Scientific Data Application Profile Scoping Study Report

Document details

Author:	Alexander Ball, UKOLN, University of Bath
Date:	3rd June 2009
Version:	1.1
Document Name:	sdapss.pdf
Notes:	Changes from version 1.0: Typographical corrections made. References added. Conclusions expanded.

Acknowledgements

Acknowledgement to contributors

The author would like to thank the various people who contributed to the report by completing an interview or through e-mail correspondence, in particular: Linda Ault (BGS), Tim Booth (NEBC), Helen Campbell (BAS), Talat Chaudhri (UKOLN), Harry Dixon (National River Flow Archive), Steve Donegan (STFC), Monica Duke (UKOLN), Jane Greenberg (UNC Chapel Hill), Justin Hayes (MIMAS), Rachel Heaven (BGS), Stuart Jeffrey (ADS), Roy Lowry (BODC), Brian Matthews (STFC), Ken Miller (UKDA), Pam Miller (UKDA), Cameron Neylon (STFC), Manjula Patel (UKOLN), James Reid (EDINA), Robin Rice (EDINA), Belinda Robinson (BADC), Pauline Simpson (Central Caribbean Marine Institute), Humphrey Southall (University of Portsmouth), Mark Thorley (NERC), Matt Wood (Sanger Institute/Mekentosj Inc.), and Weimin Zhu (European Bioinformatics Institute). The author takes responsibility for interpreting the answers and for any change of emphasis that comes with collating the viewpoints of the various contributors.

Acknowledgement to funders

This work was funded by the JISC as part of the Repositories and Preservation Programme. UKOLN is funded by the Joint Information Systems Committee (JISC) of the Higher and Further Education Funding Councils and the Museums, Libraries and Archives Council (MLA), as well as by project funding from the JISC, the Research Councils and the European Union. UKOLN also receives support from the University of Bath where it is based.

Contents

Chapter 1

Introduction

This study was undertaken by UKOLN on behalf of the Joint Information Systems Committee (JISC).

1.1 Background

The background for this study as provided in the brief from the JISC is as follows.

Application profiles are metadata schemata which consist of data elements drawn from one or more namespaces, optimized for a particular local application [HP00]. They offer a way for particular communities to base the interoperability specifications they create and use for their digital material on established open standards. This offers the potential for digital materials to be accessed, used and curated effectively both within and beyond the communities in which they were created. The JISC has recently recognized that there is a requirement to undertake a scoping study to investigate metadata application profile requirements for scientific data[1] in relation to digital repositories, and specifically concerning descriptive metadata to support resource discovery and other functions such as preservation. This follows the development of the Scholarly Works Application Profile (SWAP) undertaken within the JISC Digital Repositories Programme and led by Andy Powell (Eduserv Foundation) and Julie Allinson (RRT UKOLN) on behalf of the JISC [AP08]. Application profiles for images (AHDS – Visual Arts), time based media (Manchester) and geospatial data (EDINA) are also being commissioned. Arguably, scientific data encompass a much wider range of resource types and are far more complex than other kinds of material, and so this scoping requires the consultants to explore whether harmonization around an application profile to improve resource discovery and reuse of scientific (research) data in the repository landscape can be achieved or is even desirable. The consultants will make recommendations for the repository landscape.

1.2 Aims and Objectives

As stated in the brief from the JISC, the aims and objectives of the study are:

1. to assess whether a single metadata AP for research data, or a small number thereof, would improve resource discovery or discovery-to-delivery in any useful or significant way;
2. if so, then to:
   a) assess whether the development of such AP(s) is practical and if so, how much effort it would take;

[1] By 'scientific data' is meant the evidence base on which academic researchers build their analytic or other work, where this evidence base is typically gathered, collated and structured according to declared and accepted protocols.

b) scope a community uptake strategy that is likely to be successful, identifying the main barriers and key stakeholders;

3. otherwise, to investigate how best to improve cross-discipline, cross-community discovery-to-delivery for research data, and make recommendations to the JISC and others as appropriate.

1.3 Scope

Scientific data as defined in this study covers a broad range of data types and contributes to the evidence base of many disciplines outside what is commonly termed Science.[2] Engineering and Technology, Medicine, Veterinary Sciences, Social Sciences, Humanities and the Arts all contain disciplines and modes of study where analyses are performed on carefully collected evidence bases. The scope of this study, therefore, is not limited to a particular set of disciplines, but rather to data gathered, collated, structured and analysed using a recognizably scientific method, with a bias towards quantitative methods.

The Scientific Metadata Model produced by CCLRC (now STFC) defines three types of scientific investigation: experiment, measurement and simulation [SM04]. Loosely speaking, an *experiment* measures the behaviour of a sample or system under controlled conditions, while a *measurement* measures the behaviour of a sample or system with as little as possible disturbance or intervention from the researcher. A *simulation* attempts to mimic or predict a measurement investigation using mathematical modelling. The common characteristics of experiments and measurements is that the researcher selects or collects a sample or system to study, applies an instrument to it, and generates a raw set of data. These data may then be collated, cleaned and otherwise processed to form a dataset more amenable to study. It will be seen that this paradigm applies equally well to census data, where the instrument is a census return, as it does to crystallography data.

Derived datasets, where one or more existing datasets are manipulated to create a new dataset, are not specifically excluded from the study, but the focus of the study is on primary data.

1.4 Methodology

The report *Dealing with Data* was used as a starting point for identifying UK data repositories and projects of relevance to this study [Lyo07]. Further repositories and projects were identified through desk research. Where possible, each repository and project was contacted and asked for details of the discovery-to-delivery metadata they use or have defined, alongside any insights they had from working with this metadata.

A series of unstructured interviews were conducted, discussing use cases for a Scientific Data Application Profile, and how widely a single profile might be applied. On the latter point, matters of granularity, the experimental/measurement contrast, the quantitative/qualitative contrast, the raw/derived data contrast, and the homogeneous/heterogeneous data collection contrast were discussed. The interviewees were:

- Cameron Neylon, STFC
- James Reid, EDINA (Geospatial Application Profile, Go-Geo!)
- Robin Rice, EDINA (DataShare)

[2] The traditional conception of Science probably coincides with the disciplines grouped by the Joint Academic Coding System under the categories Biological Sciences, Physical Sciences, and Mathematical and Computer Sciences [HES06].

- Brian Matthews, STFC
- Weimin Zhu, European Bioinformatics Institute
- Ken Miller, UKDA
- Humphrey Southall, University of Portsmouth (Great Britain Historical GIS)
- Justin Hayes, MIMAS (CAIRD)
- Monica Duke, UKOLN (Repository Search)

In addition a meeting was held with Mark Thorley (NERC) about the NERC DataGrid. Further use cases and issues were uncovered through desk research. This report represents a synthesis of the information and opinion gathered throughout the study. The conclusions were validated through a reference group of stakeholders.

1.5 Outline

The study begins in Chapter 2 with a summary of the repository landscape, giving details of repositories of scientific data in the UK and the metadata they use to describe their data. The chapter also gives details of projects that have defined or are defining discovery metadata for scientific data. Chapter 3 presents some use cases to illustrate the kinds of discovery-to-delivery activities a Scientific Data Application Profile would be expected to support. Chapter 4 compares and contrasts a number of different data models arising from the repository landscape, alongside some additional models of potential relevance, in order to determine if there is sufficient common ground on which to base a Scientific Data Application Profile. Chapter 5 presents common elements of the description metadata standards and profiles found to be in use in UK data repositories. The purpose of this is not only to uncover which metadata elements are most commonly found (or thought) to be useful in a discovery-to-delivery scenario, but also to indicate possible points of interoperability between a Scientific Data Application Profile and existing metadata standards and profiles. Further discussion, on topics such as implementation, is presented in Chapter 6 and the conclusions of the study may be found in Chapter 7.

Chapter 2

Repository Landscape

As noted in the report *Dealing with Data* [Lyo07], the data repository landscape is continually shifting and evolving. This study presents a snapshot from September 2008 to January 2009. *Dealing with Data* identified three funding bodies which fund dedicated data centres: the Arts and Humanities Research Council (AHRC), the Economic and Social Research Council (ESRC) and the Natural Environment Research Council (NERC). A further two funding bodies were identified that contribute to the European Bio-informatics Institute: MRC and the Wellcome Trust. In addition to these data centres, a number of institutions host their own data repositories, collecting and curating the scientific data outputs of their staff.

The following sections outline the descriptive metadata in use in a number of these data centres and data repositories.

2.1 NERC funded data centres

NERC funds a range of different data centres, of which the following are perhaps the most important. In addition, the NERC DataGrid Project is developing an infrastructure and interface that provides an aggregated view of the holdings of both NERC data centre holdings and those of other national and international data centres [Lat+09; Woo+07]. The NERC DataGrid currently interoperates using the DIF metadata schema as the common discovery metadata profile, but future versions will use ISO 19115 metadata [Law+09].

2.1.1 Centre for Environmental Data Archival

The Centre for Environmental Data Archival (CEDA) is based at the STFC's Rutherford Appleton Laboratory, near Didcot in Oxfordshire. It operates two NERC-funded Data Centres: the British Atmospheric Data Centre (BADC) and the NERC Earth Observation Data Centre (NEODC). The BADC was established as the Geophysical Data Facility (GDF) in 1984; its current name dates from 1994 when it became the NERC's Designated Data Centre for the Atmospheric Sciences. The NEODC is the Designated Data Centre for satellite and airborne sensor data about the surface of the Earth, and was established in 1985.

The BADC holds the outputs of NERC-funded projects dealing with measurements of weather, atmospheric chemistry and ocean surface temperature as well as numerical models and climate models. It also holds a significant amount of data from third parties such as the Met Office and the European Centre for Medium-Range Weather Forecasts (ECMWF), and provides access to data held at other centres, notably those held by NASA and ESA.

The NEODC holds data produced by NERC airborne Earth Observation surveys, and also acquires satellite data from ESA, NASA/NOAA and other parties. The types of data held range from aerial photography, multi-spectral scans and thematic mapping to sea surface temperatures, terrain mapping and mapping of vegetation, chlorophyll and ozone.

Both the BADC and NEODC use the MOLES (Metadata Object Links in Environment Sciences) schema for discovery metadata. MOLES was developed as part of the NERC DataGrid project, and currently stands at version 1.3 (see Appendix E). Version 2 is nearing completion, at which point it will deployed at the BADC and NEODC. The NERC DataGrid project has started the development of the information model for Version 3. Mappings exist via XQuery [W3C07] from MOLES to a number of other metadata schemata, including Directory Interchange Format (DIF) [GCM08a], Dublin Core [DCM08] and various profiles and serializations of the ISO 19115 standard [ISO03].

2.1.2 British Geological Survey National Geoscience Data Centre

The National Geoscience Data Centre (NGDC) is the Designated Data Centre for the Earth Sciences. It is funded and operated by the British Geological Survey (BGS), itself part of NERC, at the BGS' Keyworth site near Nottingham.

The NGDC holds over four hundred datasets: from BGS activities, from NERC-funded projects and NERC thematic programmes, and from industrial organizations, as well as those deposited under the Mining Industry Act 1926 and Water Resources Act 1991. The data cover all aspects of Earth Sciences; the NGDC holds both geographic and non-geographic data, and both digital data and physical samples.

The discovery metadata used by the NGDC is a custom profile made up of the core elements of the ISO 19115 standard, along with some non-core elements and some local elements not present in the ISO standard (see Appendix F). The profile was heavily influenced by the need to support metadata written according to the National Geospatial Data Framework's *Discovery Metadata Guidelines* [NGD00], which was used prior to ISO 19115 being published, and also metadata used internally for three dimensional models. The profile does not currently include GIS layer metadata, though this is set to change; in the meantime, the GIS layer metadata are being loaded into the catalogue database as read-only XML documents. The database can display and export the metadata in DIF (see Appendix B); further transformations are planned to enable export to ISO 19139 [ISO07], the XML version of ISO 19115, and UK GEMINI [EGU04], the UK Government profile of ISO 19115 (see Appendix G).

2.1.3 British Oceanographic Data Centre

The British Oceanographic Data Centre (BODC) is the Designated Data Centre for data concerning the marine environment, and also one of the Intergovernmental Oceanographic Commission's network of sixty national data centres. It is funded by the NERC and operated by the NERC's Proudman Oceanographic Laboratory (POL), based in Liverpool. It is a well established data centre, having been formed as the British Oceanographic Data Service (BODS) in 1969; its current name and form derive from a restructuring in 1989.

As well as managing the data of many UK and international marine research projects, the BODC also holds the National Oceanographic Database (NODB) – which collects together marine data sets mainly from UK research establishments – and the data generated by the UK Tide Gauge Network.

The BODC currently uses EDMED (European Directory of Marine Environmental Datasets) version 0 [SS00] as its native discovery metadata format (see Appendix H), but it is in the process of converting to EDMED version 1. The primary differences between the two versions are that some plain text fields in version 0 are replaced by structured text fields in version 1, and that the XML serialization of version 1 is also a profile of ISO 19115. DIF records are generated automatically from the EDMED version 0 records, and will soon be generated from EDMED version 1 records as well; in the meantime, some native DIF records are also held. In addition, the BODC also holds some data for which the only discovery metadata are those contained in the Cruise Summary Report (see Appendix I).

2.1.4 Antarctic Environmental Data Centre

The Antarctic Environmental Data Centre (AEDC) is the Designated Data Centre for Polar Science, and as the UK's National Antarctic Data Centre, its Manager is a member of the Joint Committee on Antarctic Data Management (JCADM). It is funded and operated by the British Antarctic Survey (BAS), itself part of NERC, at the BAS' site in Cambridge.

The AEDC holds data collected by UK scientists in Antarctica and the Southern Ocean, and therefore manages a wide variety of both physical samples and digital datasets: from plasmasphere measurements through flora and fauna monitoring data to geophysical data. Under the terms of the Antarctic Treaty, all AEDC data is also catalogued in the Antarctic Master Directory, operated by NASA's Global Change Master Directory (GCMD).

The descriptive metadata held by the AEDC is in a local profile designed to map easily to both DIF and ISO 19115.

2.1.5 National Water Archive

The National Water Archive (NWA) was formerly the Designated Data Centre for Hydrology, but has now been subsumed under the Centre for Ecology and Hydrology's (CEH) Environmental Information Data Centre (EIDC), the Designated Data Centre for Terrestrial and Freshwater Science, Hydrology and Bioinformatics. The NWA is made up of two data centres: the UK National River Flow Archive (NRFA), operated by the CEH, and the National Groundwater Level Archive (NGLA) maintained by the BGS. Both are based at the CEH site at Wallingford, Oxfordshire.

