.nerc.ac.uk

Website www.bgs.ac.uk

Shop online at www.geologyshop.com

Acknowledgements

This project has succeeded due to the significant input of a number of people. The input of the respondents of the on-line questionnaire is gratefully acknowledged. In particular, the following are gratefully acknowledged:

- Helen James from the Environment Agency for co-ordinating the response from her organisation.
- Gary Baker, BGS, and Gwyn Rees, CEH, from the respective NERC data centres for providing answers to the questionnaire and subsequent time for further explanation.
- Prof. Andrew Wade, Reading University and Dr Debroah Hemming, Met Office Hadley Centre for participating in follow up phone calls on the questionnaire.
- Finally to Carl Watson, BGS, for providing a review and helpful suggestion for improvement.

Contents

FIGURES

TABLES

Summary

This work is a response to the challenge posed by the NERC Environmental Data call, and is designed to scope out how to meet the following objectives:

1. Ensure that the data used to create models are recorded and their source known.
2. The models produced are themselves available.
3. The results produced by these models can be obtained.

To scope out how to fulfil these objectives a series of visits, phone calls and meetings were undertaken, alongside a Survey Monkey (on-line) questionnaire. The latter involved sending out a request to fill out the questionnaire to over three hundred contacts from institutions covering the UK, Europe and America, of which 106 responded. The responses have been analysed in conjunction with the information gained from other sources.

There are a significant number of standards for both discovery and technical metadata. There are also a range of services by which metadata can be recorded and the data stored alongside these data. NERC itself puts a significant amount of effort into storing data and model results and making the metadata available. For example there are seven Data Centres and the Data Catalogue Service (DCS) to search metadata for datasets stored in the NERC data centres.

Whilst there has been a significant amount of time and effort put into standards, the use is variable. There are a number of different standards, which are mainly related to ISO standards, WaterML, GEMINI, MEDIN, climate based standards as well as bespoke standards for data, but there is a lack of formal standards for model metadata. Storage of data and its associated metadata is facilitated via the NERC data centres with a reasonable uptake.

Whilst the standards and approaches for discovery and technical metadata for data are well advanced and, in theory, well used there are a number of issues:

- Recognition of what the user wants rather than what the data manager feels is required.
- Consolidation of discovery metadata schema based on ISO19115
- Recording different file formats and tools to allow ease of transfer from different file formats
- Retrospective capture of metadata for data and models
- Incorporation of time based information into metadata

However for model metadata, the situation is less well advanced. There is no internationally recognised standard for model metadata, and one should be developed to include features such as: model code and version; code guardian contact details; Links to further information (URL to papers, manuals, etc.); details on how to run the models, etc.; spatial extent of the model instance.

Other considerations include: an assessment of data quality and uncertainty needs to be recorded to enable model uncertainty to be quantified and there is the issue of storage of the models themselves. The latter could either be the model code (via standard repositories) or the executable.

These gaps could be filled by a work programme that would consist of the development of a metadata standard for models, a portal for the recording and supply of these metadata, testing this with appropriate user organisations and liaising with international standards organisation to ensure that the development could be recognised. The results of the whole process should be disseminated through as many channels as possible.

1 Introduction

1.1 RATIONALE

This work is a response to the challenge posed by the NERC Environmental Data (NED) call. It is designed to address the issue of ensuring that models and the data that are used to drive them and what they produce are properly recorded and made available. This is particularly important given the amount of investment that organisations make in producing both data and models.

This project sought to investigate providing wider accessibility by scoping out how to undertake the following:

1. Ensure that the data used to create models are recorded and their source known. (Input Data)
2. The models produced are themselves available. (Model Engine/Instance)
3. The results produced by these models can be obtained. (Output Data)

1.2 IMPORTANCE OF METADATA

Data are fed into a model engine and the results produced as data files. The data used to drive process models can come from a variety of organisations, e.g. NERC Centre Surveys. The models used to process these data can be developed in-house or purchased from a software provider. They can also be a collection of linked models; a composition. For this report the term model code is used to describe the algorithm and its encapsulation into a complied code. A model instance is a combination of the model and its input data where it is applied to a particular area.

These models can typically produce a number of data files which can get multiplied if sensitivity analysis or full uncertainty analysis is undertaken. Therefore, methods have to be developed to store how the data used to drive the models, the models themselves and the resulting output files.

This can benefit both single model as well as linked model compositions. The latter can be formed from components, e.g. models that can be found and linked by knowledge of their metadata.

The meta-model NED project aims to scope out how to produce a metadata catalogue which can store the information to solve these problems.

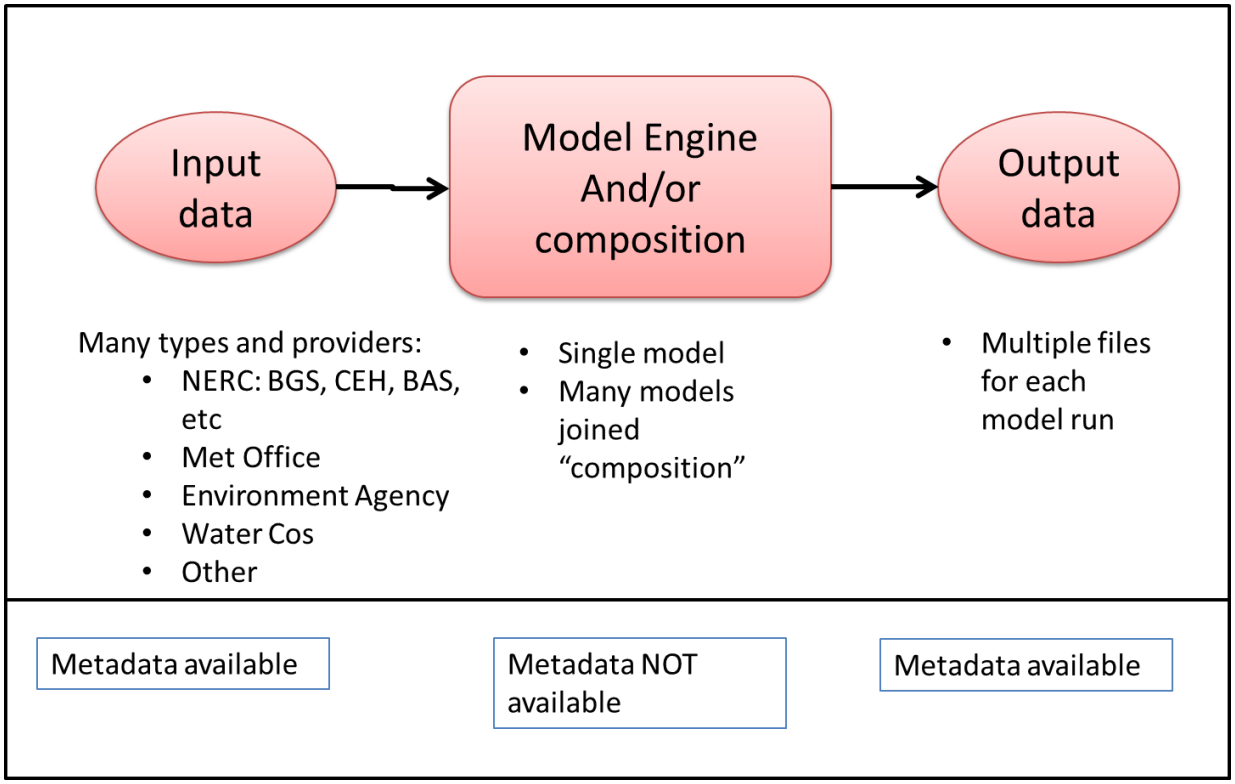


Figure 1. Data flow into and out of a model.

1.3 STRUCTURE OF REPORT

The report has four further sections: Section 2 describes what was undertaken during the project, Section 3 details what was discovered and what is missing in current practice, Section 4 provides a summary of best practice and Section 5 summarises the main conclusions of the report and outlines what we should do.

2 Methodology

2.1 ON-LINE QUESTIONNAIRE

In order to capture the views of a wide spectrum of stakeholders on how they are currently managing metadata for integrated modelling and what gaps exist, an on-line survey was constructed using the "Survey Monkey" tool. This questionnaire was structured to understand both how users approach metadata for data sets used in modelling, and also to explore issues relating to metadata for the models themselves. Accordingly the survey was circulated to over three hundred and twenty stakeholders in universities, commercial organisations, other research organisations, in addition to the NERC data centres.

A total of 108 responses were collected over a four week period. The majority of the respondents held senior positions in their organisations giving weight to the findings of the study. In order to facilitate good take up of the questionnaire the number of "mandatory" questions was kept to the minimum, and so respondents were free to "skip" questions as appropriate. Nevertheless the majority of the respondents completed most of the questions, providing a useful set of data on which to base conclusions. As another aid to maximising the level of response most of the questions were of a multiple choice format where respondents simply selected an option on screen, but scope was provided for users to also record "free text" responses(for example additional comments or opinions on gaps in provision) and very useful additional information was also captured in this way.

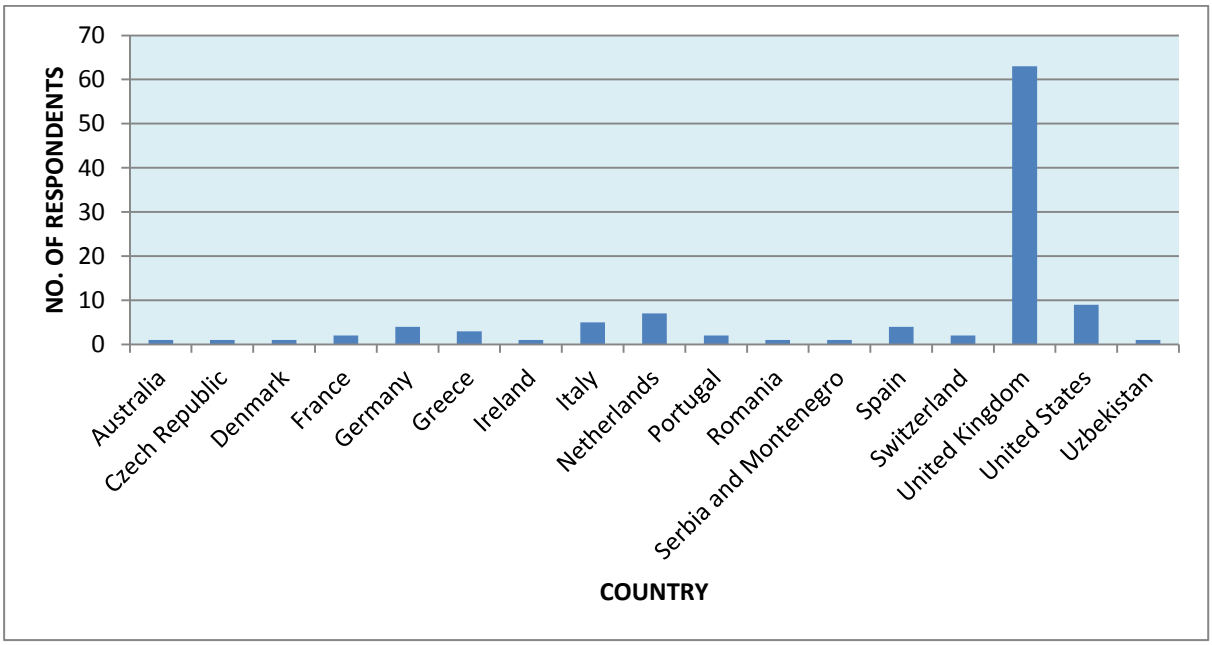


Figure 2. Number of Respondents by Country

The survey was sent out to the extensive contacts networks of BGS and HR Wallingford mainly within the UK but also further afield, and links to the survey were also enabled from relevant websites to maximise take up. The graph in Figure 2 shows that a number of responses were also received from other parts of Europe, as well as the United States and Australia.

In order to better understand differences in metadata requirements between different environmental disciplines respondents were also asked to indicate their primary science discipline. Users were asked to select their discipline from a predefined list. Overall the results indicate that a variety of disciplines are represented including climate change, earth system modelling, ground water and land use modelling (Figure 3). An option was also provided to record disciplines not listed, these also indicate a very wide variation including a number of

individuals involved in IT and systems development to support environmental modelling, CO2 storage and reservoir modelling, as well as a small number of people involved in biodiversity and also catastrophe modelling.

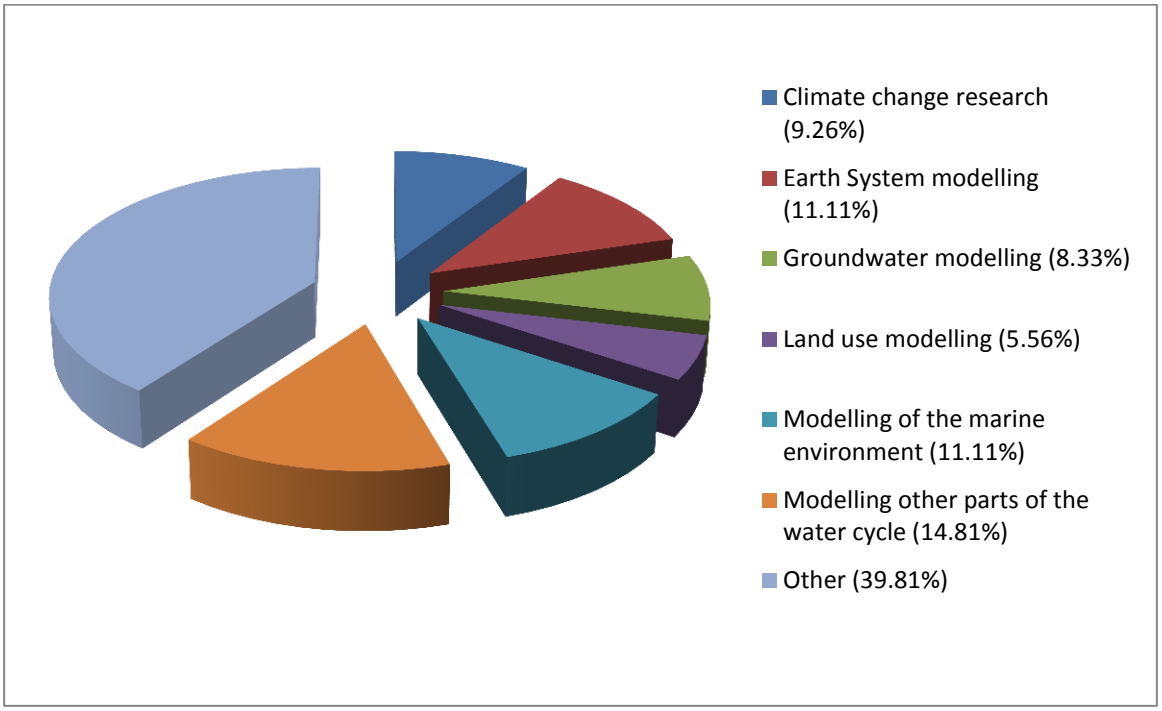


Figure 3. Scientific Disciplines Represented

Respondents were also asked to indicate their organisational roles e.g. data supplier, end user of models, model developer (i.e. Involved in creating model code and systems to support modelling, and those actively involved in the process of integrated environmental modelling. The respondents included a small proportion of data suppliers with the remaining c90% split fairly equally between end user, model developer and modeller (see Figure 4).

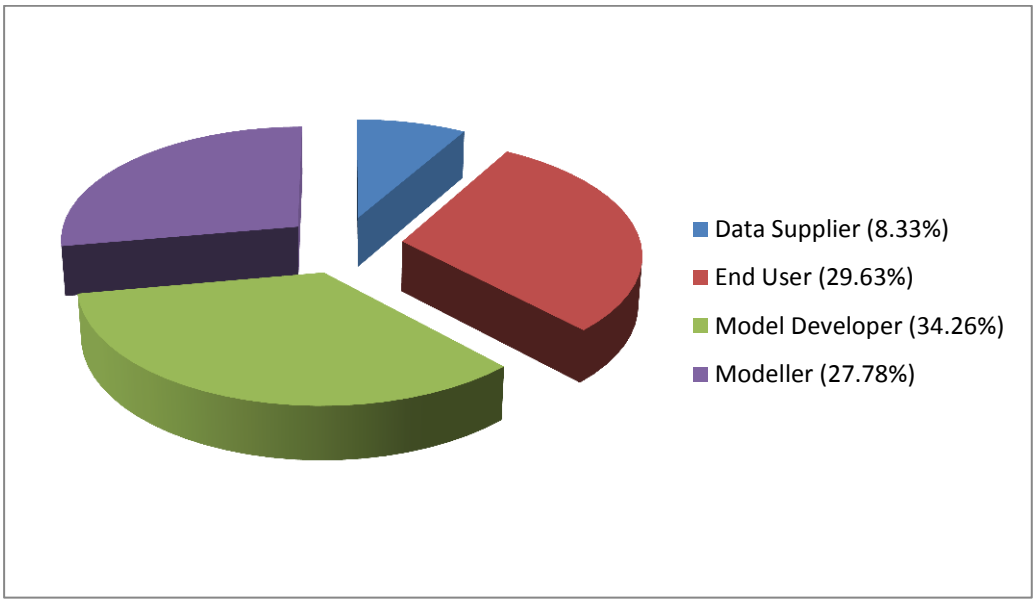


Figure 4. Respondent Roles

2.2 VISITS AND PHONE MEETINGS

2.2.1 Environment Agency

A visit to the Environment Agency HQ in Bristol was undertaken on the 16th July. The meeting was held between BGS staff (Stephanie Bricker, Geraldine Wildman, Andrew Kingdon and Andrew Hughes) and the Environment Agency staff responsible for models (Helen James), Data (Brian Wilson), Data licensing (Paul Hyatt) and Data Sharing (Chris Jarvis). The management of data, the drivers and use of metadata within the Environment Agency were explained.

The main issues presented by the Environmental Agency staff were:

- Legislative drivers are very important – both UK Government and European, e.g. Water Framework Directive and INSPIRE
- Freedom of Information (FOI) enquiries – There are a huge amount so have to reduce them, some 47000 in all at a huge cost in staff time
- Significant amount of datasets (1500 in all) and data flow mapping undertaken on them all

In terms of metadata and data use within the Environment Agency:

- A small proportion of Environment Agency metadata is made available via data.gov.uk The vast majority is held in an internal repository.
- Linked data – Bathing Water Quality collected, analysed and then checked before being made available in via linked data (e.g. a method of publishing data in a defined structure so that it can be interlinked and be used to provide extra services). These data then serve all internal and external requirements.
- All data is managed by a service provider, with spatial data held in Oracle which is distributed as 50 copies to the Environment Agency regions
- Standards are very much used with Defra open data strategy and metadata using ESRI spatial data. Currently investigating ways of dealing with both discovery and technical metadata

2.2.2 NERC Data Centres

The NERC website defines the role of its Data Centres as "It is essential that data generated through NERC supported activities are properly managed to ensure their long-term availability. Our network of data centres provide support and guidance in data management to those funded by NERC, are responsible for the long-term curation of data and provide access to NERC's data holdings. The NERC Data Policy details our commitment to support the long-term management of data and also outlines the roles and responsibilities of all those involved in the collection and management of data."

There are seven NERC data centres, relating to the following subject areas:

1. Atmospheric science
2. Earth sciences
3. Earth observation
4. Marine Science
5. Polar Science
6. Science-based archaeology
7. Terrestrial & freshwater science, Hydrology and Bioinformatics

Representatives of NERC Data Centres contributed to the project whether via the Survey Monkey questionnaire or by direct contact. Of particular interest for this project is the NERC funded "Model Core" project which aims to extend the storage of data to models themselves. The project, reporting to the NERC Science Information Strategy (SIS), is currently investigating the feasibility of a "gold standard" which will:

- Build on the current NERC policy on archiving simulations (BADC Model Data Policy)
- Ensure that rich metadata are available for the model (both discovery and technical)
- Input and output files are in standard formats and have associated Digital Object Identifiers (DOIs)
- Define how to store models, i.e. using a model code repository such as SourceForge, GitHub, etc.
- Provide a way of recording where the models are stored (Register of Code Repositories or RCR)
- Have adequate documentation with which to understand all the elements of the modelling process

2.2.3 Follow up to Survey Monkey Questionnaire

Interviews were conducted with:

Dr. Deborah Hemming, Met Office Hadley Centre,

Prof. Andrew Wade, University of Reading

The purpose of the interviews was to clarify some of the responses made to the questionnaire and potentially gather further useful information from selected individuals who were clearly engaged with the topic.

Dr Hemming mainly works with global and also regional scale climate models, whilst Dr. Wade specialises in biogeochemical and fluid flow modelling. Despite the differences in disciplines covered, both researchers were interested both in the representation of temporal and spatial information in metadata and this further highlights the interest in these areas reflected in the questionnaire results. In both cases there was an interest expressed in temporal resolution – so that a modeller had sufficient metadata to, for example, select data containing the minimum or maximum temperature parameter for a given period e.g. (month, week, day etc). It seems that some of this type of capability may already be incorporated within metadata for climate models, providing a basis for developing a scheme suitable for other environmental disciplines.

The other common theme concerned information on spatial extent. It was clear from the interviews conducted that there is an important need for information to be recorded which allows users to understand the spatial resolution before they proceed further to download the data. This is a particular issue when linking together large scale climate models with data more at a geological scale – for example soil moisture datasets, and is clearly viewed as a key issue to address in developing a metadata scheme.

Both interviews also highlighted the additional "technical metadata" information that others had outlined in the survey including information on the model code and information on how to actually run the model (including time steps and assumptions made).

3 Findings

3.1 SUMMARY OF CURRENT ACTIVITIES

3.1.1 Adoption of metadata standards

Respondents to the on-line questionnaire were asked both about how they managed metadata for the datasets used in modelling and about use of metadata for environmental models themselves. In terms of metadata standards for datasets used in modelling about 30% of respondents (see Figure 5) indicated that they adhered to INSPIRE data specifications. However the number of people who use the ISO metadata standards (ISO 19110, ISO 19115, and ISO 19119) is relatively small – generally 10% or less of those contributing to the survey for each standard, suggesting that these standards (which are adopted by the NERC data centres) are not so commonly used within the wider environmental modelling community. However, there is a significant overlap between INSPIRE and ISO 19115 and this may mask the use of the ISO standard. A further c.30% of respondents preferred to use a variety of other more domain specific standards including the metadata components within WaterML (Part 2), the GEMINI 2.1 standard, the climate and forecast metadata convention, as well as the MEDIN discovery metadata standard. In some cases, particularly for larger organisation (such as the UK Environment Agency) internal metadata schemes are used.

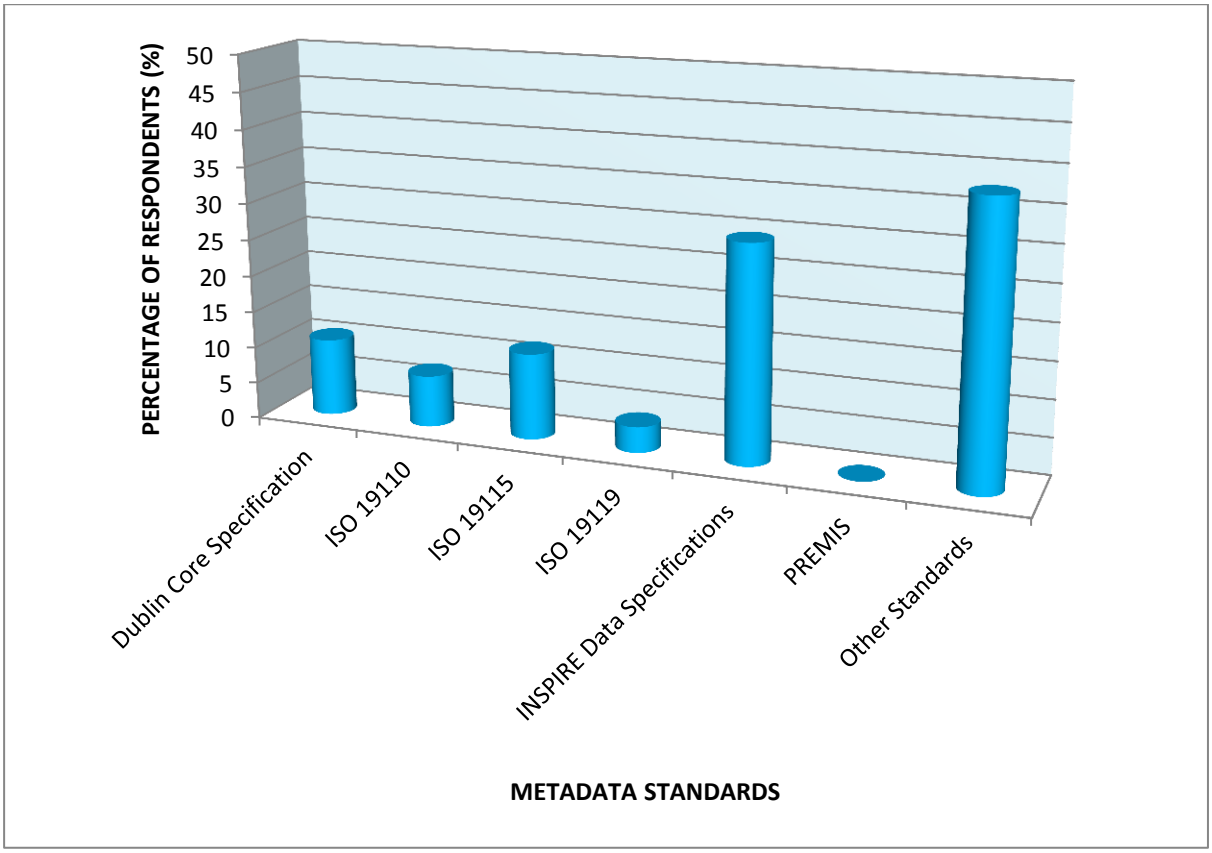


Figure 5. Metadata Standards Applied to Data

Considering metadata standards for models there is a general consensus from the questionnaire confirming our initial view that there is a lack of formal standards for model metadata. Some organisations (e.g. the Environmental Protection Agency in the United States) use their own internal standard. Organisations such as CSDMS in the United States have also proposed a system of describing model metadata.

3.1.2 Using metadata to find and locate data and models

The questionnaire results (Figure 6) indicate that whilst a fair proportion of respondents tend to use metadata catalogues to locate and identify data the most used method of finding data to use in modelling is through organisations which people already collaborate with (36% of respondents) or by approaching known suppliers of specific datasets (c.20% of respondents).