In both the NWA archives, the holdings consist of time series data from monitoring stations; in the case of the NRFA, these are gauging stations measuring river flow rates, while the NGLA collects data about the water levels in or around 170 wells and boreholes across the UK.

The NWA does not hold formal descriptive metadata about its datasets, but does hold information about the monitoring stations. For example, the NRFA records for each gauging station: its geographical location, its hydrometric area, the catchment area of the river at the point of the gauging station, and details of its history (when it was built, when its monitoring equipment was upgraded, if applicable when it was decommissioned, etc.). This information, alongside some tacit metadata, could be used to provide formal descriptive metadata if required.

2.1.6 NERC Environmental Bioinformatics Centre

The NERC Environmental Bioinformatics Centre (NEBC) was originally established to provide data management and Bioinformatics services for the NERC's Environmental Genomics Thematic Programme, though this remit was later extended to serve the Post-Genomic and Proteomics Thematic Programme. Its data curation activities are now performed under the aegis of the Centre for Ecology and Hydrology's (CEH) Environmental Information Data Centre (EIDC), the Designated Data Centre for Terrestrial and Freshwater Science, Hydrology and Bioinformatics, based at the CEH site at Wallingford, Oxfordshire.

The NEBC holds both 'omic' data – nucleotide sequences, protein sequences, proteomic data, metabolomic data, microarray results, etc. – and non-'omic' data – such as phenotypic or demographic data. It holds metadata at two different levels: dataset-type-independent and dataset-type specific. The dataset-type-specific metadata uses existing standards wherever possible: dbEST for expressed sequence tag DNA sequences, MIAME/Env for Transcriptomics and MIAPE for Proteomics. The dataset-type-independent metadata conforms to a local schema loosely based on the UK Environmental Data Index catalogue (see Appendix J). Output filters exist to translate this metadata into DIF.

2.2 ESRC and AHRC funded data centres

2.2.1 UK Data Archive

The UK Data Archive (UKDA) specializes in Social Science and Humanities data, and is a designated legal place of deposit for digital records from the National Archives (TNA). It was established in 1964 and is funded by the Economic and Social Research Council (ESRC), the JISC, and the University of Essex where it is based. The UKDA has strong international links: it is a member of the Council of European Social Science Data Archives (CESSDA), the International Association of Social Science Information Service and Technology (IASSIST), the US Inter-university Consortium for Political and Social Research (ICPSR) and the International Federation of Data Organizations (IFDO).

The UKDA is the lead partner in the Economic and Social Data Service (ESDS) and hosts the History Data Service (formerly part of the Arts and Humanities Data Service) and the Census.ac.uk service. It is responsible for over 5000 datasets, both qualitative and quantitative, as well as collections of multimedia and physical resources. The metadata it holds for these datasets conform to a local profile of the Data Documentation Initiative metadata standard (see Appendix C and Appendix K); the profile uses elements agreed as mandatory by all CESSDA members.

2.2.2 Archaeology Data Service

The Archaeology Data Service (ADS) formerly hosted Arts and Humanities Data Service Archaeology (AHDS Archaeology) and is now part-funded directly by the Arts and Humanities Research Council (AHRC). It was established in 1996 by a consortium of eight UK universities and the Council for British Archaeology, and is hosted by the University of York.

As well as holding an aggregated catalogue of over a million archaeological interventions, sites and monuments, the ADS holds data from both fieldwork projects and laboratory investigations. Examples include geophysical and topographic survey data, stratigraphic databases, metrical data and specialist finds reports. The corresponding metadata is modelled on the MIDAS Heritage format, the UK historic environment data standard [FIS07] (see Appendix D). The ADS are also in the process of implementing a faceted classification scheme for resource discovery with geographical location, temporal coverage, site type and medium/format as the four facets covered. The facet structure draws its controlled vocabularies from thesauri created by the National Monuments Record as well as MIDAS and the UK Government administrative area lists.

2.3 European Bioinformatics Institute

The European Bioinformatics Institute (EBI) is part of the European Molecular Biology Laboratory (EMBL), but receives additional funding from the European Commission, the Wellcome Trust, the US National Institutes of Health, UK Research Councils, the UK Department of Trade and Industry, and members of the EBI Industry Programme. It was established at the Wellcome Trust's Genome Campus in Hinxton, Cambridgeshire in 1992, taking over from and expanding the role of the EMBL Nucleotide Sequence Data Library, which had been running in Heidelberg, Germany since 1980.

The EBI holds a wide range of datasets from the field of molecular biology, including nucleotide sequences, genomes and gene-expression data, protein sequences, protein–protein interactions, biological pathways and macromolecular structures. Each database held by the EBI has its own structure and data model, and consequently its own metadata profile, although some consistency is achieved using controlled vocabularies. For cross-searching, the EBI's BioWisdom SRS platform and a Lucene engine perform full-text indexing of the database tables; some more sophisticated text mining techniques are also being considered. Further linkages are provided using citations to published papers (using PubMed identifiers or DOIs) and other cross-references. Because of the rapid evolution of understanding in the field and consequent changes in the data models used, the maintenance of formal mappings between the various metadata profiles in use is costly, and while the EBI is involved in standardization work, it does not rely on such mappings at a technical level.

2.4 STFC e-Science Centre

The e-Science Centre at the Science and Technology Facilities Council (STFC) provides a range of services for both internal and external customers, including data management and curation. Centre staff are based either at the Daresbury Laboratory in Halton, or at the Rutherford Appleton Laboratory, near Didcot in Oxfordshire.

The Data Management Group at the e-Science Centre has produced a scientific metadata model for the purposes of interoperable resource discovery of datasets [SM04]. The model has been adopted for the data assets held by the STFC and by various other projects including the Australian Research Enabling Environment (ARCHER). A summary of the metadata defined by the model may be found in Appendix L. The model itself is discussed further in Chapter 4.

2.5 Data repository projects

2.5.1 Go-Geo!

Go-Geo! is a JISC-funded service to provide a portal for accessing metadata about geospatial data collections. The first three phases of the Go-Geo! Project ran between 2001 and 2004, and were a collaboration between EDINA National Data Centre, University of Edinburgh, and the UK Data Archive, University of Essex. From November 2004, the Go-Geo! portal was run as a trial service to the UK Higher and Further Education communities by EDINA, and in November 2008 became a full JISC service. The portal works by querying a number of different data repositories using the Z39.50 protocol [ANS95] and collating the results. This in turn relies on the provision of consistent metadata by each of the Z39.50 targets. The standard used is the Academic Geospatial Metadata Application Profile (AGMAP, see Appendix M), a profile of ISO 19115 and a superset of UK GEMINI, specifically produced for the UK Higher and Further Education communities by the Go-Geo! Project.

2.5.2 eBank UK

The eBank UK Project ran between 2003 and 2007 with funding from the JISC. It was led by UKOLN, University of Bath, with partners from the University of Southampton, the Digital Curation Centre and the University of Manchester. The project brought together an interdisciplinary team, drawn from chemistry, computer science and the digital library sphere, to explore how crystallographic data describing molecular structures could be archived and disseminated using a distributed network of digital repositories. One of the key deliverables from the project was the eCrystals repository at the University of Southampton, the further development of which is now the domain of the follow-on eCrystals Federation Project.

One of the outcomes of the initial implementation of the eCrystals repository was an application profile for the metadata to accompany the crystallographic data in the repository (see Appendix N). The final form of this profile was strongly based on Dublin Core metadata elements, simplifying its implementation in ePrints, the repository software in use at the University of Southampton. The eCrystals Federation Project is in the process of creating a more extensive application profile, using the eBank profile as a starting point.

2.5.3 SPECTRa

The SPECTRa (Submission, Preservation and Exposure of Chemistry Teaching and Research Data) Project ran between 2005 and 2007 with funding from the JISC [TM07]. It was a partnership between the University of Cambridge and Imperial College London, and produced Open Source tools to enable high volumes of chemical data – specifically from the fields of synthetic organic chemistry, crystallography and computational chemistry – to be deposited into a DSpace repository and subsequently reused. It was a high priority of the project that these tools should integrate with established workflows of the chemists producing the data.

The metadata profile used by the project was an extended version of the eBank profile (see Section N.3). The principal differences were that the SPECTRA profile: a) distinguished the chemist owning the data from the spectroscoper or crystallographer who generated it; b) was adapted to apply to organic and computational chemistry, and c) included information relating to open access embargoes.

2.5.4 DataShare

The Data Information Specialists Committee – United Kingdom (DISC-UK) DataShare Project is a JISC-funded project investigating ways of incorporating research data into institutional repositories; it runs from 2007 to 2009. The three institutions taking part are the Universities of Edinburgh, Oxford and Southampton. One of the aims of the Project is to produce exemplars for the technical handling of research data by repository software – specifically EPrints, DSpace and Fedora – including the metadata that should be held for each dataset. The Edinburgh DataShare has constructed a profile of Dublin Core Metadata Terms and DSpace metadata with the specific needs of multidisciplinary research data in mind; this profile may be found in Appendix O. This profile has been adapted by ePrints Soton to conform with its existing metadata profile for research publications [Gib09].

2.5.5 Data Audit Framework

The Data Audit Framework (DAF) is a methodology and accompanying tools to enable universities and other institutions to identify and evaluate the data they hold and their policies and procedures for managing it. It was developed with JISC funding during 2008 and early 2009 by HATII at the University of Glasgow, in partnership with the University of Edinburgh, UKOLN at the University of Bath, Kings College London, Imperial College London and University College London.

Phase 3 of the DAF Methodology involves the collection of management metadata concerning the data assets identified by the audit [JRR08, pp. 31–36]. While the emphasis of the metadata set is on data management, the set does contain elements relevant for the discovery and access of the data asset (see Appendix P).

2.6 Other projects

2.6.1 Dryad Project

The Dryad Project is a collaboration between the Metadata Research Center at the University of North Carolina at Chapel Hill and the (US) National Evolutionary Synthesis Center (NESCent), with funding from the National Science Foundation. The aim of the project is to set up a repository for evolutionary biology datasets, particularly published ones, and to link those datasets with major journals and databases in the field. The project is putting together an application profile suitable for the diverse range of datasets the repository will host [Whi+08].

2.6.2 DCMI Science and Metadata Community

In January 2009, following discussions at a workshop at the 2008 International Conference on Dublin Core and Metadata Applications, the Dublin Core Metadata Initiative set up a Science and Metadata Community as a forum for individuals and organizations to exchange information and knowledge about metadata describing scientific data. A workplan for the Community will be established at the 2009 International Conference on Dublin Core and Metadata Applications [DCM09].

Chapter 3

Use Cases for a Scientific Data Application Profile

The Singapore Framework for Dublin Core Application Profiles [NBJ08] defines an application profile as consisting of a set of functional requirements, a domain model, a description set profile, syntax guidelines, data formats and usage guidelines. This chapter identifies potential use cases for a Scientific Data Application Profile; Chapter 4 looks at common features of domain models already in use for scientific metadata profiles, while Chapter 5 identifies common features of existing description set profiles for scientific data. Syntax guidelines, usage guidelines and data formats are not compared as they are too specific to the application and/or description set profile in question to be instructive for the purposes of this study.

The use cases below are presented under four headings: broad enquiry, specific follow-up work, detailed enquiry, data mining and aggregator functions [Rya07]. The use cases themselves are based on interviews conducted with data producers, managers and librarians, as well as previous work [DAD07; Gre07; OE07; Woo+07].

As this study is concerned with the discovery to delivery of scientific data, these use cases focus predominantly on researchers seeking out data relevant to their studies. A common factor across the use cases is an aggregator that allows multiple repositories to be searched simultaneously, as this is the primary benefit of many repositories sharing a single application profile, at least in the discovery context. The use cases are agnostic as to whether the data are held in discipline-based data centres or institutional data repositories. Use cases involving the management, curation and preservation of scientific data are not considered as they are out of scope.

3.1 Broad enquiry

Use Case 1. A researcher is seeking a broad overview of data available for a topic. The researcher chooses a data repository aggregator or cross-search service and searches for relevant data, using entry points such as the field of research, a keyword or identifier for an experimental subject, geographical area (e.g. latitude/longitude 'square') or time period. The search returns a list of datasets, along with a brief abstract or summary. The researcher may narrow down the result set by performing further searches within the result set or a subset thereof. The items in the result set link through to fuller records, which include information on spatiotemporal resolution that the researcher can use to assess if the data are suitable. Where available, the researcher makes use of preview images or data to make comparisons and gain a preliminary understanding of the data. If this is not sufficient, the researcher uses details of how to access the data, also part of the detailed records, to obtain a copies of the most relevant and useful datasets in a suitable format.

Use Case 2. A researcher is seeking inspiration from previous studies on a suitable methodology (e.g. which questions to ask in a questionnaire, which coding scheme to use). The researcher chooses a data repository aggregator or cross-search service and searches for relevant data, using entry points such as the field of research and the variable or keyword to be measured. The search returns a list of datasets, along with a brief abstract or summary. The researcher may narrow down the result set by performing further searches within the result set or a subset thereof. The researcher uses the descriptions to determine which data are relevant. The items in the result set link through to fuller records; those which do not include information on the methodology directly, provide further avenues of enquiry, e.g. contact details for the investigators, discipline specific metadata packages which may describe the methodology, links to published papers where the methodology is described.

3.2 Specific follow-up work

Use Case 3. A researcher is reviewing existing literature on a topic of study and discovers a previous piece of research; this may be through a published paper, a project description, and/or through contact with other researchers or those close to the previous research. The researcher wishes to review the data to verify the findings, but there is no direct link to the data. The researcher chooses a data repository aggregator or cross-search service and searches for the relevant data; search entry points include the names of the investigators, the name of the project, the grant number of the project, and the citation for a corresponding published paper. The search returns a list of possible datasets, along with a brief abstract or summary. The researcher may narrow down the result set by performing further searches within the result set or a subset thereof. The items in the result set link through to fuller records, which include details of how to access the data (e.g. URL). The researcher uses the descriptions to determine which data are relevant, and then uses the access information to obtain a copy of the data in a suitable format.