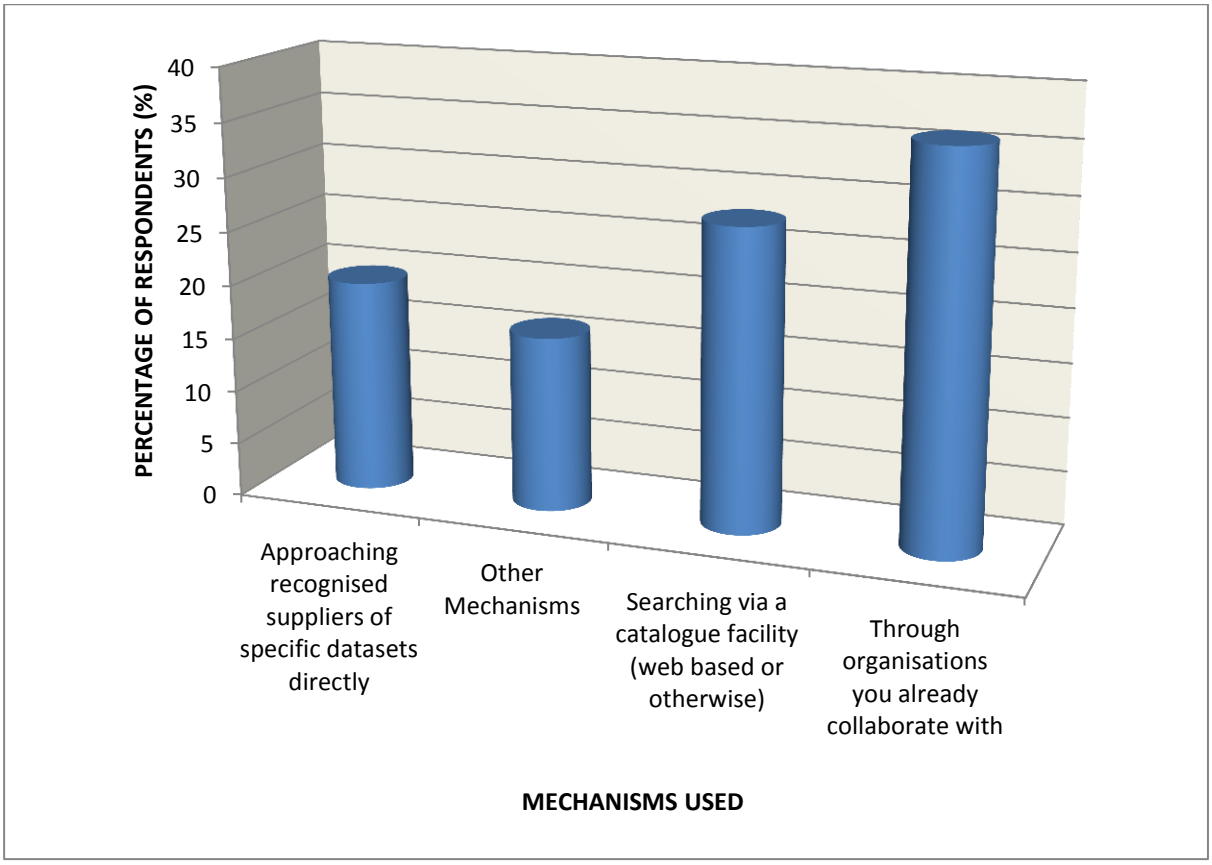


Figure 6. Mechanisms Used to Locate and Identify Data

This clearly implies a lack of take up of on-line metadata catalogues within environmental modelling and a reliance on personal contact or recommendation. There is also a perceived lack of metadata to help people find and locate the datasets they need (Figure 7) with only c20% of respondents indicating that the level of metadata supplied is sufficient. This suggests a number of gaps in provision which are discussed further in section 3.3.

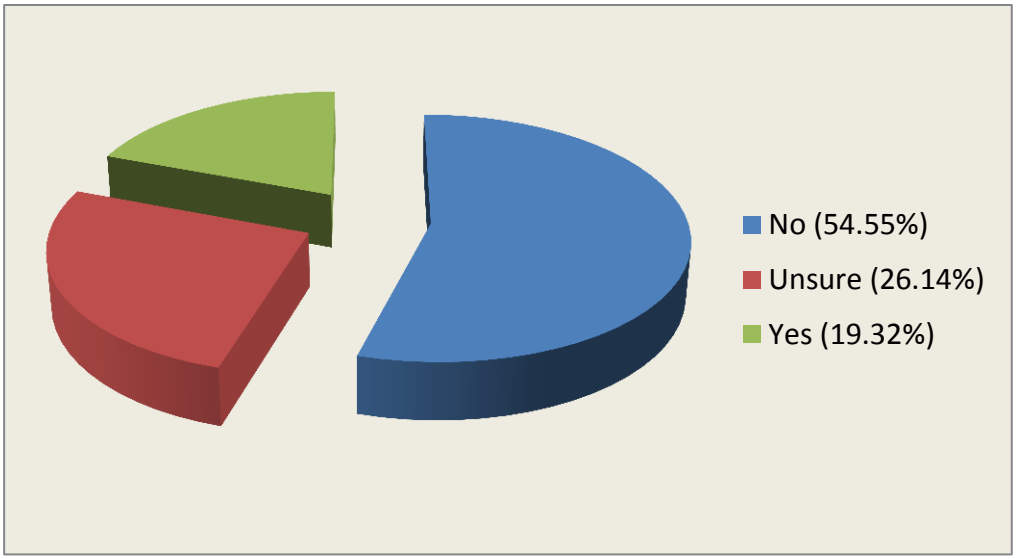


Figure 7. Sufficient metadata supplied with data – is this the case?

When asked whether it was easy to find models produced within other environmental disciplines c.55% of respondents reported that this was a difficult process, with c.28% unsure. However several people highlighted the dangers in using a model developed in another discipline without fully understanding the model. There was also some perception that the atmospheric science community may be better at identifying appropriate models within other disciplines.

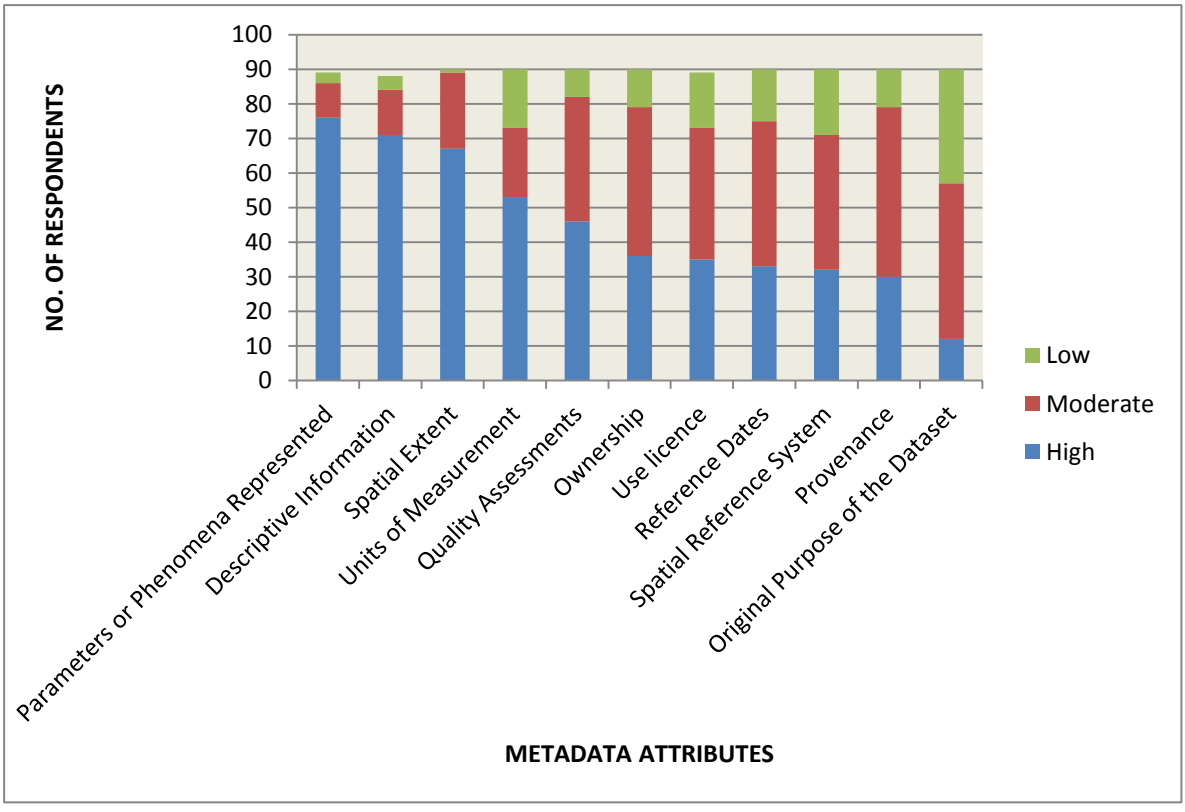


Figure 8. Searching for Data - Relative Importance of Metadata Attributes

Respondents were also asked which metadata attributes are viewed as most important in finding data and models. The results (Figures 8 and 9) indicate that both for datasets and models the most important attributes are descriptive information, the parameters or phenomena involved, and the spatial extent of the dataset or model, followed by quality assessments. For models the type of model (e.g. whether deterministic, probabilistic etc.) and the technical platform are viewed as fairly important metadata attributes (over 30 respondents rate these as of high

importance). The programming languages used and the typical runtime are seen as relatively less important for models. Otherwise the trends for finding data and models are relatively similar, although recording the original purpose of the activity is viewed as more important for models than for datasets. An interesting feature of both datasets and models is that items such as reference dates and provenance (the history of the model or dataset) fall lower in relative importance compared to other attributes. These are "typical" metadata attributes which tend to be advocated by data management specialists, and there is a general view expressed that there should be more emphasis on inclusion of metadata items of interest to the end user.

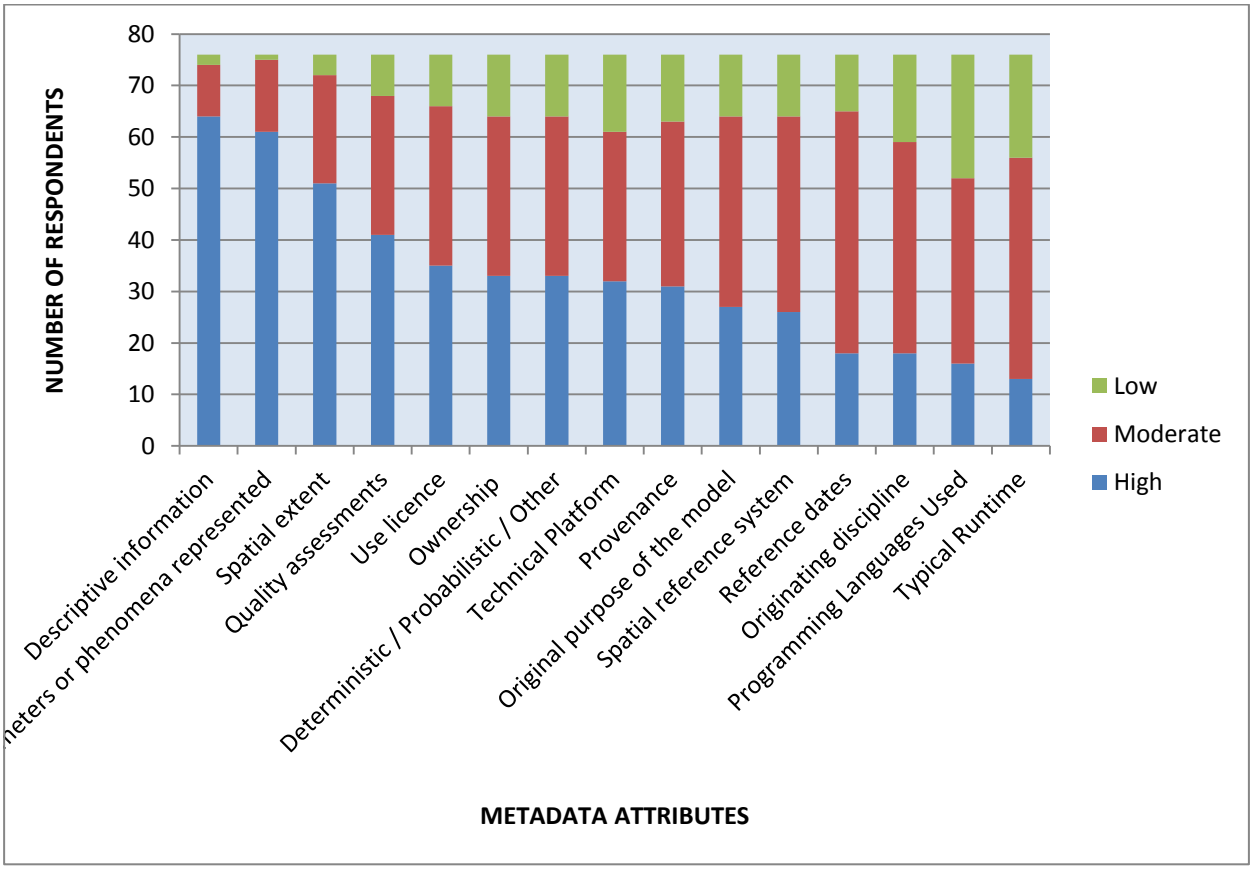


Figure 9. Searching for Models - Relative Importance of Metadata Attributes

Other metadata attributes which respondents wanted to record included the temporal resolution, including the date and time of measurements within the dataset, and the time period to which measurements relate (e.g. month, week, day, minute etc.). There is also interest in recording any associated and derived datasets. For models additional metadata attributes desired included an indication of ease of use, to avoid spending an inordinate amount of time configuring an unfamiliar model. Although the programming languages used was ranked fairly low in relative importance overall, several respondents indicated that to know if the source code for the model was available was an important factor, particularly for developing compositions of linked models. The minimal data requirements were also regarded as an important element to include in metadata for models.

3.1.3 The role of metadata in making use of data and models

The majority of questionnaire respondents who supply metadata (c.40%) indicate that their primary reason for providing metadata is to assist others in using the dataset or model (Figure 10), and this seems to be a more important driver than providing access.