3.3 Detailed enquiry

Use Case 4. A researcher is interested in a particular type of measurement made within a defined geographical area. The researcher chooses a data repository aggregator or cross-search service and searches for relevant data; search entry points include geographical area (e.g. latitude/longitude 'square'), time period, the field of research and the variable or keyword measured. The search returns a list of possible datasets, along with a brief abstract or summary. The researcher may narrow down the result set by performing further searches within the result set or a subset thereof. The items in the result set link through to fuller records, which include information on spatiotemporal resolution, data quality, provenance and data collection methodology that the researcher can use to assess if the data are suitable. Where available, the researcher makes use of preview images or data to make comparisons and gain a preliminary understanding of the data. The researcher uses details of how to access the data, also part of the detailed records, to obtain a copies of the most relevant and useful datasets in a suitable format.

Use Case 5. A researcher is interested in a particular type of measurement made of a defined sample. The researcher chooses a data repository aggregator or cross-search service and searches for relevant data; search entry points include the field of research, the variable or keyword measured, a keyword or identifier for the sample measured and run numbers if known. The search returns a list of possible datasets, along with a brief abstract or summary. The researcher may narrow down the result set by performing further searches within the result set or a subset thereof. The items in the result set link through to fuller records, which include information on data quality, provenance and data collection methodology that the researcher can use to assess if the data are suitable. Where available, the researcher makes use of preview images or data to make comparisons and gain a preliminary understanding of the data. The researcher uses details of how to access the data, also part of the detailed records, to obtain a copies of the most relevant and useful datasets in a suitable format.

3.4 Data mining

Use Case 6. A researcher is investigating a trend in an observed phenomenon, and wishes to gain an understanding of possible links with other observed trends. The researcher therefore seeks to use data mining to discover associations between data relating to the observed phenomenon and other data from appropriate datasets. The researcher chooses a data repository aggregator or cross-search service and searches for relevant data; search entry points include geographical area (e.g. latitude/longitude 'square'), time period, and keywords identifying possible points of correspondence with the initial dataset. The researcher may narrow down the result set by performing further searches within the result set or a subset thereof. The items in the result set link through to fuller records, which include information on spatiotemporal resolution, data quality, provenance, data collection methodology, formatting and licensing that the researcher can use to assess if the data are suitable. Where available, the researcher makes use of preview images or data to make comparisons and gain a preliminary understanding of the data. The researcher uses details of how to access the data, also part of the detailed records, to obtain a copies of the most relevant and useful datasets. Additional metadata provided with the dataset allows for proper interpretation of and interaction with the data values.

Use Case 7. A researcher is investigating a trend in an observed phenomenon, and wishes to gain an understanding of possible links with other observed trends. The researcher therefore seeks to use data mining to discover associations between data relating to the observed phenomenon and other data from appropriate datasets. The researcher designs software that queries a data repository aggregator or cross-search service and searches for relevant data; search entry points include geographical area (e.g. latitude/longitude 'square'), time period, and keywords identifying possible points of correspondence with the initial dataset. The API for the aggregator allows the software to fetch machine-readable information on spatiotemporal resolution, data quality, formatting and licensing that the software can use to assess if the data are suitable. The software uses machine-readable details of how to access the data, also provided by the aggregator, to obtain a copies of the most relevant and useful datasets. Additional metadata provided with the dataset allows for proper interpretation of and interaction with the data values.

3.5 Aggregator functions

Use Case 8. An aggregation service wishes to implement cross-searches across multiple data repositories. Since many repositories are using the same application profile, either natively or as an additional metadata format generated from native metadata, the same adaptations can be made to include all these repositories. The aggregation service is already able to handle standard Dublin Core metadata, harvested using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), so in the short term it harvests the Scientific Data Application Profile metadata and uses a dumbing-down procedure to interpret it as Dublin Core. This allows the records to be searched using the existing interface. In the longer term, the aggregation service writes an interpreter and interface that takes advantage of the richness of the Scientific Data Application Profile metadata. Any discipline-specific metadata referenced by the records are not displayed directly by the aggregator, but a link is provided allowing the end user to access it, either directly or via the aggregator (if encapsulated within the more general record).

Alternative use cases could be constructed where an aggregator harvests metadata using newsfeed technologies such as RSS or Atom in place of OAI-PMH.

3.6 Discussion

3.6.1 Granularity of description

There are a number of different granularities at which one can consider data (see Chapter 4) but in the simplest case one can distinguish between a data table, a dataset and a collection. By *data table* is meant, for example, a series of data tuples (e.g. time stamp and sensor reading, or respondent and questionnaire responses). By *dataset* is meant a set of data tables that are closely linked, either by being different versions of each other or through being generated in the course of the same investigation. By *collection* is meant all the datasets held by an archive or institution. It is possible for a collection to consist of a single dataset, and for a dataset to consist of a single data table.

The use cases above portray scenarios in which metadata are provided at the dataset level. Lawrence et al. [Law+09] argue that discovery services based on metadata at the data table ('data granule') level are unworkable, since at that level one would be searching through a vast number of poorly differentiated items, and that the searcher is better served reaching the data table via the dataset. Furthermore, in some disciplines, the rate of advance means that the requirements for detailed metadata evolve rapidly, leading to difficulties maintaining interoperability. This does not mean that metadata at the data table (or even data point) level is not necessary; indeed, some applications – particularly those involving the mass refactoring of data – require such metadata. Rather, it means that at the initial discovery stage such metadata should not be used, but the discovery metadata should provide sufficient information to allow the more detailed metadata to be obtained. This reasoning underlies use cases 6 to 8 above, where detailed metadata are made available separately from the general metadata.

Another reason for separating detailed metadata from discovery metadata is to resolve the tension between providing the metadata needed to interpret data in their proper disciplinary context, and providing the metadata needed to find the data in an interdisciplinary context. While the detailed metadata may be vital to evaluating, understanding and interpreting the data, it would not be appropriate to attempt the construction of a schema to provide these metadata generically, but on the other hand, there are severe limits to the search functionality one can provide over records using multiple different discipline-specific metadata schemata.

At the other extreme, if only collection level descriptions are provided, searchers are unable to cross-search collections held by different archives or institutions. A partial use case for this scenario is provided below.

Use Case 9. A researcher is interested in a particular type of measurement made within a defined geographical area. The researcher consults a collection description database and searches for collections that might hold relevant data; search entry points include the field of research and geographical area. The search returns a list of archives and/or institutions that have datasets in the relevant field and datasets in the relevant geographical area, alongside links to the catalogues for each. The researcher follows each link in turn. For each link, the researcher familiarizes himself/herself with the search interface and proceeds to search for relevant data. . . .

There are clear scalability issues with this use case. As more datasets are added to a collection, the superset of geographical areas covered would grow, as would the range of research topics covered – particularly for generalist repositories such as those provided by Higher Education

Institutions (HEIs) – potentially leading to a high number of false positives for any given search. Indeed, given the heterogeneity of datasets produced by HEIs, it is hard to see how the value of collection-level descriptions would be maintained in the face of growing collections, without at least basic dataset-level information being provided.

3.6.2 Interdisciplinary search

One of the major barriers to interdisciplinary search is the way in which terminology varies between disciplines [Law+09; Spa01]. Obvious problems exist where different disciplines use different terms for the same concept, or use the same term for different concepts, but there are also more subtle problems where different disciplines use the same term for the same concept, but with different standards for when the term may be used.

Many of the metadata schemes examined in this study attempt to address this problem using one or more controlled vocabularies; some metadata schemes specify explicitly which vocabulary to use, while others allow the decision to be taken by the organization implementing the scheme. Of particular note is the technique used by the DIF standard [GCM08a], which uses a hierarchical vocabulary with disciplinary fields at the top level and measured parameters at the bottom level. As every intermediate level in the hierarchy is specified, the terminology describing the data is always placed in the appropriate disciplinary context. While this helps to reduce the number of false positives among search results, in order to reduce the number of false negatives, mappings would also need to be provided between terms in different disciplinary branches of the vocabulary; such mappings do tend to gloss over the more subtle disciplinary differences between terms, but this tends to be a more significant issue at the interpretation stage rather than the data discovery stage. More serious is the issue that few organizations have the resources to produce and maintain such mappings, due to the effort involved.

The World Wide Web Consortium is developing a system for providing mappings between vocabularies, ontologies and so on ('knowledge organization systems'), based on the Resource Description Framework (RDF). The Simple Knowledge Organization System (SKOS) allows the relationships between (URI-identifiable) terms in different knowledge organization systems to be stated in such a way as to permit automated reasoning [W3C09]. Use of such a system to create a network of mappings may help reduce the burden of maintenance, if the indirect mappings provided by the network are sufficient to remove the necessity for a direct mapping. Even so, this does not resolve the issue of the cost of producing and maintaining the mappings that *are* needed.

The production of controlled vocabularies of disciplinary fields, subjects and keywords, and mapping the relationships between them, is clearly beyond the scope of a Scientific Data Application Profile. Nevertheless, it would be prudent for such a profile to allow controlled terms to be used, and for the vocabulary from which they are drawn to be specified, to allow software tools to make use of any mappings or translations that may be (or become) available. In situations where an aggregator wishes to use text mining techniques over free-text descriptions of datasets, controlled keywords may help to give the text mining tools additional context.

3.6.3 Suitability of default OAI-PMH metadata

The Open Archives Initiative Protocol for Metadata Harvesting requires that compliant servers should provide at least a Simple Dublin Core record (i.e. using only the original fifteen Dublin Core elements, with no use of encoding schemes, element refinements or complex values, but allowing elements to be repeated) for each item in the associated catalogue or repository [OAI08]. They may also provide other metadata schemata, and clients may query an OAI-PMH server to ask which schemata it supports. Approximately two thirds of UK repositories provide metadata according to more than one schema [Bal09].

Given that OAI-PMH-compliant repositories must support Simple Dublin Core anyway, consideration must be given to the additional benefits a Scientific Data Application Profile could have over this generic resource description. First, OAI-PMH Simple Dublin Core does not allow for multiple resources to be described in a single record, so it would not be possible to give information about data tables within a dataset without providing separate records for them. Second, as it is not possible in OAI-PMH Simple Dublin Core for a single element to have multiple values, the logical relationship between repeated elements is ambiguous. For example, repetition of the 'format' element could mean either that the same data is available in several alternative formats, or that different parts of the data are in different formats. Third, it is not possible in OAI-PMH Simple Dublin Core to specify how an element has been encoded, or if a controlled vocabulary has been used; in the former case this makes automated handling of the data more difficult and possibly less robust, while in the latter case this loses valuable context and negates some of the advantages of using a controlled vocabulary, particularly for interdisciplinary search. Fourthly, OAI-PMH Simple Dublin Core is missing some elements called for in the above use cases, such as access methods, contact details for investigators, preview data/images, project name, data quality measures, provenance (excepting antecedent datasets) and data collection methodology. While it is true that some of this information could be provided as part of, say, a 'description' element, it would be programmatically difficult to separate out the information again for use as an access point or differentiating the information when presenting it to a user.

Chapter 4

Domain Models for Scientific Data

There are already a number of different conceptual models of scientific data in existence; some are defined independently while others are defined in the course of defining a metadata profile. The models presented in this Chapter are by no means an exhaustive set, but are intended to show how data are viewed in a variety of domains; for the most part, they are drawn from the metadata standards introduced in Chapter 2. The models from DDI and the MIDAS Heritage Standard show the perspective from the Social Sciences and Humanities domains, while the MOLES model and the Scientific Metadata Model show the perspective from observational, experimental and simulation science. Two further models are presented giving a more general perspective: the Draft International Standard ISO 2146 for Registry Services, and the Functional Requirements for Bibliographic Records (FRBR) model. In Section 4.7, the similarities and differences between the models are explored.

4.1 DDI

The DDI standard does not have a formally declared domain model, but one can be inferred from the structure of its modules. The significance of a module within DDI is that each has a schema associated with it, and therefore can be implemented on its own or within the larger hierarchy. There are eleven modules that represent the structure of the full DDI hierarchy, plus an additional five that represent different types of data structure. There are also a number of utility modules; *Reusable* contains common elements that would otherwise occur in several modules, while interface modules are provided for compatibility with Dublin Core, XML and XHTML.

The top level module in the hierarchy is *DDI Instance*. This corresponds to a programme of research consisting of one or more related studies. Within the hierarchy, a single study is represented by the *Study Unit* module. If there are close similarities between several Study Units, they may be aggregated within a *Group* module; this module provides a mechanism for mapping between the properties of Study Units using the *Comparative* module.

There are four modules that may be used directly within either the Study Unit or Group modules. The *Data Collection* module corresponds to the process by which data were collected, and hence deals with research methodology, survey instruments and data processing. The *Logical Product* module considers the data as a collection of facts, and deals with the parameters measured or controlled, the categories used, and the dependencies between the data. The *Conceptual Component* module deals with the subject matter of the data: what was measured, how it was grouped and which concepts were being tested. Finally, the *Archive* module deals with the organizations storing and curating the data, giving details on how the data are managed and how they may be accessed.

The remaining modules relate solely to Study Units within the hierarchy. The *Physical*

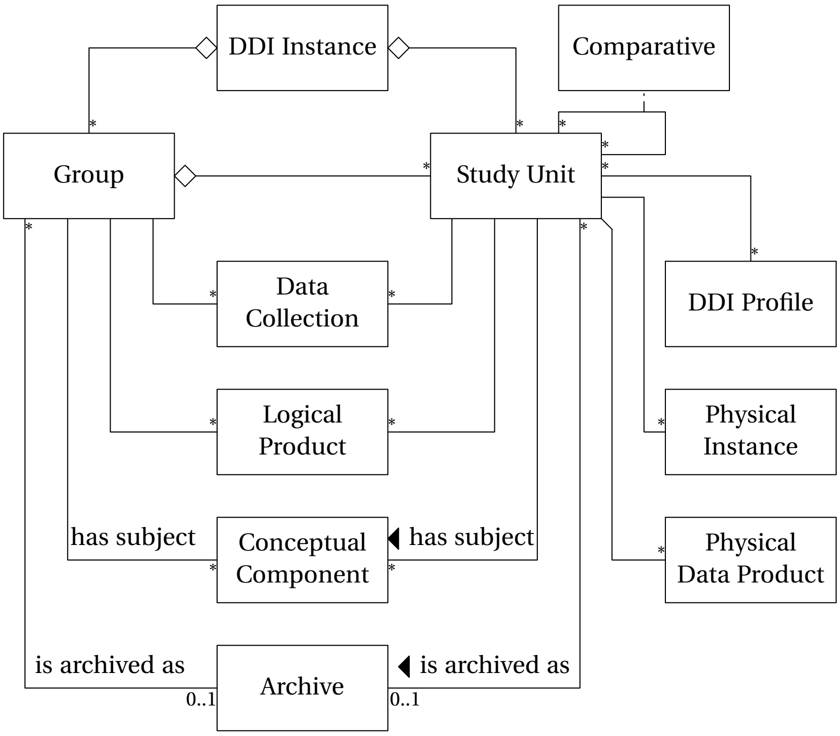

Figure 4.1: Simplified entity analysis of the DDI version 3.0 metadata standard.

Data Product module deals with how the logical product is represented as a file – file format, additional formatting conventions, etc. — whereas the *Physical Instance* module deals with a particular instance of the data product: how it was created, how it was checked, fixity checksums, etc. Finally, *DDI Profile* describes any additional conventions used in the metadata.

These modules and the relationships between them may be interpreted as an entity model as illustrated in Figure 4.1.

4.2 MIDAS Heritage Standard

The MIDAS Heritage Standard defines a number of information groups, each of which corresponds to an entity about which a catalogue record could be provided. These are:

- Heritage Assets
  - Area
  - Monument
  - Artefact and ecofact
- Activities
  - Investigative activity
  - Designation and protection
  - Heritage asset management activity
  - Casework and consultation
  - Research and analysis
  - Historical event
- Information Sources
  - Archive and bibliography

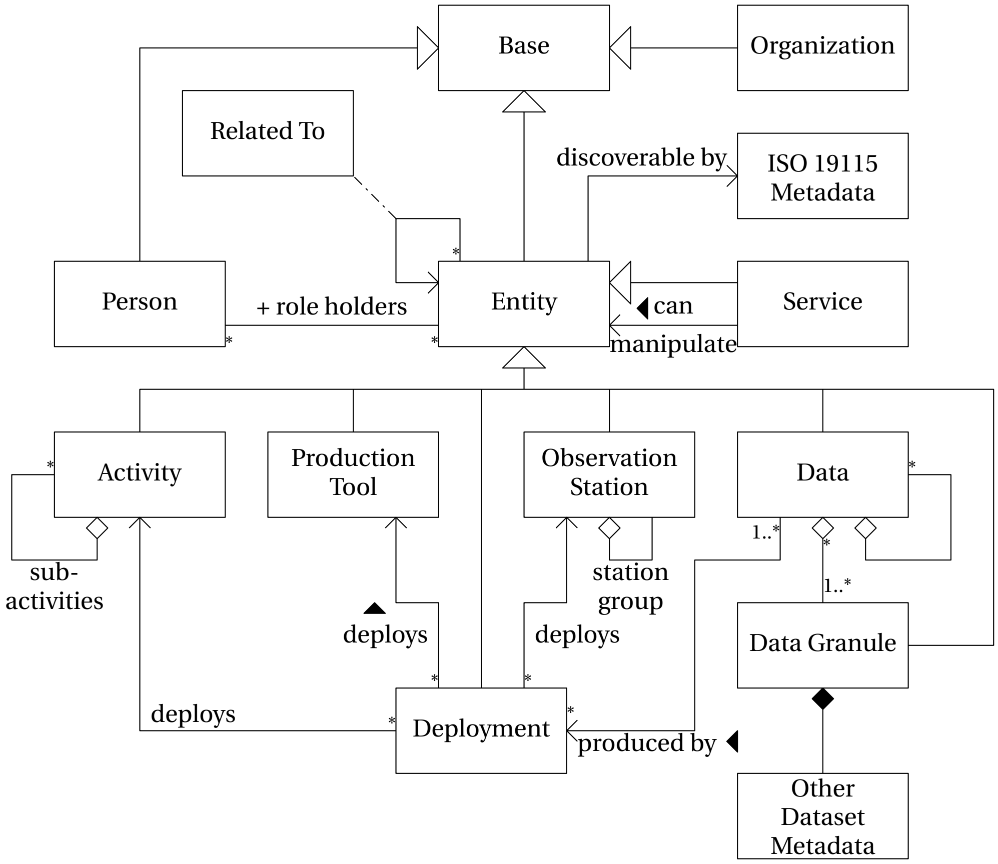

Figure 4.2: High level data model from the MOLES metadata profile (adapted from Lawrence et al. [Law+09]).

  - Narrative and synthesis
  - Management activity documentation
- Spatial Information
  - Location
  - Map depiction
- Temporal Information
  - Date and period
- Actor Information
  - Actor and role

The scope of this standard is clearly wider than just research data, but in MIDAS terms, a dataset would be an archive and bibliography object, set in the context of an investigative activity, casework/consultation activity or research and analysis activity. It would have a particular spatial and temporal coverage, be about a particular physical asset, and various actors would have responsibility for it in various ways. It would also be connected with other archive/bibliography or narrative/synthesis objects, that is, published papers and so on.

4.3 MOLES Data Model

The MOLES metadata profile, which underlies the NERC Data Grid, uses a data model with five key entities (see Figure 4.2). The *Deployment* entity represents a data gathering exercise, and links four other entities. The *Activity* entity represents the study or project under whose auspices the data gathering exercise is performed. The *Observation Station* entity represents the location(s) where the data is gathered, and is also used to record the researchers generating the data. The *Production Tool* represents the instrument and/or methodology used to collect the data. Finally, the data itself is represented by a *Data* entity; this entity is made up either of further Data entities, or of one or more *Data Granule* entities, themselves essentially atomic datasets. MOLES allows for the metadata for Data Granule entities to be supplemented with more specific (archiving-related) metadata in a different scheme, such as CSML [Woo+06].

One further entity, *Service*, represents a process which can manipulate the other high-level entities to produce either new high level entities or text/visualizations. All high-level entities may be described using metadata from another scheme such as DIF or ISO 19115 [GCM08a; ISO03], and their inter-relationships clarified by *Related To* associations.

4.4 CCLRC Scientific Metadata Model

The Scientific Metadata Model (SMDM) devised by STFC (formerly CCLRC) has a number of different models associated with it. The scientific activity model has four levels of entity. *Policy* is a governmental or company policy that drives research by initiating one or more *Programmes* of work. Each Programme represents a tranche of funding for studies or projects on a particular theme or topic. A *Study* is a piece of work performed by a principal investigator and/or institution, along with co-investigators and researchers. A Study is typically funded by a Programme, and therefore may have a grant number associated with it. An *Investigation* is a data collection exercise performed as part of a Study. The model explicitly recognizes three types of Investigation, although it leaves room for others.

- An *Experiment* typically consists of a controlled environment, where an instrument is used to measure one property of the environment or specimen while other properties are set to a known value or a series of known values.
- A *Measurement* is typically produced by a passive detector that records the state of the environment at specified intervals of time and space.
- A *Simulation* takes a mathematical model of a system, and from a set of initial parameters either calculates what a further set of parameters must be, or determines how the modelled system evolves over time.

The model also recognizes the existence of *Virtual Studies*. These are groups of studies that are related in some way – typically having the same principal investigator/institution and subject matter – other than belonging to the same Programme. Common examples are studies where one is the follow-on of the the other.

The collected data itself is covered in a different model. Each Investigation produces exactly one *Data Holding*. This Data Holding is made up of one or more *Data Collections*, each of which may be divided into further Data Collections. The concept of a Data Collection enables different sequences of data to be separated out, e.g. the raw data instrument data from the intermediate and processed sets of data. Each Data Collection is ultimately represented by a set of *Atomic Data Objects:* physical data files or database queries from which the data may be obtained.

These two models are combined and illustrated in Figure 4.3

4.5 ISO 2146

ISO 2146, *Information and documentation – Registry Services for Libraries and Related Organizations,* is a Draft International Standard that defines an information model for facilitating the collaborative construction of registries of all types [ISO05]. It underlies the Online Research

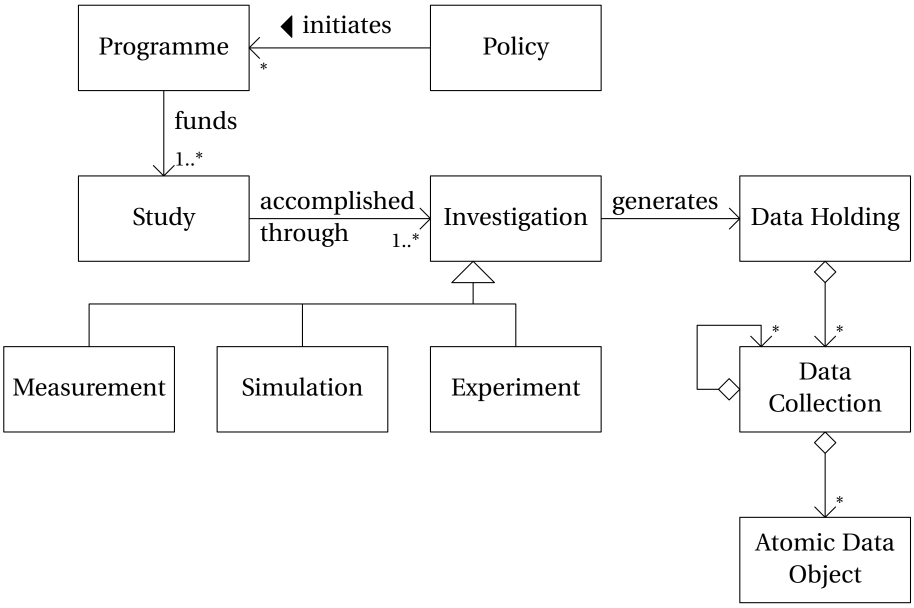

Figure 4.3: High level entity model from the CCLRC Scientific Metadata Model.

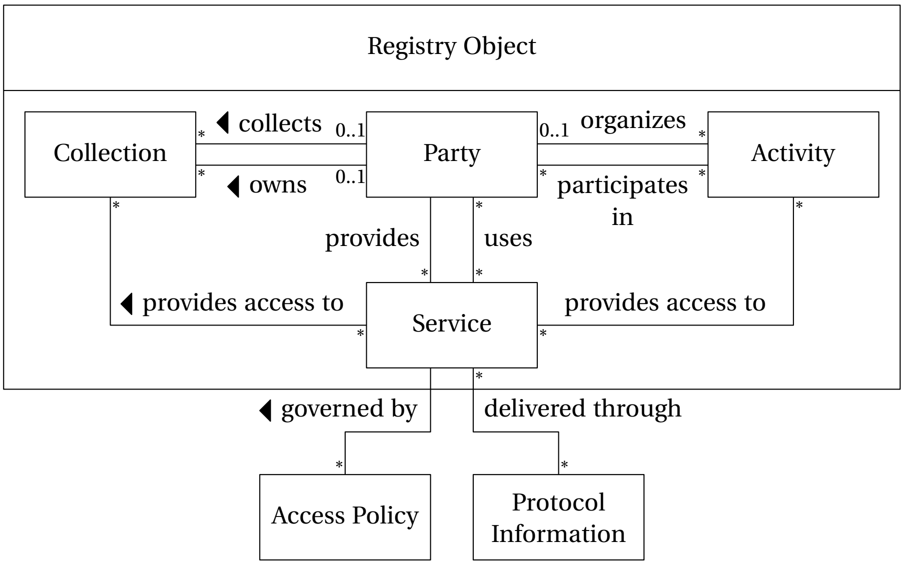

Figure 4.4: High level entity model from the Draft International Standard ISO 2146. Relationships are not predefined by the standard; those shown are examples.

Collections Australia (ORCA) Registry, which will be enhanced for use as the Australian National Data Service (ANDS) national collections registry [Cat08; TW08].

The model defines four types of *Registry Object:* Activities, Collections, Parties and Services. *Activities* occur over time and generate one or more outputs. *Collections* are aggregations of physical or digital objects, considered as a unit for business purposes. *Parties* are people or groups performing a role in relation to the business of a specific community or domain, such as owning a collection, providing a service or performing an activity. *Services* are systems that provide functions of value to an end user, such as providing access to collections or activities. The model also defines an additional three entities that describe how a service is used. *Access Policies* define who may access a service, what functions and resources they can access and what conditions and obligations apply, while other policies may be described using a *Service*

Description entity. *Protocol Information* describes a protocol supported by a service and level at which the service conforms to the protocol. These entities and the relationships between them are illustrated in Figure 4.4.

The metadata associated with these entities in the standard are not specifically aimed at discovery, but do contain some pertinent information such as subject, spatial and temporal coverage, and information supporting access/delivery.

4.6 FRBR

In 1990, the Stockholm Seminar on Bibliographic Records called for a study of the functions performed by bibliographic records with respect to different media, applications and users. This was in response to increasing pressure to reduce cataloguing costs and to allow greater sharing and interoperability of records through standardization. The International Federation of Library Associations (IFLA) formally commissioned the study in 1992, and *Functional Requirements for Bibliographic Records* (FRBR) was published in 1998 [IFL98].

The methodology of the study included an entity analysis technique, in order to determine what 'things' were described by bibliographic records, and to establish the kinds of relationship that exist between them. The result of this was a domain model consisting of three groups of entities. The Group One entities, characterized as products of intellectual or artistic endeavour, are *Work*, *Expression*, *Manifestation* and *Item*. A work is realized through one or more expressions, each of which is embodied in one or more manifestations, each of which is exemplified by one or more items. As a concrete example, a novel (the work) may be expressed in a number of different languages, or may be expressed in an abridged or extended form. Each expression may be published in several different editions (manifestations), while conversely an omnibus edition collects together several different expressions (of different works). Finally, each edition is made available as a series of physical copies (items).

The Group Two entities, characterized as those agents responsible in some way for the Group One entities, are *Person* and *Corporate Body*. These agents may create works, realize expressions, produce manifestations and/or own items.

While a work may have any entity mentioned so far as its subject, the full range of subject matter is far broader. Group Three entities are intended to represent the remainder of possible subject matter; they are *Concept*, *Object*, *Event* and *Place*.

These entities and the relationships between them are summarized in Figure 4.5.

As the name implies, FRBR models the kinds of objects typically held as library stock, and does not have any special relevance to scientific data. It has, however, been used in modified forms as the domain model for the Dublin Core Application Profiles already commissioned by the JISC, and is included here to explore how mappings could be made, if desired, between those Application Profiles and a Scientific Data Application Profile [Cha09].