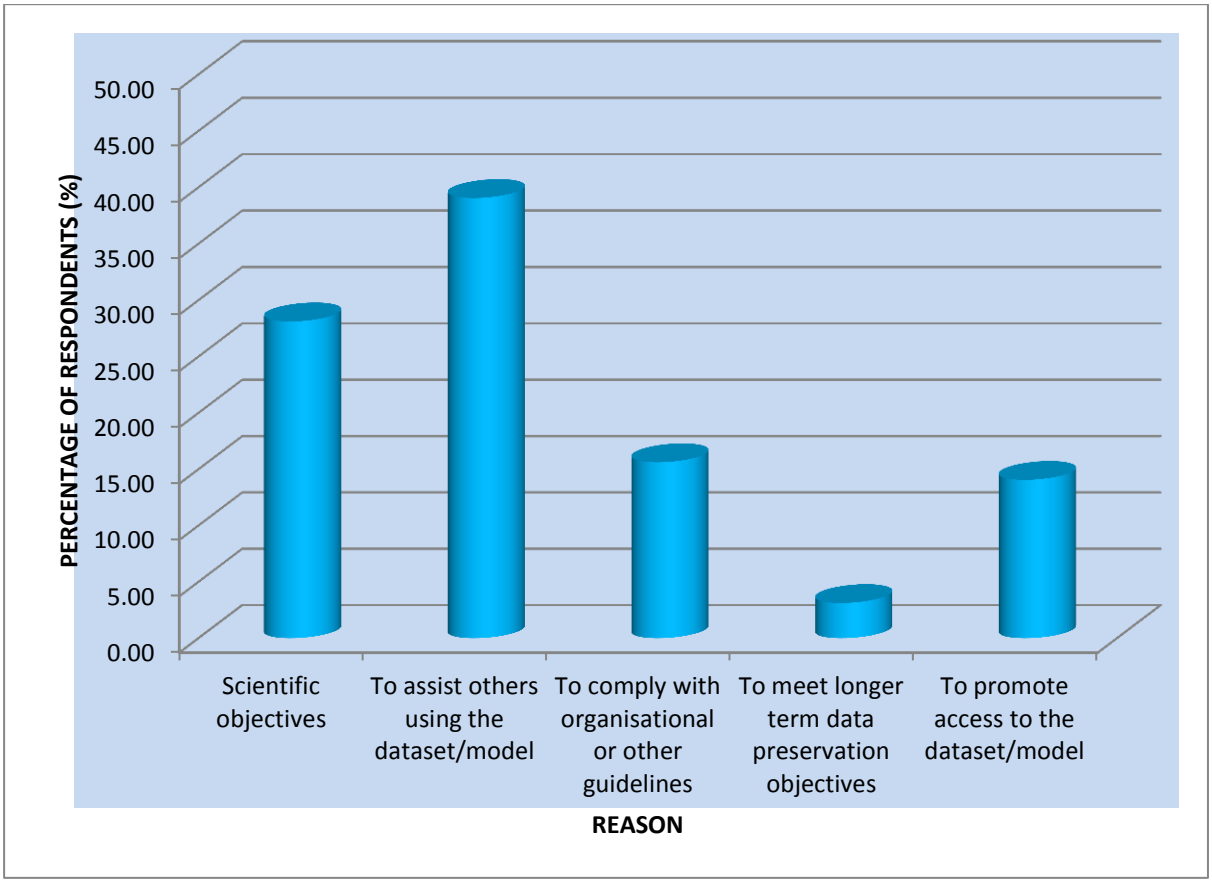


Figure 10. Primary Reason for Providing Metadata

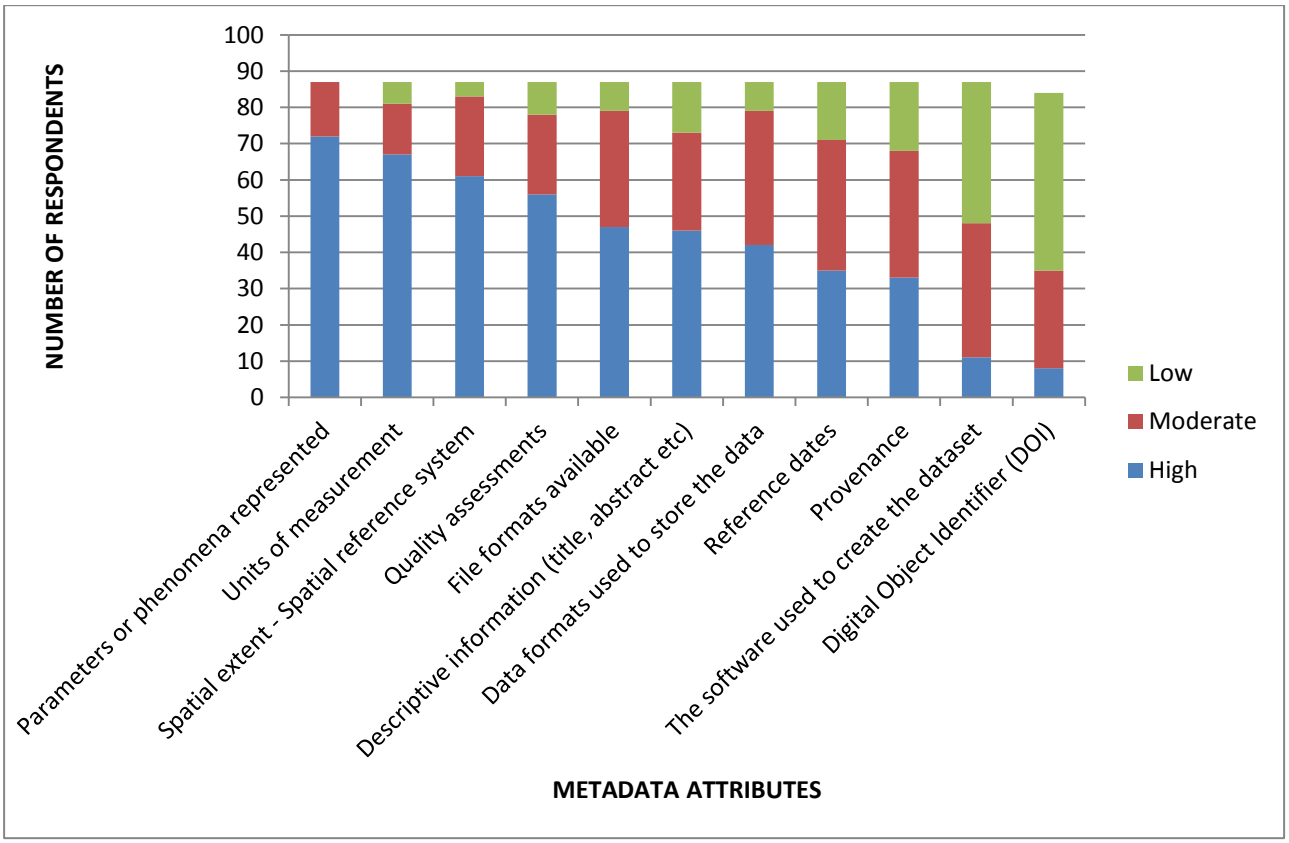


Figure 11. Making use of Data - Relative Importance of Metadata Attributes

For users aiming to making use of data (using descriptive and technical metadata) the parameters represented and units of measurement, together with spatial details and quality assessments are viewed as the most important metadata attributes (Figure 11) with about 60 respondents ranking these as important. Data and file formats are also considered to be reasonably important. Similar to the metadata attributes for finding and locating data reference dates (date created etc.) and provenance are also regarded as relatively less important in the ranking. The inclusion of a digital object identifier (DOI) within the metadata schema is regarded as relatively unimportant, and this is an interesting trend considering the increasing interest in using DOI's to uniquely identify datasets within data management generally.

In the case of models (Figure 12) the metadata attributes perceived as most important were:

- information on the datasets used as inputs
- details of the parameters represented in the data
- the assumptions made in building the model, and
- information on the models used as inputs.

Approximately 60 respondents ranked these four attributes as important. A total of 35 respondents ranked information on the details of the software or model code as of high importance, there was also an indication from the free text comments that this is important information to have particularly for customising code when linking models together. Most of the remaining attributes on the right hand side of Figure 12, including file formats available for input and output and compatible model coupling technologies fall further down the order of relative importance for models with only 20% of respondents regarding these as of high importance to record in metadata (though generally a good proportion of respondents do regard these attributes as of at least moderate importance). Again information on input and output file formats and coupling technologies would seem to be quite important to know about when selecting models for coupling together in a composition, and the relative importance of these attributes would be expected to increase in linked modelling scenarios.

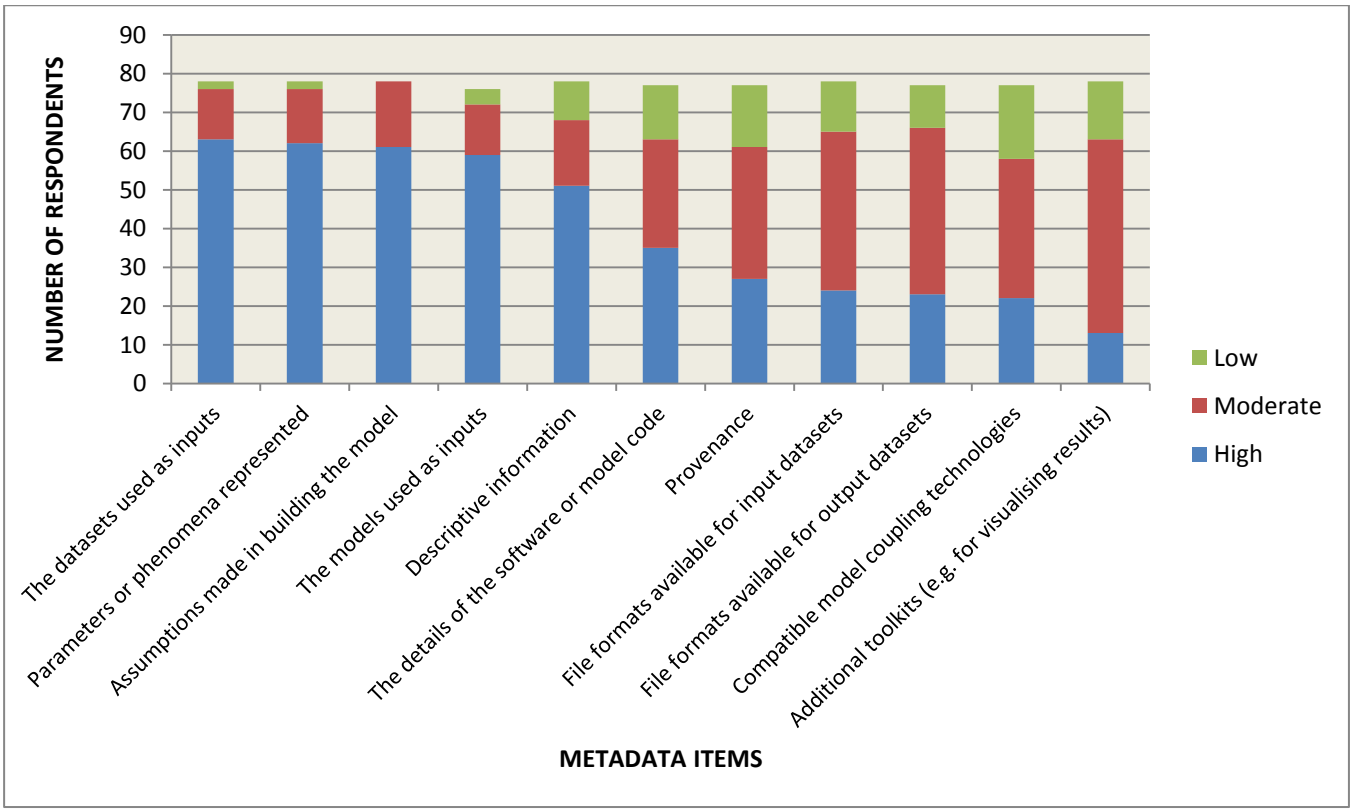


Figure 12. Making use of Models -Relative Importance of Metadata Attributes

Other attributes recommended for inclusion in metadata for data include temporal descriptors and resolution, and also an improved means of describing units. Although provenance was ranked as relatively less important overall there was some interest in being able to access information on the history of use (e.g. what the model had been used for and whether it had met previous requirements). Availability of documentation on the model (for example possibly a link to documentation) was also mentioned several times as being a required metadata attribute.

3.2 BEST PRACTICE

The questionnaire results highlight a number of current trends in best practice concerning how metadata data is used within environmental modelling.

3.2.1 Metadata Standards in Environmental Modelling

The questionnaire results suggest relatively low levels of adoption of the ISO metadata standards (e.g. ISO 19115) which are in general use by NERC data centres for discovery metadata. However, domain specific standards tend to be more commonly adopted for example the metadata elements within Water ML 2.0 Part 1, and the climate and forecast metadata convention applicable to climate modelling.

At the same time when asked about the most important attributes to assist discovering and using data and models, many of the attributes commonly found in for example the ISO19115 schema are regarded as important elements of schemes for discovery and descriptive or technical metadata. This suggests that the ISO 19115 schema for discovery metadata (possibly with appropriate extensions) may provide a good basis for developing a readily adoptable discovery metadata scheme to support environmental modelling.

A number of researchers suggest that there is an over emphasis on spatial metadata attributes in the ISO metadata schemes and that more information, particularly on attributes such as temporal resolution, and the units in which parameters are expressed, should be included.

3.2.2 Best practice issues relating to discovering and accessing data and models

The overall impression is that a metadata schema to support environmental modelling must be easy both to populate and to obtain access to for search and discovery purposes. Such a scheme should easily support the minimum requirements of various environmental disciplines. It is evident that such a schema is not available at the moment but that there are strong drivers within the modelling and IT community to create such a schema (see further discussion in section 3.3)

There is clearly a strong interest in users being able to access the data they need in a format which is useful for them, even if they have to convert from one format to another. There is therefore a need for metadata profiles to include file format information.

The need to be able to capture metadata retrospectively from legacy projects has been mentioned by a number of respondents. There is an indication that this may be less of a problem with NERC funded projects over recent years because of NERC's metadata requirements for submitting data.

3.2.3 Issues relating to model usage

The metadata provided for each dataset should include some documentation on how to use the model. This could for example be in the form of a URL link to appropriate documentation. Related to this the possibility of recording a dataset owner or expert user was also highlighted in the questionnaire, this would provide a means of obtaining advice on the appropriateness of the dataset for various purposes. Technical information on how to configure the model for use is particularly desirable. As one respondent remarked:

"there is little point in being able to access a model and then have to spend several days configuring it to run on your own system."

Other specific information should include whether the source code for a model is available and how to access this, particularly for users wanting to develop compositions.

A number of respondents were interested in indications of data quality being present in the discovery metadata, and also indications of uncertainty. The quality information is particularly valuable when using data or models of course and should include some estimate of accuracy (for example for data items) and also an indication of any limitations with the model.

Where "real" measured data has been mixed with modelled estimates, for example in an input dataset or model this information should be included in the metadata accompanying the dataset or model.