4.7 Comparisons

At the finest level, it is possible to draw a rough equivalence between a DDI Physical Instance (with aspects of a DDI Archive) and a FRBR Item. Both represent a single copy of a set of data, having a single location on a storage medium and a unique history in terms of processing operations and responsible parties. The first level of abstraction above this is an entity representing the set of bitwise identical copies of a set of data; at this level there is a rough equivalence between a DDI Physical Data Product (with aspects of a DDI Archive), an SMDM Atomic Data Object and a FRBR Manifestation. Note that while DDI Archive and FRBR Manifestation allow nesting, in order to aggregate files into a single dataset, SMDM achieves this with Data Collections; MOLES achieves it at the next level of abstraction with Data entities. Properties at this level include

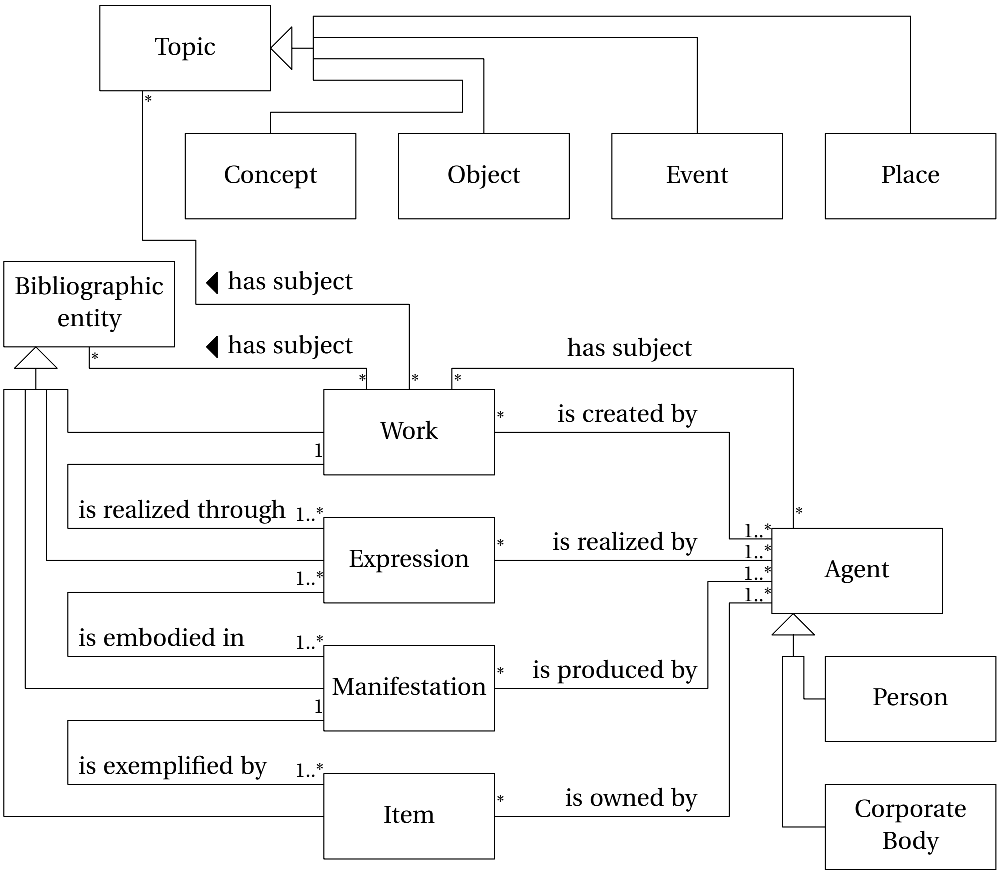

Figure 4.5: FRBR entity model expressed as a UML class diagram. The high level classes 'Bibliographic entity', 'Agent' and 'Topic' are known as Group One, Two and Three entities respectively in the original model.

formatting and encoding, and URLs or database queries that allow a user to retrieve (i.e. make) a copy of the data. The retrieval aspects of this level may be modelled as an ISO 2146 Service.

The next level of abstraction is an entity representing the set of versions of the data that may not be bitwise identical (perhaps because they are in different formats or encodings) but which convey exactly the same information; there is a rough equivalence between a DDI Logical Product (with aspects of a DDI Data Collection), a MOLES Data Granule, an SMDM Data Collection and a FRBR Expression. Properties at this level will include the level of processing – whether the data are raw, fully processed, or in an intermediate stage – and the equipment/ software used to generate/process the data.

The level of abstraction above this is an entity representing the set of versions of the data that may not convey exactly the same information (because one version may be raw data and another version cleaned data) but would normally be considered the same data nevertheless; at this level there is a rough equivalence between a MIDAS Archive and Bibliography, a MOLES Data entity, an SMDM Data Holding, an ISO 2146 Collection and a FRBR Work. DDI does not have a direct equivalent because it does not make a straightforward distinction between a data gathering exercise and the dataset it produces; properties associated with entities mentioned here may be found as aspects of a DDI Data Collection or Conceptual Component.

The context for a dataset is a data gathering exercise; such an exercise may be represented by a DDI Study, a MIDAS Investigative Activity or Research and Analysis Activity, a MOLES Deployment, an SMDM Investigation or an appropriate ISO 2146 Activity. Data gathering exercises usually take place in the context of a project or a long-term data gathering commitment;

these may be represented by a DDI Instance, a MIDAS Casework/Consultation Activity, a MOLES Activity, an SMDM Study or an appropriate ISO 2146 Activity.

4.8 Alternative metadata models

The following models did not arise from the survey of the repository landscape, but may nevertheless be considered suitable inspiration for a domain model for scientific data.

4.8.1 CIDOC CRM

The CIDOC Conceptual Reference Model (CRM) is a formal ontology for interchanging and integrating cultural heritage information, developed by the International Committee for Documentation (CIDOC) and the International Council of Museums (ICOM) [Cro+09].

CIDOC CRM is heavily influenced by RDF, and therefore may be completely expressed in terms of classes and relationships (properties) between them. It models not only entities to be described, but also the entities and values used to describe them. It has two base classes, 'Primitive Value' and 'Entity'. Primitive values include Numbers, Strings and Time Primitives (strings used to identify a temporal extent). The entities defined by the model include Persistent Items (things, people, groups of people), Temporal Entities (events, condition states), Time Spans (abstract temporal extents), Places (abstract spatial extents) and Dimensions (abstract quantities, e.g. of distance, time, or mass). Properties are defined such that the inverse relationships are automatically implied; for example: *is identified by (identifies), carried out by (performed), right held on (has right on),* etc.

The Model does not make explicit provision for datasets, although it does associate them with the Information Object class or a subclass thereof. It is designed to be extensible, though, so a compatible model specifically tailored to the needs of data sets could be constructed. Its principle attraction from the scientific data perspective is the flexibility afforded by its strong RDF flavour; it acts more as a metamodel, providing a system for modelling real world resources.

4.8.2 Open Provenance Model

The Open Provenance Model was an output of the Second Provenance Challenge, a community undertaking arising out of the 2006 International Provenance and Annotation Workshop [Mor+08]. Like CIDOC CRM, it is more of a metamodel than a model of a particular type of thing or system. It defines three entities – *Artifact, Process* and *Agent* – and the following relationships:

- Process *used* Artifact (the role played by the artifact must be specified);
- Artifact *was generated by* Process (the role of the artifact must be specified);
- Process *was controlled by* Agent (the role of the agent must be specified);
- Process *was triggered by* Process;
- Artifact *was derived from* Artifact.

In order to remain as general as possible, the model does not specify additional properties for entities beyond unique identifiers, although it does introduce the notion of temporal annotations in the context of causality graphs. It would clearly provide a flexible way of describing how a set of data came to be produced, but does not provide a complete set of metadata for discovery to delivery purposes.

4.8.3 FuGE

The Functional Genomics Experiment (FuGE) model was developed with two aims: to provide a model of common components within genomics investigations, and to provide a framework for capturing laboratory workflows [Jon+07]. FuGE metadata is arranged into packages of classes and relations with the following structure:

Common Components used to develop models for high-throughput or data intensive experimental processes:

- **Audit** Contacts, auditing and security settings for all objects.
- **Description** Additional annotations and free-text descriptions for all objects.
- **Measurement** Classes for providing measurements within FuGE, including slots for the measurement value, the unit and the data type.
- **Ontology** A mechanism for referencing external ontologies or terms from a controlled vocabulary.
- **Protocol** A model of procedures, software, hardware and parameters. The package can define workflows by relating input and output materials and/or data to the protocols that act on them.
- **Reference** External bibliographic or database references that can be applied to many objects across the FuGE model.

Bio Biological materials, data formats, investigational structure and database representations of bio-molecules, used in an omics experiment:

- **Conceptual Molecule** Captures database entries of biological molecules or other molecule types.
- **Data** Defines the dimensions of data and storage matrices, or references to external data formats.
- **Investigation** Defines an overview of the investigation structure by capturing the overall design and the experimental variables and by providing associations to related data.
- **Material** Models material types such as organisms, samples or solutions, using ontology terms or an extension of the package.

While there are clearly parts of the model that are specific to genomics – most conspicuously metadata in the *Conceptual Molecule* and *Material* packages – much of it is applicable to quantitative research in general.

Chapter 5

Existing metadata standards and profiles

A number of different metadata schemes are already in use by data repositories in the UK and elsewhere. Fifteen of these, arising from the survey of the repository landscape in Chapter 2, have been included with this study as Appendices B to P. These schemes represent a variety of target data types, research communities and organizational contexts, and yet certain similarities may be traced between them. It is on the basis of these similarities that any Scientific Data Application Profile would have to be constructed, not just because they represent a consensus on which elements are most useful in a discovery-to-delivery scenario, but also because they indicate where a Profile could interoperate with existing metadata standards and profiles.

The purpose of this section is to draw out the similarities between the description sets specified by the metadata standards and profiles under consideration, and note any divergence in practice. Included below are metadata elements, or groups of elements, included by at least three out the fifteen schemes; the number in parentheses following the name of the element represents the number of schemes that include that element or a near equivalent.

5.1 Identification

Dataset Identifier (12) Most of the metadata schemes provide a mechanism for recording a locally unique identifier for the dataset. In the majority of cases this serves to identify both the dataset and the metadata record corresponding to it.

Dataset Name (15) All the metadata schemes record at least one name for the dataset. Around half allow alternative names to be recorded, while those with an appropriate information model allow a small hierarchy of names to be used.

Dataset Version (4) Only four of the metadata schemes record a version number for the dataset. This may be because most archives only expect to keep one version of the dataset, or it may be considered redundant when a date of last revision is available.

Dataset Date (13) Only two metadata schemes do not explicitly record a date for the dataset. Among those that do, there is some variety in the types of date recorded: date of creation, date of last modification, date of accession to archive, date of publication, date last checked, date of next review, and date of deletion.

Metadata Record Identifier (3) Three metadata schemes provide a mechanism for explicitly recording a separate identifier for the metadata record itself. The primary usefulness of such an identifier is to enable one to distinguish between multiple different records for the same dataset, perhaps for the purposes of quality control or audience-targeted metadata.

Metadata Record Date (10) Two-thirds of the metadata schemes record one or more dates from the history of the metadata record; most of these record the date of last update, with some recording when the record was first created and a few recording when the record is next due for review. Only DDI (see Appendix C) allows the use of a version number for the record, in addition to a date.

Metadata Scheme Name (7) Just under half of the metadata schemes provide a mechanism for identifying themselves within records using a metadata element. In the majority of cases this is to ease compatibility with the ISO 19115 standard [ISO03].

Metadata Scheme Version (5) Most of those metadata schemes that identify themselves within records also record the metadata scheme version. The two that do not, use a controlled vocabulary for the scheme name and therefore could impose version control using that vocabulary.

5.2 Responsibility

Project/Study/Series Name (9) Nine of the metadata schemes allow a dataset to be placed in the context of a project, study or series. One further scheme, the NEBC profile (see Appendix J), achieves this indirectly by allowing one to specify an associated grant number.

Project/Study/Series Status (4) The status of the associated project or study – that is, whether it is still operating or has been wound up – is only recorded in four of the metadata schemes. It is primarily of use in a data management context, rather than a discovery or delivery context.

Agent (15) All the metadata schemes provide a mechanism for associating people and organizations with the dataset. Some do this using explicit role-based elements, while some use a generic agent element qualified by a role sub-element. Among the many roles provided are: funder, principal investigator, co-investigator, originator/originating centre, project manager, data collector/creator/contributor, researcher, technical contact, depositor, data curator, metadata creator/editor, asset manager and distributor.

Agent Contact Details (11) All but four of the metadata schemes provide a way of recording the contact details (address, telephone number, fax number, e-mail address, Web address) of at least one of the agents associated with the dataset. The four that do not, rely on the catalogue holder having an alternative mechanism for storing contact details.

Rights/Restrictions (14) All the metadata schemes apart from the Cruise Summary Report scheme (see Appendix I) provide a way to record access constraints, use constraints, licensing information, embargo information and/or other pieces of rights-based information. A few of these explicitly provide for machine-readable versions of licensing information.

5.3 Archiving

Location (15) All the metadata schemes provide a way of recording a location where one may find a copy of the data. Most schemes specify a data centre responsible for distributing the dataset, while some record a URL for direct or mediated access to the data, and a few specify a path within the archive file system or the physical location of the media on which the dataset is stored. Four schemes also give additional instructions on how to access the data.

File Format(s) (10) Of the ten metadata schemes that specify a file format for the dataset, two explicitly give details of the available dissemination format(s), rather than the archival format.

Storage Medium (6) Four of the metadata schemes record the type of medium on which the dataset is stored, while three record the available dissemination media instead. One scheme, the EDMED profile (see Appendix H), records both.

Size (7) Seven metadata schemes record the size of digital data sets. Two explicitly require the size be measured in MB, one requires the size to be measured in bytes, one allows kB, MB or GB, and the remainder allow any size of byte unit.

Data quality information (5) The DIF metadata standard (see Appendix B) allows for free text descriptions of data quality and/or quality control procedures. DDI explicitly provides for information on sampling quality, and also allows free text descriptions of other data appraisal processes. AGMAP (see Appendix M) provides for four different data quality statements: logical consistency, completeness, positional accuracy and attribute accuracy. The CCLRC Scientific Metadata Model (see Appendix L) provides for a data quality rating confirmed by certificate, to which a link should be given. The NGDC metadata profile (see Appendix F) allows for a structured statement of data quality, providing elements for the name of a data quality measure, a description of that measure, and the quantitative results of the measurement.