3.2.4 IPR and policy matters

Although IPR and policy matters relating to environmental metadata were not specifically examined in the questionnaire, a number of comments on this area were offered, and have a bearing on the development of metadata systems. It was a widely held view by respondents that data provided by academics or public bodies should be available without cost, the view was expressed that tax payers have already paid for the capture or production of that data and therefore should not have to pay again. There is also a need to encourage more data to be made available in the public domain, an issue was noted by some public sector organisations that although they were aware of high quality commercial data they sometimes had to nevertheless use alternative public sector data which were considered inferior, because they could not obtain access to the higher quality commercial data.

3.3 GAPS IN METADATA PROVISION

3.3.1 Discovering data and models

The survey results confirm our initial supposition that there are conspicuously few widely used metadata schemes for models. However many respondents do regard a number of the metadata attributes included in ISO19115 for example as being important and useful both for discovering data and models even though they may not currently use this standard formally. Metadata elements already contained within the ISO schemes included the spatial extent and spatial reference system, which were viewed as critical for determining the spatial resolution of models when for example linking regional or global models with lower resolution models. The responses overall indicate that the definition of a minimum set of required metadata attributes which are applicable across discipline boundaries is a key requirement. It may be that this could be based for example on the ISO19115 schema.

Metadata attributes which were of particular interest to environmental modellers and which go beyond the level of detail provided in the current ISO schemas are described in Table 1. Additional metadata elements suggested within the questionnaire are further summarised in Figure 13.

Table 1. Additional metadata items to assist discovery of data and models

Attribute	Information Required
Data/Model Quality Assessments	For datasets -Including estimates of accuracy and also measurements of uncertainty For models – limitations and assumptions For models – Scientific Pedigree (e.g. peer reviewed publications) For models – Does the model answer the questions it was designed to address
Additional description of temporal parameters	Temporal resolution and scale (e.g. period of time over which measurements have been made, years, months, weeks etc.) Also what statistical information (if any) is available over a given time period More information on dates and times when measurements were made, this is considered more useful than dates when the metadata record was submitted

There was a strong interest in having more metadata about the computing environment and model code when using models (see section 3.3.2) but at the discovery level there was an interest in simply recording whether the model code was available, in order to assist modellers seeking to build linked model compositions.

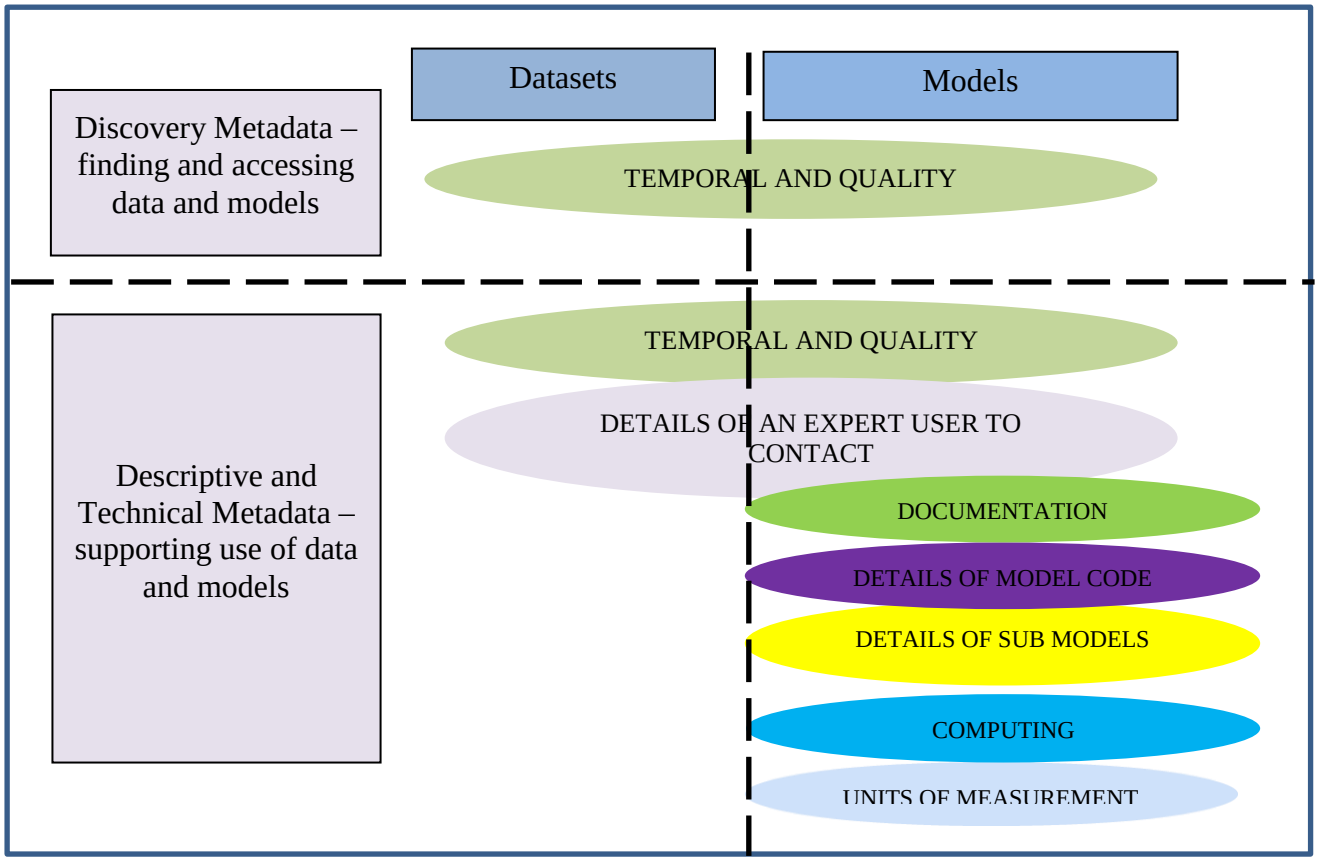


Figure 13. Some additional metadata elements recommended

3.3.2 Supporting the use of data and models

As described above there is a lack of established metadata schemes for models. Some discipline specific schemes are available (as described in Section 3.1.1), and some organisations use their own internally developed schemes. However, as with discovery metadata, there is a clear recognition within the overall environmental modelling community that a usable scheme supporting dataset and model usage that is not constrained by discipline boundaries is required.

The availability of better descriptions of temporal and quality information within the metadata is seen as particularly important when using data for environmental modelling as well as when discovering data. Improved information on the units used is also desirable (e.g. for molecular ratios it is important to state whether the units are Mol/Mol or g/g, or "%").

Another major area which requires metadata development to facilitate effective model use is details of the computing and modelling environment including:-

- Information about the code used to create the model
- Information on the computing environment used
- Which sub models were used in a linked ensemble
- Documentation on how to use the model (e.g. what assumptions were made, and any limitations on its intended usage
- Information on the required input and output data
- Information for input and output data should cover all data types (e.g. constants, parameters and variables) and how their variation over time and space is recorded.

Additional metadata elements desired in a metadata scheme to support environmental modelling are further summarised in Figure 13.

An additional recommendation from the questionnaire was that each dataset is assigned an owner or expert user who can be contacted for further information on the dataset if required. This is actually already a component of NERC's own data management policy and could be extended to a wider metadata scheme.

3.3.3 Additional requirements arising from environmental modelling workflows

There is a common trend in the development of e-infrastructures for environmental sciences within Europe and beyond for users to rationalise the number of web portals for access to models and data, for example to create portals that federates together other existing catalogues. This aspiration is reflected in a number of our questionnaire responses.

In addition to providing better access to metadata to enable other researchers to locate and use them, there is also a perceived need for better systems for model developers and dataset providers to supply the metadata in the first place. These could include for example improved methods for automatically extracting certain metadata, or integrating metadata collection more with the modelling process, to reduce the time/resource impact on the modeller. Some of this information is recorded as part of the modelling workflow, but it often resides in reports and is not systematically made available for model discovery and access, and so mechanisms to make this information more widely available are needed.

The questionnaire results also imply a general lack of availability of software tools to create or access metadata. Tools that are used include Arc GIS which has its own tools for managing spatial metadata. NERC research centres (particularly BGS and CEH) provide research centre catalogues and contribute to the NERC data catalogue. One solution could be for easily available open source tools to access metadata. There is also an interest in improved tools to readily select data on geographic criteria and in time slices, which can export the selected data ready for use. The NERC Centre for Ecology and Hydrology (CEH) have developed internal systems for this, and further development of such technologies will rely on the availability of suitable metadata, particularly including appropriate temporal information.

There is also a view expressed that as a long term aim a metadata standard for modelling should contain the information to build, either manually or automatically, a composition or series of linked models, and should contain sufficient information to detect errors in such a composition.

The requirement for various types of semantic support within a metadata system, for example across different disciplines, or between different countries and languages was also highlighted by a number of people.

3.3.4 Breaking down barriers to more integrated cross discipline modelling

Over 50% of respondents reported that it was not easy to locate models produced in other environmental disciplines, with a further 28% being unsure how easy this was, demonstrating a clear need for better systems to locate models. Lack of searchable catalogues and common ways to describe models were also viewed as important barriers to making models more widely available (Figure 14).

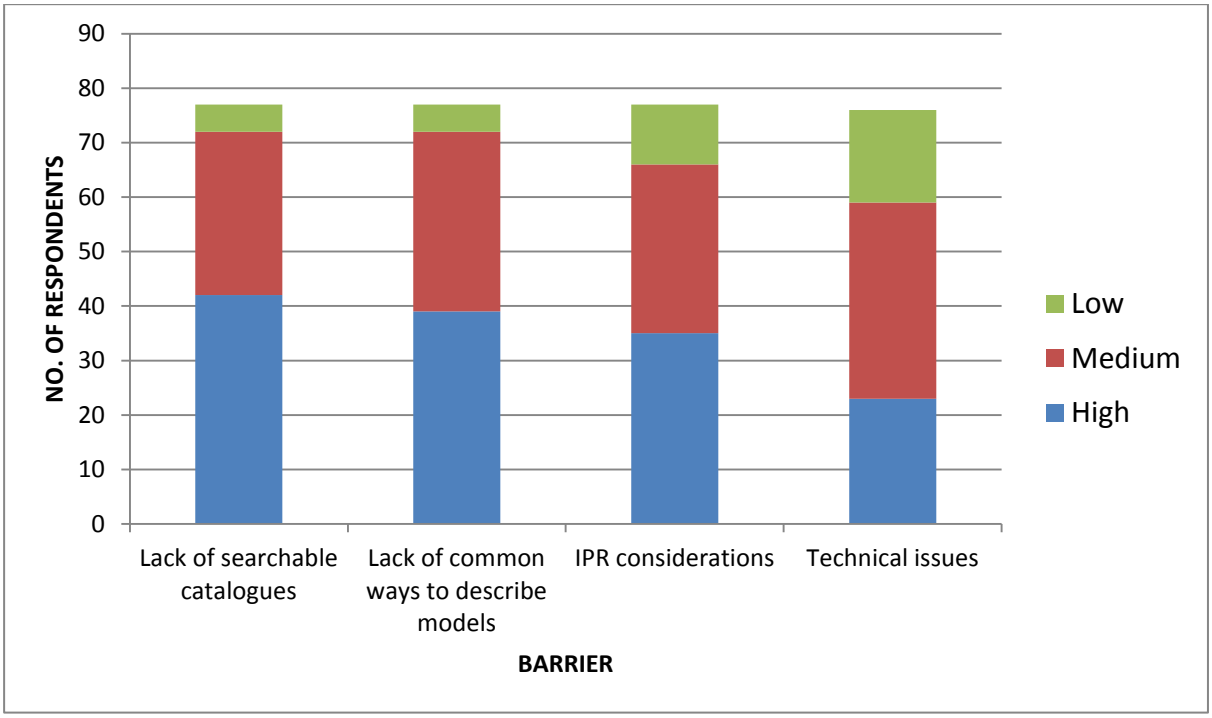


Figure 14. Barriers to the wider availability of models

3.3.5 Summary of essential gaps to be addressed

The key gaps in provision identified are summarised in Figure 13 and include:

- Metadata elements describing the temporal information available in datasets and models both for discovery and use of data and models
- More information needs to be provided on data and (particularly) on model quality issues, to assist users in selecting models which are suitable for their purposes
- With regard to metadata for researchers using models there is a definite need for more "technical metadata" information. A number of required elements have been suggested in the questionnaire and clearly to some extent reflect individual preference. But the key emphasis is on information on how to configure and use the model. A requirement for the metadata simply to contain a link to existing information about the model, whether in a user manual or research paper etc. was a fairly common requirement
- In terms of being able to find out what other models are available there is a clear lack of suitable metadata catalogues (presumably because the metadata itself is not available)
- Users are not aware of suitable software tools to capture metadata within their domains, and there is an indication that such tools need to be developed. Clearly for a metadata scheme to support environmental modelling to work then users need to be able to enter and supply their metadata easily.

4 Summary of Best Practice

4.1 CURRENT METADATA STANDARDS

Since datasets form the boundary condition inputs and resultant outputs of modelling studies, the authors consider that a study of metadata for models should also include that of the metadata for the supporting datasets. Indeed, it is expected that the two will be very similar and derived from the same base standards.

We begin by looking at the metadata elements associated with various forms of typical environmental monitoring data divided into categories based around the geospatial structure of the dataset. This allows modelling data to be included easily alongside that of its measured equivalents. Early versions of the Climate Science Modelling Language (CSML) identified 13 geospatial data feature types for describing measured and modelled datasets. CSML version 3 offers 10 (OGC, 2011), with an additional 'observation' feature type. Version 3 feature types are specialisations of the O&M model (ISO19156) with the exception of 'observation' which is a direct usage. The 10 feature types are as follows (OGC, 2011):

- Point – A single observation at a point e.g. a single raingauge measurement.
- PointSeries – A time series of single datum observations at a fixed location e.g. a stream of measurements of a single parameter from a tide gauge, buoy or weather station.
- Profile – An observation of a parameter along a vertical line in space e.g. a wind sounding or radiosonde.
- ProfileSeries – A time-series of profiles on fixed vertical levels at a fixed location e.g. vertical radar timeseries.
- Grid – Single time-snapshot of a gridded field.
- GridSeries – Time-series of gridded parameter fields e.g. a numerical weather prediction model output.
- Trajectory – An observation along a discrete path in time and space e.g. aerosol measurements along an aircraft's flight path.
- Section – A series of profiles from a trajectory in time and space e.g. marine Conductivity and Temperature Data (CTD) measurements along a ship's track.
- Swath – Two-dimensional grid of data along a satellite ground path. E.g. AVHRR satellite imagery.
- ScanningRadar – Backscatter profiles along a look direction at fixed elevation but rotating in azimuth e.g. a weather radar output.

Data produced according to each of these feature types is typically combined with, either a separate metadata file, or metadata incorporated into the data file itself. This metadata can be divided into an number of categories, some more closely related to information required to find the dataset ('discovery metadata') and some more closely related to information required in order to use the dataset once it has been obtained ('use metadata'). Of course, some information is useful for both locating and using the dataset. Table 2 gives examples of the metadata elements given by three environmental datasets, one set of model results and two from sensors and together they represent six of the ten geospatial feature types given in CSML version 3:

- **NetCDF CF:** Meteorological model results stored as the Climate and Forecasting version of NetCDF (NetCDF CF) as part of the DRIHM project [http://www.unidata.ucar.edu/software/netcdf/].
- **WaterML 2.0:** Data served via a web service from the HydroServer instance at SDSC in San Diego [http://www.opengeospatial.org/standards/waterml].

- **Satellite:** Altimeter data from the Envisat mission stored as part of the GlobWave dataset, stored as NetCDF [http://www.unidata.ucar.edu/software/netcdf/].

Table 2. Metadata elements exhibited in three example environmental datasets

MetaData	NetCDF CF DRIHM Implementation (CSML Grid and GridSeries)	WaterML 2.0 (CSML Point and PointSeries)	Satellite NetCDF Envisat Altimeter Data (CSML Swath)
Ownership and Contact Details	:email, :institution	gmd:organisationName, gmd:pointOfContact, gmd:individualName, gmd:role, gmd:onlineResource, gmd:address, gmd:phone, gmd:electronicMailAddress	:institution, :contact, :processing_center, :source_provider
Title and Abstract	:title, :comment, :filename	gmd:title, gmd:abstract, gmd:citation	:title
Provenance	:projectinfo, :algorithm, :history, :source, :model_name, :model_description,	wml2:generationSystem, om:featureOfInterest, wml2:ObservationProcess, wml2:processing, wml2:processType	:source, :project, :history, :mission_name, :source_name
Reference Dates and Times	:calendar, :time, :time_bounds, :time_date, :datestart, :dateend, :filedate, :julday, :julyear, :GMT, :dt, :units	wml2:generationDate, om:phenomenonTime, om:resultTime, gml:TimePeriod, gml:TimeInstant, gml:timePosition, gml:beginPosition, gml:endPosition, wml2:time, wml2:temporalExtent, wml2:aggregationDuration	:start_date, :stop_date, :calendar
Spatial Extent, Geometry, SRS	:coordinates, :grid_mapping_name, :dx, :dy, :griddim_bottomtop, :griddim_southnorth, :griddim_westeast, :units, :epsg_code, :bounding_box, :inverse_flattening, :semi_major_axis, :longitude_of_prime_meridian, :grid_mapping_name	wml2:samplingFeatureMember, sams:shape, gml:Point, gml:pos srsName, wml2:MonitoringPoint, wml/siteProperty/elevation_m	:comment, :coordinates, :scale_factor, :add_offset
Phenomenon / Parameters	:standard_name, :long_name	om:observedProperty, wml2:parameter, wml2:qualifier, wml2:processReference, wml2:sampledMedium	:altimeter_sensor_name, :radiometer_sensor_name, :long_name, :standard_name, :calibration_formula, :calibration_reference
Units of Measurement	:units	wml2:uom	:units
Technical	:cell_method, :_Netcdf4Dimid, :_FillValue	wml2:interpolationType, wml2:source, wml2:cumulative, name xlink:title="noDataValue", wml2:aggregationDuration	:software_version, :source_software, :source_version, :acq_station_name, :cycle_number, :pass_number, :equator_crossing_time, :equator_crossing_longitude, :product_version, :_FillValue, :flag_masks, :flag_meanings, :valid_min, :valid_max
Quality Measures		wml2:quality	:quality_flag
Licence and IP			

Standards Definitions	:references	gmd:language, gmd:CI_RoleCode, gml:Dictionary, gml:dictionaryEntry	:references

Table 3 compares these common metadata categories, established by looking at reasonably mature implementations of modelled and measured data, with the core metadata for geographic datasets found in ISO19115 [OGC, 2003; Table 4]. This core metadata from ISO19115 constitutes mandatory elements (M) recommended but option elements (O) and elements mandatory under certain conditions (C).

Table 3. Common metadata categories and their representation in core ISO19115

MetaData	ISO19115
Ownership and Contact Details	**Dataset responsible party** (O) (MD_Metadata > MD_DataIdentification.pointOfContact > CI_ResponsibleParty)
Title and Abstract	**Dataset title** (M) (MD_Metadata > MD_DataIdentification.citation > CI_Citation.title) **Dataset topic category** (M) (MD_Metadata > MD_DataIdentification.topicCategory) **Abstract describing the dataset** (M) (MD_Metadata > MD_DataIdentification.abstract)
Provenance	**Lineage** (O) (MD_Metadata > DQ_DataQuality.lineage > LI_Lineage)
Reference Dates and Times	**Dataset reference date** (M) (MD_Metadata > MD_DataIdentification.citation > CI_Citation.date) **Additional extent information for the dataset (temporal)** (O) (MD_Metadata > MD_DataIdentification.extent > EX_Extent > EX_TemporalExtent or EX_VerticalExtent) **Reference system** (O) (MD_Metadata > MD_ReferenceSystem)
Spatial Extent, Geometry, SRS	**Geographic location of the dataset (by four coordinates or by geographic identifier)** (C) (MD_Metadata > MD_DataIdentification.extent > EX_Extent > EX_GeographicExtent > EX_GeographicBoundingBox or EX_GeographicDescription) **Spatial resolution of the dataset** (O) (MD_Metadata > MD_DataIdentification.spatialResolution > MD_Resolution.equivalentScale or MD_Resolution.distance) **Additional extent information for the dataset (vertical)** (O) (MD_Metadata > MD_DataIdentification.extent > EX_Extent > EX_TemporalExtent or EX_VerticalExtent) **Spatial representation type** (O) (MD_Metadata > MD_DataIdentification.spatialRepresentationType) **Reference system** (O) (MD_Metadata > MD_ReferenceSystem)
Phenomenon / Parameters	
Units of Measurement	
Technical	**Dataset character set** (C) (MD_Metadata > MD_DataIdentification.characterSet) **Distribution format** (O)

	(MD_Metadata > MD_Distribution > MD_Format.name and MD_Format.version) **On-line resource** (O) (MD_Metadata > MD_Distribution > MD_DigitalTransferOption.onLine > CI_OnlineResource) **Metadata file identifier** (O) (MD_Metadata.fileIdentifier) **Dataset language** (M) (MD_Metadata > MD_DataIdentification.language)
Quality Measures	
Licence and IP	
Standards Definitions	**Metadata standard name** (O) (MD_Metadata.metadataStandardName) **Metadata standard version** (O) (MD_Metadata.metadataStandardVersion) **Metadata language** (C) (MD_Metadata.language) **Metadata character set** (C) (MD_Metadata.characterSet) **Metadata point of contact** (M) (MD_Metadata.contact > CI_ResponsibleParty) **Metadata date stamp** (M) (MD_Metadata.dateStamp)

It can be seen that all but four of the common metadata categories are covered by what is considered 'core' ISO19115: Phenomenon / Parameters, Units of Measurement, Quality Measures, Licence and IP. Quality Measures are covered by the optional element DQ_DataQuality and Licence and IP through the optional constraints element MD_Constraints.

Handling physical, chemical or biological parameters and their units of measurement is best achieved through the use of phenomenon and unit dictionaries such as climate and forecasting (CF) standard names [OGC, 2011] (see Table 4) or the BODC parameter code units definition [http://www.bodc.ac.uk/] (see Table 5).

Table 4. CF Standard Names Entry

Entry ID	Canonical Units	Description
wave_frequency	s-1	Frequency is the number of oscillations of a wave per unit time

Table 5. BODC Parameter Code Units Definition Entry

BODC Parameter Code	Units	Definition	Minimum Permissible Value	Maximum Permissible Value	Absent Data Value
CTMPZZ01	Degrees Celsius	Temperature of the atmosphere	-100	60	-999

4.2 CURRENT USAGE

There are a significant amount of tools, initiatives and technologies related to metadata. A literature search is summarised in Appendix 2 under the following headings:

- Metadata tools

- Repository technologies
- Storage technologies
- Data preservation technologies – summary and main trends
- Data discovery and access
- Technologies and frameworks for processing data

The main findings from this search include:

- Availability of tools for producing metadata in XML format and the use of the OpenSource GeoNetwork which is used within the FluidEarth Model Catalogue to display the location of model instances
- Availability of system to store digital objects such as FEDORA
- NERC has put significant resources into a data store under the JASMIN project
- Use of Open Geospatial Consortium standards to define catalogue services
- There are a significant number of portals including the INSPIRE (see below) Geoportal to display spatial data
- Once models become more readily available then there is the potential to be linked using frameworks such as ESMF and standards such as OpenMI

4.3 INSPIRE

The INSPIRE Directive aims at creating an infrastructure for geographical information interoperability in Europe. In this context data holders should publish their geographic datasets through a range of Network Services. INSPIRE Transformation services provide a means to transform a given dataset through the invoking of a service implementing a standardized procedure on a remote machine. Typical examples of transformation services are the schema transformation which transforms the structure of the input dataset and the Coordinate Reference System (CRS) transformation which can be used to bring together datasets based on different CRS.

4.4 NERC INITIATIVES

4.4.1 Data Catalogue Service (DCS)

The DCS (see http://data-search.nerc.ac.uk/) allows you to search a catalogue of metadata (information describing data) to discover and gain access to NERC's data holdings and information products. The metadata are prepared to a common NERC Metadata Standard and are provided to the catalogue by the NERC Data Centres.

Data Providers create metadata documents describing data resources. These are published by each data provider to make them available for others to access. An automatic process gathers or harvests these documents from each data provider, and ingests them into a database where they are stored alongside those from other data providers. Data providers have control over their publishing tool via the Data Providers Admin Interface. A web service carries out searches of this database in response to search requests received from a search interface, possibly hosted by a third party as part of a web portal. The web service returns results back to the search interface, for presentation by the search interface to display to the user. Search tools included in the search interface help the user construct search requests based on time periods, geographic areas and text terms from controlled vocabularies, provided by a vocab server.

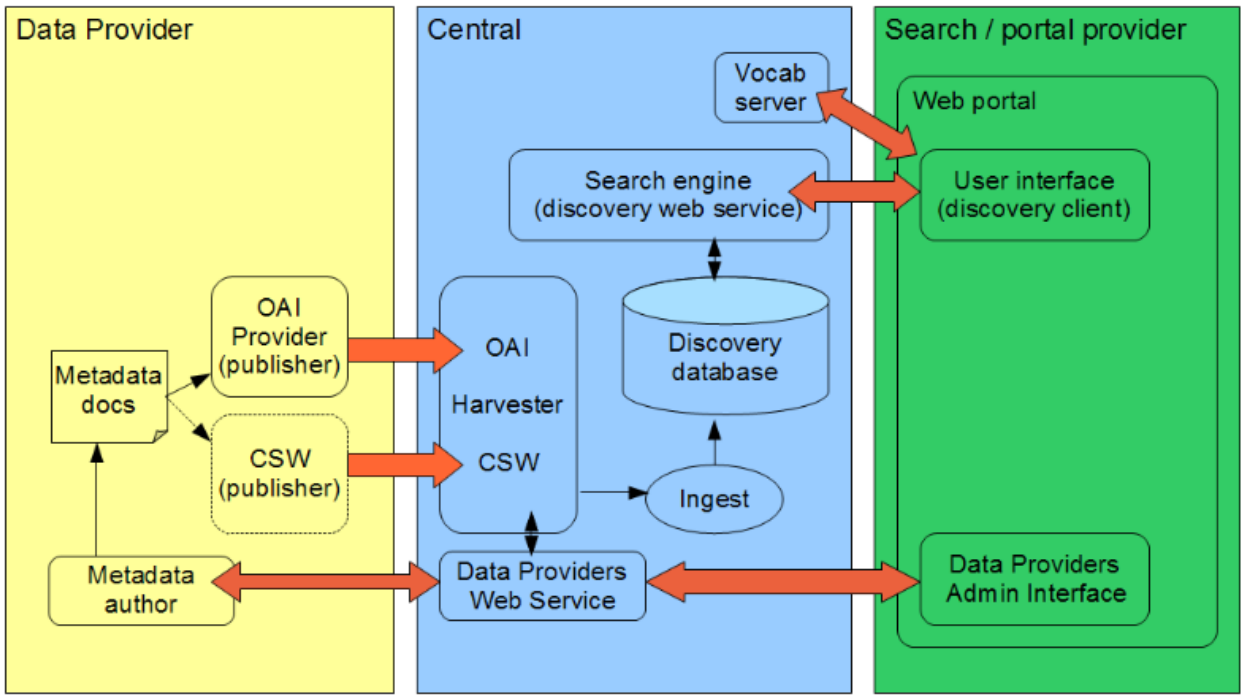


Figure 15. Components of the NERC data discovery service

4.4.2 Lowland Catchment Research (LOCAR) Data Management

The LOCAR data Centre was set up to manage the scientific data produced by the NERC LOCAR Thematic Programme, which finished in March 2006. The aim of the Data Centre was to create an integrated, quality controlled, quality assured database readily accessible to LOCAR scientists, and to the wider scientific community.