Data Preview (4) Four of the metadata schemes provide for a graphical preview of the dataset for use in searching and browsing contexts. These are the DIF standard and three profiles of ISO 19115 (UK GEMINI, AGMAP and the NGDC profile).

Dataset Language (7) The language of the text within the dataset is recorded in just under half the metadata schemes. In the target context of most of the others, use of English as the dataset language is either mandated or considered a safe assumption.

Dataset Status (8) Eight of the metadata schemes record the status of the dataset – whether it is in its planning stages, in progress or complete. Two further schemes hold this information indirectly, using the dataset completion date.

5.4 Topical Coverage and Derivation

Dataset Type (12) All but three of the metadata schemes record a type for the dataset. Half of these, however, use this element to distinguish databases/data tables from images, text documents, sound recordings and so on. The MOLES profile (see Appendix E) and the CCLRC Scientific Metadata Model distinguish simulations, measurements and analyses/experiments. The NEBC profile distinguishes at the top level between raw -omic data, raw non-omic data, physical holdings, publications and software, although it does have several additional levels of classification. The DDI standard and the UKDA profile (see Appendix K) distinguish surveys, census data, measurement data, demographic data and so on. The eBank profiles establish a level of granularity where crystal structure data is a type.

Subject/Keywords (13) All but two of the metadata schemes – the MOLES Profile and Cruise Summary Reports – record subject headings and/or subject keywords for the dataset. In most cases one or both of these elements use controlled vocabularies.

Abstract/Summary/Description (14) All the metadata schemes allow for a free text summary or description of the dataset, with the exception of the eBank profiles (see Appendix N). The latter is aimed at highly specific data types for which a free text description would not add much useful information. Four metadata schemes have explicit elements for recording the original purpose of the dataset.

Parameters Used (6) Six of the metadata schemes allow a list of the parameters measured, and in some cases controlled, as part of the data creation/collection process. None of them use a controlled vocabulary, although Cruise Summary Reports use a controlled list of data types (e.g. [dissolved] 'Oxygen', 'pH', 'Incident radiation') to help classify parameters, and the DIF standard records parameters as subject terms one level below the lowest for which a controlled vocabulary is available.

Methodology/Instrumentation (8) Just over half the metadata schemes record descriptions of how the data were collected or created. Some have explicit elements for the measurement instrument or platform. A further two metadata schemes prefer that this information is recorded in a separately archived document, to which the metadata record for the dataset can refer.

Processing Steps (6) Six metadata schemes record information about the processing history of the data, of which half have explicit elements for detailing the software used.

Related Datasets (11) All but four of the metadata schemes record relationships with other datasets. Of these, only two do not allow a generic relationship to be specified, while seven provide dedicated elements for particular types of relationships, such as: parent/part, source/derivation, new version/old version.

Derived Publications (11) Eleven of the metadata schemes provide a mechanism for linking the dataset with published material that reports the results. In almost all cases this is achieved by a straightforward textual citation; The MIDAS Heritage standard (see Appendix D) splits the citation across several elements, while the DAF profile (see Appendix P) accepts a URL or other location as well as or instead of a citation.

5.5 Spatiotemporal Coverage

Spatial Extent (12) All but three of the metadata schemes allow the spatial coverage of the dataset to be recorded. All twelve allow the use of area names to define the area; seven allow a latitude/longitude-based bounding rectangle, while five also allow the vertical limits of the area to be recorded. The MIDAS Heritage standard allows a map object to define the area, while Cruise Summary Reports can use numbered grid squares and AGMAP can use a latitude/longitude-based bounding polygon.

Spatial Resolution (7) Just under half of the metadata schemes record the spatial resolution of the dataset. Five schemes use absolute distances or fractions to specify the distance between adjacent data points or the size of the sampling areas, while four can use keywords to represent the smallest geographical unit used as a sampling area (e.g. 'household', 'parish').

Temporal Extent (15) All of the metadata schemes have some indication of the temporal extent of the dataset. With seven of them, this explicitly refers to the time period over which the data were collected, whereas with three of them, it refers to the time period covered by the dataset; five schemes allowed both periods to be recorded. While all the schemes can specify temporal extent by means of start and end dates, only four can specify historical or chronostratigraphic periods.

Temporal Resolution (5) Only five of the metadata schemes allow the temporal resolution of the dataset to be recorded, either by textual description or by specifying an approximate time interval (representing the time between data points or the sample duration). A further three schemes record the frequency with which a dataset is updated.

Chapter 6

Further Discussion

6.1 Implementation

The success of an application profile lies in its adoption by the community it is designed to serve; therefore, any Scientific Data Application Profile must be practical for repositories to adopt. This condition has several implications. First, it must be possible for repositories to collect the information required by the profile, either automatically or by eliciting it from data depositors. Second, it must be possible to store the information required by the profile in a machine-accessible way. Third, it must be possible to serve metadata according to the syntax of the profile, for harvesting by OAI-PMH or through some other means. Fourth, it must be possible for an aggregator to parse and make use of the profiled metadata.

On the first point, institutional commitment is key to ensuring all the required information is collected when the data are deposited. Researchers should be encouraged to assemble the required metadata in the normal course of research, so that submitting the metadata at the time of deposit is not too onerous. The SPECTRa Project identified that a 'golden moment' exists for deposit, where the researcher best understands the data, has the information to describe it and the motivation to deposit it [TM07]; institutional policies and procedures that are in harmony with this are more likely to succeed. Furthermore, refusing to accept datasets without sufficient metadata may encourage more deposits, if deposit in the data repository is thereby seen as a mark of academic achievement.

Any Scientific Data Application Profile must be sensitive to the practicalities of collecting the metadata; collection is likely to be more successful where the required elements can either be extracted automatically or quickly copied from documentation already produced in the course of the research. A possible approach for assisted deposit is exemplified by the SeaDataNet EDMED metadata profile (see Appendix H) where the data depositor fills out a simplified set of metadata fields, and these are later translated into more structured and controlled metadata fields by data curators.

On the second point, the natural place for most Higher Education Institutions to store metadata about data holdings will be in existing repository systems. According to the Directory of Open Access Repositories, the two most popular repository software systems in use are DSpace and EPrints; figures from the Fedora Commons suggest Fedora may be the third most popular [Ope09; SDS08].

DSpace version 1.5.2, current at the time of writing, has a fixed internal data model based on the Dublin Core Libraries Working Group Application Profile [CG04; DSpF; DSpM], which means that any metadata to be stored in the system must be adapted to fit in the fields already in the system. The Edinburgh DataShare profile provides an example of a profile designed within these limitations (see Appendix O). It is likely that this restriction will be removed in the next release of DSpace [DSp09]. The internal data model of EPrints 3 – specifically, the metadata fields of the User Object and EPrint Object – may be modified, although it is considerably easier to do this prior to creating the underlying database than to retrospectively alter an existing database [EPr07]. Note that it would not be possible to alter the data model fundamentally, for example to add a hierarchy of object types, without considerable programming effort. Fedora, in contrast, allows arbitrary XML metadata to be associated with a repository object.

The fixed metadata elements within DSpace represent a limiting factor for a universally adoptable application profile. This means that in order to encourage the most widespread adoption, a Scientific Data Application Profile would have to be representable in terms of DSpace metadata elements. There are two alternatives to this. One is that DSpace repositories only implement a subset of the Profile, leading to incomplete records being harvested. The other is that institutions using DSpace would have to use an alternative system for handling the Scientific Data Application Profile metadata; while not impossible, this would represent a disincentive to adopting the profile at those institutions.

Use cases 6 to 8 in Chapter 3 outline circumstances in which more detailed, discipline-specific metadata are referenced from within the metadata provided by a Scientific Data Application Profile. This raises the issue of how these additional metadata are stored and managed; possibilities include as embedded XML within the metadata record, as a separate item in the repository, or as an external file on a Web site or in another Internet-accessible system.[1] The first of these may be difficult to implement in existing repository software, whereas the others imply an additional administrative burden.

National data centres and subject-specific repositories already store metadata that suits their needs, and realistically will not change what they store to suit a Scientific Data Application Profile. In order to encourage adoption of the Profile among such centres, therefore, the Profile would have to be sufficiently flexible to allow the metadata standards and profiles in use to map to it with a minimum of metadata manipulation.

The third point is largely dependent on the second: provided the required metadata are available in the system, it should be possible to adapt repository software to export them in an appropriate format. For example, EPrints 3.1 introduced a Scholarly Works Application Profile (SWAP) export plugin that allows as much SWAP metadata as can be stored in EPrints about an item to be surfaced appropriately.

On the fourth point, OAI-PMH requires that at least simple Dublin Core metadata are available on request, but allows arbitrary metadata to be requested and returned, so long as it can be expressed in valid XML [OAI08]. There is no fundamental reason why a service such as Repository Search [Int] could not be programmed to handle metadata records returned in an XML expression of a Scientific Data Application Profile.

Across all four of these points, a Scientific Data Application Profile would be easier to implement if written from the perspective of describing a single type of object, rather than a hierarchy of different types. Nevertheless, some complexity, such as associating several different data products with a single dataset record, may be possible.

6.2 Relationship with other application profiles

The review of existing metadata standards and profiles in Chapter 5 shows that spatiotemporal coverage is a common element, and would need to feature in a Scientific Data Application Profile, even if it is not required in all cases. As work has already been done to develop a Geospatial Application Profile (GAP), it would save re-invention if a Scientific Data Application Profile were to incorporate GAP as a component part.

[1] The condition of Internet accessibility is necessary for use cases in which the metadata are retrieved automatically.

Chapter 7

Conclusions

From the preceding discussions, the following conclusions and recommendations may be drawn.

1. There exist some compelling use cases for a Scientific Data Application Profile in the area of discovery-to-delivery. A comparison of data models and metadata schemes from a variety of disciplines suggests that a carefully generalized metadata profile could be constructed that is both widely applicable and yet still fulfils the requirements of the use cases.
2. While the comparison of several different data models shows sufficient common ground for a relatively detailed data model on which to base a Scientific Data Application Profile, from an implementation perspective a simple model is preferred. It is recommended that the primary entity described by a Scientific Data Application Profile should be that described by MOLES as a Data entity and by the CCLRC Scientific Metadata Model as a Data Holding; the entities referred to as Data Granules or Data Collections respectively could be nested within this entity, but preferably not recursively. This is to ease implementation both at the repository level and at the aggregator level. This recommendation does not preclude the use of other entities, but they should be nested within the main entity rather than specified independently, at least at the point of exporting the metadata to an aggregator.
3. The use cases and practicalities of implementation place a number of demands on the description set profile used by a Scientific Data Application Profile, some of which are in conflict:
   - The profile should provide sufficient entry points onto the data to support a wide range of discovery-to-delivery use cases.
   - It should be possible to map between elements of the profile and elements of metadata schemes already in use by data repositories, in particular ISO 19115, Dublin Core and DDI. A corollary of this is that the profile must be able to support multiple controlled vocabularies for some of its elements.
   - It should be possible to record the metadata needed by the profile in most repository systems.
   - Given sufficient preparation, it should be possible for a data depositor to provide the information needed by the profile quickly and easily, possibly with the aid of an intermediary and/or automated tools.
   - It should be possible to associate several different data products with a single dataset record.
   - It should be possible to associate an alternative set or sets of metadata with a single dataset record.
   - The profile should make use of the Geospatial Application Profile for its geospatial metadata elements.

- The profile should include information on intellectual property rights, preferably in a form or forms readable by both humans and machines.

One possible technique to use to achieve this compromise is the notion of levels of compliance, as used in the CCLRC Scientific Metadata Model [SM04]. This allows different requirements to be set for different circumstances without sacrificing interoperability. It may also allow aggregators to deal more cleanly with incomplete records marked at a lower level of compliance.

4. The usage guidelines supporting a Scientific Data Application Profile should advise on how the metadata collected by the Profile can be recorded in the process of an investigation (e.g. at what stage to document a particular piece of information). This may need to be presented from several different disciplinary contexts, and will be more important in disciplines which do not already have a culture of high quality data curation.

5. The usage guidelines supporting a Scientific Data Application Profile may advise, if applicable, on which fields may be suitable for mediated entry, that is, filled out in free text by a data depositor and converted to an encoding or controlled vocabulary by a data curator.

6. The syntax guidelines supporting a Scientific Data Application Profile should provide at least an XML serialization of the profile, for use in OAI-PMH or other XML-based technologies such as Atom newsfeeds.

7.1 Alternatives to constructing a Scientific Data Application Profile

In academic disciplines served by national and global data centres, researchers are well catered for in terms of the discovery and delivery of data pertaining to those fields. A Scientific Data Application Profile would therefore be of principal benefit in interdisciplinary research and generalist data repositories such as those provided by HE institutions.

It is not currently possible to cross-search simultaneously data repositories specializing in different disciplines. Without a Scientific Data Application Profile, any attempt to provide such a service would most likely have to rely on Simple Dublin Core, which is unsatisfactory for the reasons set out in Section 3.6.3. Therefore, effective cross-disciplinary searching would have to be carried out one specialism at a time, incurring significant expenditure of time and effort on the part of researchers familiarizing themselves with the interface and terminology of each data repository or specialist aggregator.

With small, generalist data repositories, the issues for cross-searching are the same but on a greater scale. It is much less feasible to search each institutional repository in turn for relevant data, not least because of the lower likelihood of retrieving useful data in any one search. Thus without aggregation, the data holdings of such repositories would remain largely invisible. As with data centre holdings, aggregation on the basis of Simple Dublin Core would be possible but unsatisfactory.

An additional issue to consider is that institutional data repositories are still in their infancy, and in the absence of a Scientific Data Application Profile, there is a real danger of each repository devising its own metadata profile for data. Not only would this involve considerable duplication of effort, it would also increase the difficulty of providing a data-specific cross-repository search service.

Taken together, these points suggest that failure to construct a Scientific Data Application Profile would have a negative impact on secondary interdisciplinary research and on the development of institutional data repositories.