To create the database the Data Centre was responsible for specifying procedures, formats and media in which data will be received from the field and disseminated to users, setting up a data management policy, and ensuring that data were held securely. The Data Centre actively sought out existing NERC and third party datasets, and was responsible for disseminating field data as it become available, and for storage and dissemination of the datasets created by LOCAR researchers.

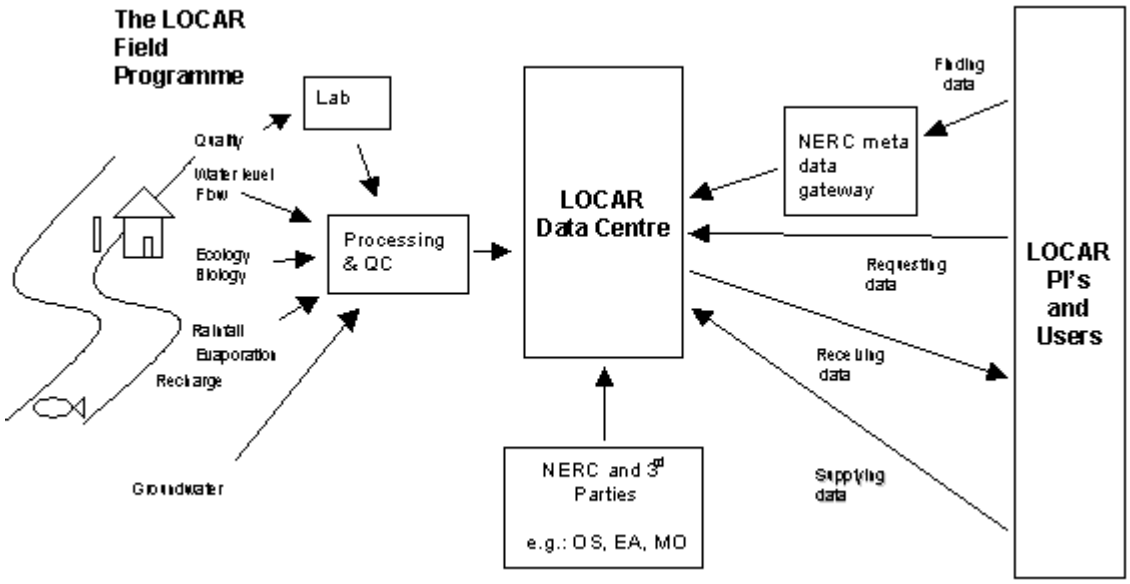


Figure 16. Flow of data for the LOCAR Data Centre

4.4.3 Earth Science Academic Archive

The Earth Science Academic Archive has been set up as part of the National Geoscience Data Centre (NGDC) to deposit the results of the relevant Earth Science research. The ESAA accepts results from NERC research and any other similar research projects to ensure their long term safe keeping and future use.

The Earth Science Academic Archive is responsible for:

- liaising with principal investigators and other NERC grant holders to ensure that appropriate data are offered to the NGDC
- selection of data for inclusion in the NGDC in liaison with BGS scientists and other stakeholders
- long-term curation and preservation of analogue and digital data (including samples)
- publicising the holdings and making available information on the web

Examples of types of data submitted to the ESAA:

- research reports
- photographs
- spreadsheet data
- figures and diagrams
- 3D models

It is essential that all data gatherers/generators provide appropriate metadata to their Data Centre, in line with current metadata standards, such as the "working standards" provided by Holmes et al, 1999. These "working standards" are in turn derived from more comprehensive National Geospatial framework Archive and ISO standards.

5 Summary of findings and proposed work

5.1 SUMMARY OF FINDINGS

The following summarises the main findings of the work, the gaps that have been identified and how they could be tackled.

5.1.1 What does exist?

There are a significant number of standards for both discovery and technical metadata. There are also a range of services by which metadata can be recorded and the data stored alongside these data. NERC itself puts a significant amount of effort into storing data and model results and making the metadata available. For example there are seven Data Centres and the Data Catalogue Service (DCS) to search metadata for datasets stored in the NERC data centres.

5.1.2 What is used?

Whilst there has been a significant amount of time and effort put into standards, the use is variable. There are a number of different standards, which are mainly related to ISO standards, WaterML GEMINI and MEDIN, climate based standards as well as bespoke standards for data, but there is a lack of formal standards for model metadata. Storage of data and its associated metadata is facilitated via the NERC data centres with a reasonable uptake.

5.1.3 What gaps are there?

Whilst the standards and approaches for discovery and technical metadata for data are well advanced and, in theory, well used there are a number of issues:

- Recognition of what the user wants rather than what the data manager feels is required.
- Consolidation of discovery metadata schema based on ISO19115
- Recording different file formats and tools to allow ease of transfer from different file formats
- Retrospective capture of metadata for data and models
- Incorporation of time based information into metadata

However for model metadata, the situation is less well advanced. There is no internationally recognised standard for model metadata, which should include, but not necessarily be limited to:

- Model code and version
- Code Guardian who they are and contact details
- Links to further information (URL to papers, manuals, etc),
- details on how to run the models, etc.
- Spatial extent of the model instance
- Information on mixing data of different types (observed and modelled).

Other considerations include an assessment of data quality and uncertainty needs to be recorded to enable model uncertainty to be quantified and there is the issue of storage of the models themselves. The latter could either be the model code (via standard repositories) or the executable.

5.1.4 How could they be filled?

To assist the development of successful uptake of the storage and discovery of data and models, the following activities are required:

- Development of a metadata standard for models based on ISO19115
- Creation or extension of a tool to record metadata including for time based data
- Use of NERC's DCS to store and serve discovery metadata for models
- File conversion tools made more readily available
- Provide for storage of models in an accessible form: code and executable

As well as this, there a number of initiatives that could provide tools and techniques to fill these gaps.

5.2 DETAILS OF ACTIVITIES

There is a need for a system that allows the storage and interrogation of metadata for model codes and their instances. A project is envisaged that would build on existing metadata standards (i.e. ISO19115) to provide a suitable standard that could be used in conjunction with existing tools to provide a system to store model metadata. The development of this system should be undertaken in conjunction with suitable project partners, for example the Environment Agency (EA) and water companies. The work should be undertaken in conjunction with the NERC SIS "Model Code" project. Whilst this initiative is using climate models as their example, this could be extended to included models of the terrestrial water environment, i.e. hydrologic and hydraulic models. The likely activities in this project are outlined below.

5.2.1 Activity 1 – Metadata standards for model discovery and use

The initial task will be to determine the metadata standard to be used for data along with the metadata standard to be used for models (as an extension of the standard used for data). It is likely to be related to ISO standards and be INSPIRE compliant, so develop recommendations for extensions to the ISO 19115 discovery metadata standard in prototype form. This could build on the user consultation information gained during this scoping study, for example. Develop a suitable schema for technical/descriptive metadata to support using models (as distinct from discovery) models.

If the ISO standard is used then there is a need to investigate how to progress adding extensions with relevant ISO committees at an early stage.

5.2.2 Activity 2 – Stakeholder workshop and initial engagement

Establish discipline specific and also cross discipline stakeholder focus groups to review the proposed schemes (in a short time scale). These could include for example commercial users and relevant NERC data centre staff. Develop user requirements for capturing and also accessing model metadata (what applications are needed) - in association with stakeholder focus groups.

Liaise with NERC's Model Code project via the Science and Information Strategy Board.

5.2.3 Activity 3 – Investigate Feasibility of Approach

Identify test-bed models/datasets, to test proposed scheme and applications. Develop a prototype application to allow searching selected models with input from NERC data centres, such as NGDC and EIDC. Develop tools to aid metadata capture - based on user requirements - maybe focus on a couple of common modelling environments.

5.2.4 Activity 4 – Cataloguing technology for model metadata

Building on existing technology such as the FluidEarth Model Catalogue and NERC's DCS, a model catalogue will be developed. This will use a mix of input form and map based searches to enable the user to find model codes and their instances.

Alongside this, finalise metadata scheme and liaise with ISO committees to get standard extended/adopted.

5.2.5 Activity 5 – Investigate Commercial Feasibility

Application testing and release of appropriate model metadata capture applications. Create demonstration examples of how these datasets can be used to create commercial products, either by reprocessing and adding value, or by running further models against the data.

5.2.6 Activity 6 – Dissemination and stakeholder feedback

Dissemination of activities will be by the project partners usual channels e.g. FluidEarth, OpenMI, OGC, BGS webpages, EA standard processes etc. This will be supplemented by using the stakeholder group developed in activity 2. Alongside this then one or two showcase examples of the data in action will be created and used as exemplars to show the utility of the approach. To compliment this, a set of quotations from opinion formers will be used to build the impact case.

Set up user group to assist project direction which will include: Project staff, Environment Agency, water companies such as Thames Water, representative of NERC's Model Code project. This will help steer the project.

5.2.7 Activity 7 – Project management

The project will be managed using the internal procedures of each organisation, for example PRINCE2. A project lead will be identified with overall responsibility for delivery and who will liaise with the funding body and ensure proper communication with the project partners.

References

Holmes, KA, Dobinson, A, Giles, J RA, Johnson, C C, Lawrence, D J D and McInnes, J L, 1999. BGSgeoIDS Metadata - Issues Document. British Geological SurveyTechnical Report, WO/99/01R.

ISO19115, 2003. Geographic information — Metadata. International Standards Organisation, ref: ISO 19115:2003(E).

OGC, 2011. Climate Science Modelling Language (CSML): Sampling Coverage Observations for the met/ocean domain, Open Geospatial Organisation, OGC 11-021.

Appendix 1 Data obtained from on-line questionnaire

Section One – Background information Country, science discipline etc

Country Affiliation

COUNTRY	NO. RESPONDENTS
Australia	1
Czech Republic	1
Denmark	1
France	2
Germany	4
Greece	3
Ireland	1
Italy	5
Netherlands	7
Portugal	2
Romania	1
Serbia and Montenegro	1
Spain	4
Switzerland	2
United Kingdom	63
United States	9
Uzbekistan	1
TOTAL	**108**

Science Discipline

DISCIPLINE	NO. OF RESPONDENTS	% OF RESPONDENTS
Climate change research	10	9.26
Earth System modelling	12	11.11
Groundwater modelling	9	8.33
Land use modelling	6	5.56
Modelling of the marine environment	12	11.11
Modelling other parts of the water cycle	16	14.81
Other	43	39.81
TOTAL	**108**	**100**

Primary Activity

PRIMARY ACTIVITY	NO. OF RESPONDENTS	% OF RESPONDENTS
Data Supplier	9	8.33
End User	32	29.63
Model Developer	37	34.26
Modeller	30	27.78
TOTALS	**108**	**100.00**

Section Two –Metadata for Data

Question 2.1 For people searching for datasets which information is most important to identify and locate the data they need? Please rank the options below in relative importance (High, Moderate or Low)

	NO. OF RESPONSES			
	High	**Moderate**	**Low**	**Total**
Parameters or Phenomena Represented	76	10	3	89
Descriptive Information	71	13	4	88
Spatial Extent	67	22	1	90
Units of Measurement	53	20	17	90
Quality Assessments	46	36	8	90
Ownership	36	43	11	90
Use licence	35	38	16	89
Reference Dates	33	42	15	90
Spatial Reference System	32	39	19	90
Provenance	30	49	11	90
Original Purpose of the Dataset	12	45	33	90

Question 2.2 Which of the following metadata standards does your data comply with?

	% OF RESPONDENTS	NO. OF RESPONDENTS
Dublin Core Specification	10.59	9
ISO 19110	7.06	6
ISO 19115	11.76	10
ISO 19119	3.53	3
INSPIRE Data Specifications	29.41	25
PREMIS	0.00	0
Other Standards	37.65	32
TOTAL	**100**	**85**

Question 2.3 Do you think that there is sufficient metadata and other supporting information made available to help people find and locate the datasets you use or supply?

	NO. OF RESPONDENTS	% OF RESPONDENTS
No	48	54.55
Unsure	23	26.14
Yes	17	19.32
TOTALS	**88**	**100**

Question 2.4 Which of the following mechanisms do you use most often to locate and identify the data you use?

	NO. OF RESPONDENTS	% OF RESPONDENTS
Approaching recognised suppliers of specific datasets directly	17	19.77
Other Mechanisms	14	16.28
Searching via a catalogue facility (web based or otherwise)	24	27.91
Through organisations you already collaborate with	31	36.05
TOTALS	**86**	**100**

Question 2.5 When working with environmental datasets what metadata or other supporting information is most important to enable you to make effective use of the data? Please rank the options in relative importance (High, Moderate or Low)

	NO. OF RESPONSES			
	High	Moderate	Low	Total
Parameters or phenomena represented	72	15	0	87
Units of measurement	67	14	6	87
Spatial extent - Spatial reference system	61	22	4	87
Quality assessments	56	22	9	87
File formats available	47	32	8	87
Descriptive information (title, abstract etc)	46	27	14	87
Data formats used to store the data	42	37	8	87
Reference dates	35	36	16	87
Provenance	33	35	19	87
The software used to create the dataset	11	37	39	87
Digital Object Identifier (DOI)	8	27	49	84

Question 2.6 Do you think that there is sufficient metadata and other supporting information made available to help people make effective use of the datasets you use or supply?

	NO. OF RESPONDENTS	% OF RESPONDENTS
No	45	51.72
Yes	20	22.99
Unsure	22	25.29
TOTALS	87	100

Question 2.7 In terms of being able to both locate and use the datasets you need – if there was one thing you could improve in this process, what would it be? (64 free text responses)

Section Three – Metadata for Models

Question 3.1 For people searching for models which information is most important to identify and locate the model(s) they need. Please rank the options below in relative importance (High, Moderate or Low)

	NO. OF RESPONSES			
	High	Moderate	Low	TOTAL
Descriptive information	64	10	2	76
Parameters or phenomena represented	61	14	1	76
Spatial extent	51	21	4	76
Quality assessments	41	27	8	76
Use licence	35	31	10	76
Ownership	33	31	12	76
Deterministic / Probabilistic / Other	33	31	12	76
Technical Platform	32	29	15	76
Provenance	31	32	13	76
Original purpose of the model	27	37	12	76
Spatial reference system	26	38	12	76
Reference dates	18	47	11	76
Originating discipline	18	41	17	76
Programming Languages Used	16	36	24	76
Typical Runtime	13	43	20	76

Question 3.2 Do you assign metadata to models according to a formal standard? If so, which standard? (43 free text responses)

Question 3.3 Do you think that there is sufficient metadata and other supporting information made available to help people make use of the models you use or supply – what could be improved?