While the JISC is ideally placed to construct a Scientific Data Application Profile, at least within the UK context, it should be noted that other bodies or projects may construct one if the JISC does not. The DataShare Project (Section 2.5.4) has already produced simple metadata profiles for data, and as mentioned in Section 2.6.2, the Dublin Core Metdata Initiative (DCMI) has set up a Science and Metadata Community to look at general metadata requirements for scientific data. Therefore, as an alternative to the JISC developing a Scientific Data Application Profile independently, a further option would be to collaborate with the DCMI through the Science and Metadata Community to produce such a Profile. Given the DCMI's historic concentration on theoretical aspects of metadata, the JISC could usefully bring to the collaboration a focus on marketing in the broad sense, that is, issues such as implementability, the fulfilment of user requirements, and outreach. The DCMI, in turn, would provide a global perspective.

7.2 Development effort required

From work on previous Dublin Core Application Profiles, it is estimated that the development effort required to produce a Scientific Data Application Profile, along with supporting documentation, will be twelve months full time equivalent. If the work is conducted in partnership with the DCMI, a somewhat reduced amount of effort would be required.

7.3 Community uptake strategy

In order to ensure community uptake, any Scientific Data Application Profile must have the support of key stakeholders. In particular, it must both satisfy a need within the community and be seen to do so. The key stakeholders in the uptake of a Scientific Data Application Profile are:

- repository software developers, who would either have to modify their software to support the Profile, or enable repository managers to adjust their installation to support the Profile;
- institutional repository managers, who would have to implement the Profile for the repository, change the workflows of the repository to suit, and engage with data creators to ensure that they understand the metadata requirements and can satisfy them;
- cross-search tools (e.g. Repository Search), who would have to implement support for the Profile, most likely in a two-stage process as outlined in use case 8;
- data centres, who would have to adjust their systems to support the output of metadata in a form compliant with the Profile;
- researchers producing scientific data, who would have to supply the metadata required by the Profile, and who may gain wider recognition through the wider visibility of their data;
- researchers seeking scientific data, who would use the metadata provided by the Profile to discover and re-use existing data.

Even with community support, uptake of the Profile will slow or non-existent if the barriers for implementing it are too high. The main barriers to the adoption of a Scientific Data Application Profile are:

1. the investment of both time and money required by repositories to implement the Profile in terms of metadata input, and metadata storage;
2. the investment of both time and money required by repositories and data centres to implement the Profile in terms of metadata output;
3. the investment of both time and money required by aggregators to develop a parser and possibly new interfaces to support the Profile;
4. the adjustment of researchers' workflows to ensure the required level of metadata is available for encoding according to the Profile.

The first three of these barriers may be reduced if sufficient priority is given to ensuring the Profile is straightforward to implement at a technical level. The second barrier may be reduced for institutional repositories if the implementation can be performed as part of repository software development instead of local customization. The fourth barrier may be reduced if consideration is given to ensuring the metadata encoded by the Profile are either able to be collected automatically or may easily be recorded in the course of research. In all cases, the barriers may be overcome if sufficient benefit may be argued to come from using the Profile, in terms of the enhanced visibility and discoverability of data.

It is therefore recommended that the author(s) of the Profile should:

1. consult with repository software developers and cross-search providers early on, to determine the technical implications of choosing certain domain models, metadata elements and encodings;
2. hold a multidisciplinary workshop to refine and verify the use cases presented in this report;
3. additionally or alternatively, hold a multidisciplinary workshop to test whether the proposed Profile meets the requirements illustrated in the use cases;
4. work with researchers to produce example workflows in at least two disciplines showing how the metadata encoded by the Profile may be collected;
5. participate in the DCMI Science and Metadata Community, to pursue compatibility between the Profile and other similar efforts globally, and to ensure issues of implementation, user testing and outreach are addressed in that Community as well.

Following completion of the Profile, the author(s) should promote it by publishing papers or articles in appropriate journals and presenting papers at appropriate conferences. The JISC should consider funding the implementation of the Profile in at least one repository software product and at least one cross-search provider.

Appendix A

Presentation of metadata schemes

In the following appendices, metadata schemes are presented as structured lists. These lists may be split into several sections, with each section representing an entity within the associated information model, or else a set of elements used repeatedly within the scheme. Where an element is presented at a greater indent than the previous element, this indicates that the element forms part of the structure of the preceding element.

Each item in the list of metadata elements has the following structure.

- A pair of numbers in square brackets indicates the minimum and maximum number of times, respectively, that the element may occur in a conforming instance of the metadata scheme. A dash indicated that there is no upper bound to the number of times an element may occur. A single number within square brackets indicates that only the minimum number of occurrences is known/specified.
- The element name is given in a sans serif typeface.
- If the element has special formatting rules, this is given in roman type after the element.
  - 'complex' indicates that the fine structure of the element has been omitted for brevity.
  - 'encoded' means that the content of the element must conform to a particular syntax; where possible the syntax is given in parentheses.
  - 'controlled' means that the content of the element must be one of a number of predefined choices; where possible the authority list of choices is given in parentheses.
- Where the meaning of the element is not clear, or where the fine structure of the element is presented elsewhere, additional information is provided *in italics*.

Appendix B

Directory Interchange Format metadata standard

The following is taken from version 2.7.1 of the Directory Interchange Format [GCM08a].

- [1,1] Entry ID
- [1,1] Entry Title
- [1,–] Parameters (Science Keywords): *Hierarchical path of subject headings*
    - [1,1] Category: controlled (GCMD Science Keywords [Ols+07])
    - [1,1] Topic: controlled (GCMD Science Keywords)
    - [1,1] Term: controlled (GCMD Science Keywords)
    - [0,1] Variable Level 1: controlled (GCMD Science Keywords)
    - [0,1] Variable Level 2: controlled (GCMD Science Keywords)
    - [0,1] Variable Level 3: controlled (GCMD Science Keywords)
    - [0,1] Detailed Variable
- [1,–] ISO Topic Category: controlled (profile of ISO 19115 [GCM08b])
- [1,–] Data Center: *Details of the data centre distributing the data*
    - [1,1] Data Center Name
        - [1,1] Short Name: controlled (GCMD Data Centers [Ols+07])
        - [1,1] Long Name: controlled (GCMD Data Centers)
    - [1,1] Data Center URL: encoded (URI)
    - [0,–] Data Set ID
    - [1,–] Personnel
        - [1,1] Role: ‘Data Center Contact’
        - [0,1] First Name
        - [0,1] Middle Name
        - [1,1] Last Name
        - [0,–] Email: encoded (e-mail address)
        - [0,–] Phone: encoded (dash separated blocks, with country code if not US/Canada)
        - [0,–] Fax: encoded (dash separated blocks, with country code if not US/Canada)
        - [0,–] Contact Address
            - [1,–] Address
            - [1,1] City

        - [1,1] Province or State
        - [1,1] Postal Code
        - [1,1] Country
- [1,1] Summary
- [1,1] Metadata Name: controlled (ISO 19115 [ISO03]) ‘CEOS IDN DIF’
- [1,1] Metadata Version: ‘9.7’
- [0,–] Data Set Citation
    - [1,1] Dataset Creator: *Comma-separated names*
    - [1,1] Dataset Title
    - [1,1] Dataset Series Name
    - [1,1] Dataset Release Date: *yyyy-mm-dd encoding preferred*
    - [1,1] Dataset Release Place
    - [1,1] Dataset Publisher
    - [1,1] Version
    - [1,1] Issue Identification: *Volume/issue of publication*
    - [1,1] Data Presentation Form
    - [1,1] Other Citation Details
    - [1,1] Online Resource: encoded (URI)
- [0,–] Personnel
    - [1,1] Role: *e.g. ‘Investigator’, ‘Technical Contact’, ‘DIF Author’*
    - [0,1] First Name
    - [0,1] Middle Name
    - [1,1] Last Name
    - [0,–] Email: encoded (e-mail address)
    - [0,–] Phone: encoded (dash separated blocks, with country code if not US/Canada)
    - [0,–] Fax: encoded (dash separated blocks, with country code if not US/Canada)
    - [0,–] Contact Address
        - [1,–] Address
        - [1,1] City
        - [1,1] Province or State
        - [1,1] Postal Code
        - [1,1] Country
- [0,–] Instrument (Sensor Name)
    - [1,1] Short Name: controlled (GCMD Instruments [Ols+07])
    - [1,1] Long Name: controlled (GCMD Instruments)
- [0,–] Platform (Source Name)
    - [1,1] Short Name: controlled (GCMD Platforms [Ols+07])
    - [1,1] Long Name: controlled (GCMD Platforms)
- [0,–] Temporal Coverage
    - [1,1] Start Date: encoded (yyyy-mm-dd)
    - [0,1] Stop Date: encoded (yyyy-mm-dd)

- [0,–] Paleo-Temporal Coverage: *Used instead of Temporal Coverage when one or both of the years would be negative.*
  - [0,1]
    - [1,1] Paleo Start Date: encoded (number and unit: ybp, ka, Ma, Ga) *Years before ~1950*
    - [1,1] Paleo Stop Date: encoded (number and unit: ybp, ka, Ma, Ga) *Years before ~1950*
  - [0,–] Chronostratigraphic Unit: *Each unit in the list below must be accompanied by the unit above it in the list*
    - [1,1] Eon: controlled (International Union of Geological Sciences [IUGS] geologic time scale)
    - [0,1] Era: controlled (IUGS geologic time scale)
    - [0,1] Period: controlled (IUGS geologic time scale)
    - [0,1] Epoch: controlled (IUGS geologic time scale)
    - [0,1] Stage: controlled (IUGS geologic time scale)
- [0,–] Spatial Coverage
  - [0,1]
    - [1,1] Southernmost Latitude: encoded (number [0,90] as n, $+n$ or nN [North] or $-n$ or nS [South]
    - [1,1] Northernmost Latitude: encoded (number [0,90] as n, $+n$ or nN [North] or $-n$ or nS [South]
    - [1,1] Westernmost Latitude: encoded (number [0,180] as n, $+n$ or nE [East] or $-n$ or nW [West]
    - [1,1] Easternmost Latitude: encoded (number [0,180] as n, $+n$ or nE [East] or $-n$ or nW [West]
  - [0,1] Minimum Altitude: encoded (number and distance unit)
  - [0,1] Maximum Altitude: encoded (number and distance unit)
  - [0,1] Minimum Depth: encoded (number and distance unit)
  - [0,1] Maximum Depth: encoded (number and distance unit)
- [0,–] Location: *Hierarchical path of location names*
  - [1,1] Location Category: controlled (GCMD Locations [Ols+07])
  - [0,1] Location Type: controlled (GCMD Locations)
  - [0,1] Location Subregion1: controlled (GCMD Locations)
  - [0,1] Location Subregion2: controlled (GCMD Locations)
  - [0,1] Location Subregion3
- [0,–] Data Resolution
  - [0,1] Latitude Resolution: encoded (number and unit)
  - [0,1] Longitude Resolution: encoded (number and unit)
  - [0,1] Horizontal Resolution Range: controlled (GCMD Data Resolution Keywords – *see Section B.1.1*)
  - [0,1] Vertical Resolution: encoded (number and unit)
  - [0,1] Vertical Resolution Range: controlled (GCMD Data Resolution Keywords – *see Section B.1.1*)
  - [0,1] Temporal Resolution

  - [0,1] Temporal Resolution Range: controlled (GCMD Data Resolution Keywords – *see Section B.1.2*)

- [0,–] Project
  - [1,1] Short Name: controlled (GCMD Projects [Ols+07])
  - [1,1] Long Name: controlled (GCMD Projects)
- [0,1] Quality
- [0,1] Access Constraints
- [0,1] Use Constraints
- [0,–] Distribution
  - [0,1] Distribution Media: *GCMD Media Keywords [GCM08d] preferred*
  - [0,1] Distribution Size: encoded (number and unit 'KB', 'MB' or 'GB': approximate size)
  - [0,1] Distribution Format: *GCMD Format Keywords [GCM08c] preferred*
  - [0,1] Fees
- [0,–] Data Set Language: *ISO 639-2 [ISO98] vocabulary preferred*
- [0,1] Data Set Progress: controlled ('Planned', 'In Work' or 'Complete')
- [0,–] Related URL
  - [1,1] URL Content Type
  - [1,1] Type: controlled (GCMD URL Content Type vocabulary)
  - [0,1] Subtype: controlled (GCMD URL Content Type vocabulary)
  - [1,–] URL: encoded (URI)
  - [0,1] Description
- [0,1] DIF Revision History: encoded (each line begins with yyyy-mm-dd followed by free text comment)
- [0,–] Keyword (Ancillary Keyword)
- [0,1] Originating Center
- [0,1] Multimedia Sample
  - [0,1] File: *Filename*
  - [0,1] URL: encoded (URI)
  - [0,1] Format
  - [0,1] Caption
  - [0,1] Description
- [0,1] Reference (Publications)
- [0,–] Parent DIF: *Entry ID*
- [0,–] IDN Node
  - [1,1] Short Name: controlled
- [0,1] DIF Creation Date: encoded (yyyy-mm-dd)
- [0,1] Last DIF Revision Date: encoded (yyyy-mm-dd)
- [0,–] Future DIF Review Date: encoded (yyyy-mm-dd)
- [0,1] Private: Boolean

B.1 GCMD Data Resolution Keywords

B.1.1 Spatial resolution

- Point Resolution
- < 1 meter
- 1 meter – < 30 meters
- 30 meters – < 100 meters
- 100 meters – < 250 meters
- 250 meters – < 500 meters
- 500 meters – < 1 km
- 1 km – < 10 km or approximately .01 degree – < .09 degree
- 10 km – < 50 km or approximately .09 degree – < .5 degree
- 50 km – < 100 km or approximately .5 degree – < 1 degree
- 100 km – < 250 km or approximately 1 degree – < 2.5 degrees
- 250 km – < 500 km or approximately 2.5 degrees – < 5.0 degrees
- 500 km – < 1000 km or approximately 5 degrees – < 10 degrees
- > 1000 km or > 10 degrees

B.1.2 Temporal resolution

- < 1 second
- 1 second – < 1 minute
- 1 minute – < 1 hour
- Hourly – < Daily
- Daily – < Weekly
- Weekly – < Monthly
- Monthly – < Annual
- Annual
- Decadal
- Hourly Climatology
- Daily Climatology
- Pentad Climatology
- Weekly Climatology
- Monthly Climatology
- Annual Climatology
- Climate Normal (30-year climatology)

Appendix C

Data Documentation Initiative metadata standard

The following represents a summary of the metadata defined for a DDI Instance by the Data Documentation Initiative (DDI) metadata standard, version 3.0 [DDI08]. Common element attributes, such as ID, Version and the Boolean ‘is maintainable’ flag, are not shown. All elements below are complex (i.e. have additional structure) unless stated otherwise.