	NO. OF RESPONDENTS	% OF RESPONDENTS
No	40	53.33
Unsure	27	36.00
Yes	8	10.67
TOTALS	**75**	**100**

Question 3.4 When working with environmental models what metadata or other supporting information is most important to enable you to make effective use of the models?

	NO. OF RESPONSES			
	High	Moderate	Low	TOTAL
The datasets used as inputs	63	13	2	78
Parameters or phenomena represented	62	14	2	78
Assumptions made in building the model	61	17	0	78
The models used as inputs	59	13	4	76
Descriptive information	51	17	10	78
The details of the software or model code	35	28	14	77
Provenance	27	34	16	77
File formats available for input datasets	24	41	13	78
File formats available for output datasets	23	43	11	77
Compatible model coupling technologies	22	36	19	77
Additional toolkits (e.g. for visualising results)	13	50	15	78

Question 3.5 Considering information which would be useful in making effective use of environmental model, do you find this information easy to access? (47 free text responses)

Question 3.6 Do you find it easy to find models produced within other environmental disciplines?

	NO. OF RESPONDENTS	% OF RESPONDENTS
No	43	55.13
Unsure	22	28.21
Yes	13	16.67
TOTALS	**78**	**100**

Question 3.7 What are the barriers to the wider availability of models?

	NO OF. RESPONSES			
	High	Medium	Low	TOTAL
Lack of searchable catalogues	42	30	5	77
Lack of common ways to describe models	39	33	5	77
IPR considerations	35	31	11	77
Technical issues	23	36	17	76

Question 3.8 What criteria would you use to assess the quality of a model, and its fitness for purpose? (51 free text responses)

Section 4: Additional Questions

Question 4.1 If you supply metadata (or other supporting information) for your datasets and/or models what is your primary reason for doing this?

	NO. OF RESPONDENTS	% OF RESPONDENTS
Scientific objectives	18	28.13
To assist others using the dataset/model	25	39.06
To comply with organisational or other guidelines	10	15.63
To meet longer term data preservation objectives	2	3.13
To promote access to the dataset/model	9	14.06
TOTALS	**64**	**100**

Question 4.2 If you currently provide or make use of metadata in environmental modelling, do you use a particular software tool to create or access the metadata? (39 free text responses)

Appendix 2 Summary of current approaches

METADATA TOOLS

NASA's Earth Observing System (EOS) Clearinghouse (ECHO) is a metadata registry and order broker that allows query and access to data from a large number of repositories, primarily NASA repositories, though any repository can request to have their metadata included in the ECHO database, and stores data from a variety of science disciplines.

There are also several tools to assist the capture, cataloguing and retrieval of metadata in XML format, including the open source data management system – eXist; the metadata authoring tool, MATT; the Mercury web based system to retrieve metadata and associated datasets; and the open source metadata catalogue METACAT. The latter system is in use throughout the world to manage environmental data.

Another widely used geospatial metadata catalogue system is GeoNetwork OpenSource which is an open source geospatial data catalogue service host, metadata creation and management system, and basic web mapping platform. Another widely used system is the THREDDS Data Server (TDS) - a web server that provides metadata and data access for scientific datasets, using OPeNDAP, OGC WMS and WCS, HTTP, and other remote data access protocols.

REPOSITORY TECHNOLOGIES

Fedora (Flexible Extensible Digital Object Repository Architecture) is a modular architecture built on the principle that interoperability and extensibility is best achieved by the integration of data, interfaces, and mechanisms (i.e., executable programs) as clearly defined modules, and is often used in the digital library community.

EPrints is a free and open source software package for building open access repositories that are compliant with the Open Archives Initiative Protocol for Metadata Harvesting. It shares many of the features commonly seen in Document Management systems, but is primarily used for institutional repositories and scientific journals. EPrints is a Web and command-line application based on the LAMP architecture (but is written in Perl rather than PHP). It has been successfully run under Linux, Solaris and Mac OS X . A version for Microsoft Windows was released in May 2010.

D-Space is an open source tool aimed at organisations with minimal resources. The DSpace architecture is a straightforward three-layer architecture, including storage, business, and application layers, each with a documented API to allow for future customization and enhancement. The storage layer is implemented using the file system, as managed by PostgreSQL database tables.

Of relevance to the earth science community is the National Geospatial Digital Archive (NGA) which aims to create a new national federated network for archiving geospatial imagery and data, as well as collecting and archiving important digital geospatial data and images.

STORAGE TECHNOLOGIES

The JASMIN&CEMS cluster includes 4.6 Petabytes of usable fast access Panasas® parallel file storage (http://www.stfc.ac.uk/eScience/news+and+events/38663.aspx). The important aspects of the data storage design are the 1 Tb/s aggregate bandwidth from data to processors which supports the processing of very large data volumes, and the lower total cost of ownership than competing solutions due to less need for manual intervention by operators to manage and expand the system. The 1133 data blades constitute the second largest configuration that Panasas® have provided to a single installation.

Hierarchical storage management (HSM) is a data storage technique which automatically moves data between high-cost and low-cost storage media. HSM systems exist because high-speed storage devices, such as hard disk drive arrays, are more expensive (per byte stored) than slower devices, such as optical discs and magnetic tape drives. While it would be ideal to have all data available on high-speed devices all the time, this is prohibitively expensive for many organizations. Instead, HSM systems store the bulk of the data on slower devices and then copies data to faster disk drives when needed. The following link: http://www.stfc.ac.uk/e-Science/services/atlas-petabyte-storage/22459.aspx provides details of an STFC based example.

DATA PRESERVATION TECHNOLOGIES – SUMMARY AND MAIN TRENDS

Many of the software tools which are directly applicable to digital preservation are relevant to a wide variety of science (and sometimes also non-science) disciplines. Few are specific to the earth sciences, but a number of these technologies are concerned with the basic elements of files and their representation in computer systems. Hence they should be applicable to the types of file format commonly found in earth science archives. For example the EAST and DFDL data description language would potentially provide ways of describing a wide variety of data formats. Considering the aim of increasing the level of interoperability between different earth science disciplines the data dictionary (e.g. Data entity Data specification language) and semantic languages such as OWL and SKOS will be important in documenting data dictionaries and establishing new ontologies to ensure this interoperability.

The availability of emulators both for software and operating systems will be important. The Dioscuri emulator was designed by the digital preservation community and being java based can be ported to a number of platforms, and therefore seems a particularly useful tool. Important metadata tools (some of which are also referenced in the user surveys) include the open source metadata catalogue MERCAT which is widely used to manage environmental data and also the GeoNetwork metadata catalogue system which is widely used within the earth science community.

In terms of software archiving, a number of the available tools are also those commonly used by software developers during the development phase (e.g. SourceForge, and Subversion), since these provide mechanisms for documenting and version control of the code. Open source development communities (e.g. Tigris.org) also fulfil a useful function in digital preservation in that they provide a means for users to track and be informed about changes to their software, and often methods of upgrading open source applications as new versions of the underlying software become available.

Considering the technologies available for storage and archive repository development, FEDORA (Flexible Extensible Digital Object Repository Architecture) has been mentioned in the survey responses, and therefore is clearly used by the earth science community to some extent. Products such as EPrints and D-Space are probably more applicable to the digital library and academic publishing worlds, but may have some relevance to SCIDIP-ES. Repository planning tools such as the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) tool, did not come up in any of the user survey responses, but given the importance of auditing repositories and establishing the criteria for including certain data (and risks in not doing so) would seem to have a potential application in the earth science domain.

DATA DISCOVERY AND ACCESS

Portals appear to fall into two main types, those which provide a federated search across multiple archives and those which provide a dedicated search of a specific archive system. Frequently the database behind a specific portal can be accessed by federated search systems using OGC compliant standards and metadata. There is a strong indication that the facilities for federated searches across multiple archives are generally well developed.

The relevant OGC compliant standards include OGC Catalog Services (CSW) specification, Web Map Service (WMS), Interface Implementation Specification, Web Feature Service (WFS)

Implementation Specification, Web Coverage Service (WCS). These standards have been widely implemented to provide access to potentially very detailed and rich sets of geospatial information.

Of particular relevance to this project is the INSPIRE Geoportal (http://inspire-geoportal.ec.europa.eu/discovery/) which is the central discovery portal for the European geospatial data infrastructure (EU-GDI) providing a front end to an OGC compliant data catalogue, and also the GEO portal. The GEO Portal (http://www.geoportal.org/web/guest/geo_home) is the central portal and clearinghouse for Global Earth Observation System of Systems (GEO-GEOSS) providing access to geospatial and earth observation (EO) data. The GEO portal allows the user to discover, browse, edit, create and save geospatial information from GEO members around the globe. This data discovery portal accesses the OGC compliant catalogues, viewing and download services of various organizations worldwide through the use of standardized OGC-compliant protocols.

Another important project concerned with data access is GENESI-DEC (http://www.genesi-dec.eu/). The project has established open data access services allowing European and worldwide Digital Earth Communities to seamlessly access, produce and share data, information, products and knowledge. This will create a multi-dimensional, multi-temporal, and multi-layer information facility of huge value in addressing global challenges such as biodiversity, climate change, pollution and economic development. GENESI-DEC evolves and enlarges the platform developed by the predecessor GENESI-DR project by federating to and interoperating with existing infrastructures.

GENESI-DEC involves key partners of ESFRI projects and collaborates with key participants of Digital Earth and Earth Science initiatives, including the International Society of Digital Earth and GEO-GEOSS to ensure the efficient use of already existing and planned developments.

The INSPIRE, GEO-GEOS, and GENESI-DEC portals are front ends to large complex systems which allow data producers to upload data and metadata to the portal and also for users to retrieve their data.

The NERC Data Grid (http://ndg.badc.rl.ac.uk/) provides a gateway to find data and explore what is known about the datasets. The data themselves remain located with the data providers, and this provides a multi-archive search for discovering data. In a similar manner the Earth System Grid (ESGF - http://www.earthsystemgrid.org/) provides a gateway to scientific collections which may be hosted at sites around the world.

In some cases, in addition to the functionality to discover and access data, tools are also made available within the data discovery/access portal to enable visualisation of data, although it appears that this integration of visualisation and analysis tools is not currently a common feature.

The Heterogeneous Missions Accessibility (HMA) project aims to establish harmonised access to heterogeneous Earth Observation mission data from multiple missions ground segments, including national and ESA Sentinel missions. The project partners who already have a direct contractual relationship with ESA in the framework of HMA are: ASI (Italian Space Agency), CNES (French Space Agency), CSA (Canadian Space Agency), DLR (German Space Agency), EUSC (European Union Satellite Centre).

Other web portals examined are aimed at the discovery and access of earth observation data, and in many cases it is clear that the domains which these portals support are quite diverse. For example the Global Land Cover facility at www.landcover.org is commonly accessed by users from a diverse range of communities including from science (geography, earth science, ecology, climatology, conservation, education) environmental policy (global warming, sustainable development, risk management) and resource management (biodiversity assessment, forestry, protected area management). In other cases e.g. the SPOT catalogue and maps store (http://catalog.spotimage.com) and the "GMES Land Monitoring Portal" (http://www.land.eu/portal/) the portal provides access to a specific dataset or range of data sets.

As would be expected, data is generally provided in formats (e.g. GIS files or images) which are appropriate to the predominant user community. There is not a great deal of evidence of users from one discipline being able to access and use relevant data from disparate domains. In fact the form based search facilities frequently provided allow searching on the basis of terms such as location,

sensor, data type and time, some of which require a knowledge of earth observation data, and so may not encourage users of other disciplines to make use of it. This is clearly one area where the development of tools and services in the SCIDIP-ES project can contribute to making data more interoperable between disciplines.

TECHNOLOGIES AND FRAMEWORKS FOR PROCESSING DATA

These include the Web Processing Service (WPS) interface standard which provides rules for standardising inputs and outputs (requests and responses for geospatial processing services. Through WPS a generic user gains access to geospatial data processing tools provided by third parties. WPS can be seen as a way to perform standardized geospatial computations in a distributed environment. In the context of LTDP it can be used as a tool to preserve data processing algorithms and procedures in the geospatial domain as long as adequate data preservation policies are implemented on the infrastructure providing the service itself.

The OpenGIS® Web Coverage Processing Service (WCPS) Interface Standard (http://www.opengeospatial.org/standards/wcps) defines a protocol-independent language for the extraction, processing, and analysis of multi-dimensional gridded coverages representing sensor, image, or statistics data. Services implementing this language provide access to original or derived sets of geospatial coverage information, in forms that are useful for client-side rendering, input into scientific models, and other client applications.

Open virtualisation format (OVF) represents a standard vendor independent representation of virtual machines which, in turn, are a common component of data preservation strategies. A virtual machine containing all the processing chain components of a given dataset can be used to reproduce and analyse the procedures and algorithms used in data processing.

Earth System Modelling Framework (ESMF) defines an architecture for composing complex, coupled modelling systems and includes data structures and utilities for developing individual models. The ESMF framework is emerging as a standard among the modellers in the earth science domain. The standards and software tools defined by ESMF might be useful to support LTDP of model related data. Moreover, its components can be used as standardized data processing tools. ESMF is supported mainly by US organizations, universities and research centres.

Open Modelling interface (OpenMI) was developed within the EU funded projects HarmonIT and OpenMI-Life. OpenMI evolved to become a generic solution to build software components that can be applied to linking any combination of models, databases and analytical/visualisation tools. As an emerging standard in the domain of earth science will play a major role in preservation of data processing capabilities. Open MI has a similar role to the Earth Modelling Framework (ESMF) described above, although a key feature is that it is able to pass variables between models at run-time. A framework of open source components are used to “wrap” components of models and to this extent OpenMI may represent a useful means of preserving linked environmental models.

.