- [0,1] Citation: *Publication linked to this instance*
- [0,1] Coverage
  - [0,1] TopicalCoverage *or* -Reference
    - [0,–] Name
    - [0,–] Subject: may be controlled
    - [0,–] Keyword: may be controlled
  - [0,1] SpatialCoverage *or* -Reference
    - [0,–] Name
    - [0,1] BoundingBox: *Bounding longitudes and latitudes*
    - [0,–] Description
    - [0,1] GeographyStructureVariable: *Describes geographic levels available in the data*
    - [0,1] SpatialObject: controlled (‘point’, ‘line’, ‘linear ring’ or ‘polygon’) *Model for the geographic area to which a single data point refers*
    - [0,–] GeographicStructureReference: *Information on the hierarchy of the geographic structure*
    - [0,–] GeographicLocationReference: *Information on the locations covered by the data*
    - [0,–] SummaryDataReference: *Summary data for a particular geographic level*
    - [1,1] TopLevelReference: *Broadest geographic level covered by data*
    - [1,1] LowestLevelReference: *Finest geographic level covered by data*
  - [0,1] TemporalCoverage *or* -Reference
    - [0,–] Name
    - [0,–] ReferenceDate: *Time period covered by data*
- [0,–] Group: *For specifying families/hierarchies of study units*
- [0,–] ResourcePackage: *For specifying shared metadata*
- [0,–] StudyUnit
  - [1,1] Citation: *Publication linked to this study unit*

- [1,–] Abstract
- [1,–] UniversalReference: *Reference to statement detailing the population under study*
- [0,1] SeriesStatement: *Name, description, location of series to which this study belongs*
- [0,–] FundingInformation
    - [1,–] AgencyOrganizationReference: *Funding body*
    - [0,–] GrantNumber: simple string
- [1,–] Purpose
- [0,1] Coverage: *As for instance Coverage*
- [0,–] AnalysisUnit: controlled
- [0,–] AnalysisUnitsCovered: *Explanation of analysis units used*
- [0,–] KindOfData: may be controlled (e.g. ‘survey data’, ‘demographic data’)
- [0,–] OtherMaterial: *As for instance OtherMaterial*
- [0,–] Note: *As for instance Note*
- [0,–] Embargo
    - [0,–] Name
    - [1,1] Date: *Date or date range of embargo*
    - [1,–] Rationale
    - [1,1] AgencyOrganizationReference: *Agent responsible for embargo*
    - [0,–] EnforcementAgencyOrganizationReference
- [0,–] ConceptualComponent
    - [0,1] Coverage: *As for instance Coverage*
    - [0,–] OtherMaterial: *As for instance OtherMaterial*
    - [0,–] Note: *As for instance Note*
    - [0,–] ConceptScheme *or* -Reference: *Comprehensive list of concepts measured by the data*
    - [0,–] UniverseScheme *or* -Reference: *Comprehensive list of populations and sub-populations under study*
    - [0,–] GeographicStructureScheme *or* -Reference: *Information on the hierarchy of the geographic structure*
    - [0,–] GeographicLocationScheme *or* -Reference: *Information on the locations covered by the data*
- [0,–] DataCollection
    - [0,1] Coverage: *As for instance Coverage*
    - [0,–] OtherMaterial: *As for instance OtherMaterial*
    - [0,–] Note: *As for instance Note*
    - [0,1] Methodology
        - [0,–] DataCollectionMethodology
        - [0,–] TimeMethod
        - [0,–] SamplingProcedure
        - [0,–] DeviationFromSampleDesign
        - [0,–] Software: *Software used for data collection*
    - [0,–] CollectionEvent: *When, how, by whom the data was collected*
    - [0,–] QuestionScheme

    - [0,–] ControlConstructScheme
    - [0,–] InterviewerInstructionScheme
    - [0,–] Instrument
        - [0,1] Type: controlled
        - [0,–] Software: *Software package associated with collection instrument*
        - [0,–] ExternalInstrumentLocation: encoded (URI)
        - [0,1] ControlConstructReference: *Control construct initiating sequence of instrument content*
    - [0,–] ProcessingEvent
        - [0,–] ControlOperation
        - [0,–] CleaningOperation
        - [0,–] Weighting
        - [0,–] DataAppraisalInformation
        - [0,–] Coding: *Computation steps, data encoding/recoding rules*
- [0,–] BaseLogicalProduct
    - [0,1] Coverage: *As for instance Coverage*
    - [0,–] DataRelationship: *Descriptions of the relationships between records*
    - [0,–] OtherMaterial: *As for instance OtherMaterial*
    - [0,–] Note: *As for instance Note*
    - [0,–] CategoryScheme *or* -Reference: *Categories used in logical product*
    - [0,–] CodeScheme *or* -Reference: *Encoding of categories, hierarchical relationships between categories*
    - [0,–] VariableScheme *or* -Reference
- [0,–] PhysicalDataProduct
    - [0,–] OtherMaterial: *As for instance OtherMaterial*
    - [0,–] Note: *As for instance Note*
    - [0,–] PhysicalStructureScheme: *File format, most prevalent data field format, how records are collected into files*
    - [0,–] RecordLayoutScheme: *Number and type of records in the data structure*
- [0,–] PhysicalInstance
    - [0,1] Citation: *As for instance Citation*
    - [0,–] Fingerprint: *Digital fingerprint of data file*
    - [0,1] Coverage: *As for instance Coverage*
    - [0,–] OtherMaterial: *As for instance OtherMaterial*
    - [0,–] Note: *As for instance Note*
    - [1,–] RecordLayoutReference: *Number and type of records in the physical instance*
    - [1,–] DataFileIdentification: *Identifies and provides location of data file*
    - [0,1] GrossFileStructure
        - [0,1] PlaceOfProduction: simple string
        - [0,–] ProcessingCheck
        - [0,1] ProcessingStatus: simple string
        - [0,1] CreationSoftware

      - [0,1] CaseQuantity: integer
      - [0,1] OverallRecordCount: integer
    - [0,1] ProprietaryInfo: may be controlled *Key-value information proprietary to creation software*
    - [0,1] Statistics: *Variable and category statistics data documented in physical instance*
  - [0,1] Archive
    - [1,1] ArchiveSpecific: *Access restrictions, funding source, collection description, how data held*
    - [1,1] OrganizationScheme: *All organizations archiving the instance*
    - [0,1] LifecycleInformation
    - [0,–] OtherMaterial: *As for instance OtherMaterial*
    - [0,–] Note: *As for instance Note*
  - [0,–] DDIProfile: *External DDI profile used by study*
  - [0,–] DDIProfileReference: *DDI profile used by study, included internally*
- [0,–] OtherMaterial: *Related material, and how it is related*
- [0,–] Note: *Annotation attached by reference to another item of metadata*
- [0,1] TranslationInformation: *Name, code and catalogue (I18n) of translated language, description of translation*

Appendix D

MIDAS Heritage metadata standard

The following is taken from the 2007 version of MIDAS Heritage [FIS07].

Any entry may declare a relationship to another entry using the following unit of information.

⟶ [0,–] Primary Reference Number Relation: controlled (INSCRIPTION Internal Cross-Reference Qualifiers [FIS04])

All units of information may be qualified by the following unit of information.

⟶ [0,1] Controlled Vocabulary Name

D.1 Heritage Assets

All information groups within this theme have the following units of information.

⟶ [1,1] Primary Reference Number
- ⟶ [1,1] Primary Reference Number Type: controlled (MIDAS)

⟶ [0,–] Heritage Asset Name

⟶ [1,–] Description
- ⟶ [0,–] Description Type: controlled (by local archive)

⟶ [1,1] Compiler (Organization)

⟶ [0,1] Compiler (Person)

⟶ [1,1] Date of Compilation: encoded (dd-MMM-yyyy)

⟶ [1,1] Date of Last Update: encoded (dd-MMM-yyyy)

⟶ [0,1] Entry Type: controlled (by local archive)

⟶ [0,–] External Information System: controlled (by local archive) *Alternative source of documentation*

⟶ [0,–] External Information System Primary Reference Number

D.1.1 Area

⟶ [1,–] Area Type: controlled (INSCRIPTION [FIS04])

⟶ [0,–] Evidence: controlled (INSCRIPTION Evidence Thesaurus [FIS04])

⟶ [0,–] Protection Type: *E.g. 'Conservation Area'*

⟶ [1,–] Map Depiction (Section D.4.2): *Specification of area boundary*

⟶ [0,1] Research and Analysis (Section D.2.5): *Study proposing area boundary*

D.1.2 Monument

⟶ [1,–] Monument Type: controlled (English Heritage Thesaurus of Monument Types, Defence of Britain Thesaurus [FIS04])

   └⟶ [1,–] Date and Period (Section D.5.1)

⟶ [0,–] Currency: controlled ('alternate', 'former' or 'current')

⟶ [0,–] Evidence: controlled (INSCRIPTION Evidence Thesaurus [FIS04])

⟶ [0,–] Material: controlled (RCHME Thesaurus of Building Materials [FIS04])

⟶ [0,–] Material Component Note

⟶ [0,–] Material Name

⟶ [0,–] Component: controlled (English Heritage Components Thesaurus [FIS04])

⟶ [0,1] Prime Motive Power: controlled (by local archive)

⟶ [0,–] Craft Type: controlled (English Heritage Thesaurus of Maritime Craft Types, English Heritage Historic Aircraft Thesaurus [FIS04]) *Mandatory for wrecks*

⟶ [0,1] Departure (Place): *For wrecks*

⟶ [0,1] Destination: *For wrecks*

⟶ [0,1] Manner of Loss: controlled (INSCRIPTION [FIS04]) *For wrecks*

⟶ [0,1] Nationality: controlled (by local archive) *For wrecks*

⟶ [0,1] Registration Place: *For wrecks*

⟶ [0,–] Associated Goods: controlled (by local archive)

⟶ [0,–] Construction Method: controlled (by local archive)

⟶ [0,–] Registration Place

⟶ [0,–] Associated Goods

⟶ [0,–] Construction Method

⟶ [0,–] Protection Type

⟶ [0,–] Right Note: *Legal constraints*

⟶ [0,–] Right Type: controlled (by local archive)

⟶ [0,–] Dimension: controlled (by local archive)

⟶ [0,–] Dimension Measurement Unit

⟶ [0,–] Dimension Value

⟶ [0,1] Condition: controlled (REP93 Condition [FIS04])

⟶ [0,1] Condition Date: encoded (dd-MMM-yyyy)

⟶ [0,1] Inscription

⟶ [0,1] Inscription Note

⟶ [1,1] Location (Section D.4.1)

⟶ [0,–] Investigative Activity (Section D.2.1)

⟶ [0,–] Map Depiction (Section D.4.2)

⟶ [0,–] Monument: *Related structure*

⟶ [1,–] Archive and Bibliography (Section D.3.1): *Descriptions of monument*

⟶ [0,–] Actor and Role (Section D.6.1): *Owners, architects, etc.*

D.1.3 Artefact and ecofact

⟶ [1,–] Artefact/Ecofact Type: controlled (mda Archaeological Objects Thesaurus [FIS04])

⟶ [0,–] Modification State: controlled (by local archive)

⟶ [0,1] Condition: controlled (REP93 Condition [FIS04])

⟶ [0,1] Condition Date: encoded (dd-MMM-yyyy)

⟶ [0,1] Recovery Method: controlled (by local archive)

⟶ [0,1] Artefact Name Type: controlled (by local archive)

⟶ [1,1] Recovery Purpose

⟶ [0,–] Production Method: controlled (by local archive)

⟶ [0,–] Production Technique: controlled (by local archive)

⟶ [0,–] Dimension: controlled (by local archive)

⟶ [0,–] Dimension Measurement Unit

⟶ [0,–] Dimension Value

⟶ [0,1] Evidence: controlled (INSCRIPTION Evidence Thesaurus [FIS04])

⟶ [0,1] Completeness: controlled (‘Complete’, ‘Incomplete’ or ‘Fragmented’)

⟶ [0,1] Conservation Treatment Priority: controlled (by local archive)

⟶ [0,1] Environmental Condition Note

⟶ [0,–] Inscription Content

⟶ [0,–] Inscription Note

⟶ [0,–] Material: controlled (RCHME Thesaurus of Building Materials [FIS04])

⟶ [0,–] Material Component: controlled (by local archive)

⟶ [0,–] Material Component Note

⟶ [0,–] Material Name

⟶ [0,–] Component: controlled (English Heritage Components Thesaurus [FIS04])

⟶ [0,1] Collection Extent

⟶ [0,1] Storage Location

⟶ [0,–] Protection Type

⟶ [0,–] Right Note

  └⟶ [0,–] Right Type: controlled (by local archive)

⟶ [1,1] Investigative Activity (Section D.2.1): *Activity that uncovered the artefact/ecofact*

⟶ [1,–] Date and Period (Section D.5.1): *Date of manufacture, deposition, etc.*

⟶ [0,–] Research and Analysis (Section D.2.5)

⟶ [0,1] Monument (Section D.1.2)

⟶ [0,–] Artefact and Ecofact

⟶ [0,–] Heritage Asset Management Activity (Section D.2.3): *Conservation work, etc.*

⟶ [0,–] Management Activity Documentation (Section D.3.3)

⟶ [0,–] Actor and Role (Section D.6.1): *Owners, manufacturers, curators etc.*

D.2 Activities

All information groups within this theme have the following units of information.

⟶ [1,1] Primary Reference Number

  └⟶ [1,1] Primary Reference Number Type: controlled (MIDAS)

⟶ [0,–] Activity Name

⟶ [1,–] Description

  └⟶ [0,–] Description Type

⟶ [1,1] Compiler (Organization)

⟶ [0,1] Compiler (Person)

⟶ [1,1] Date of Compilation: encoded (dd-MMM-yyyy)

⟶ [1,1] Date of Last Update: encoded (dd-MMM-yyyy)

⟶ [0,–] External Information System: controlled (by local archive) *Alternative source of documentation*

⟶ [0,–] External Information System Primary Reference Number

D.2.1 Investigative activity

⟶ [1,–] Activity Type: controlled (ALGAO Event Types [FIS04])

  └⟶ [1,–] Date and Period (Section D.5.1)

⟶ [0,1] Activity Objective

⟶ [0,1] Work Status: controlled (by local archive)

⟶ [1,–] Archive and Bibliography (Section D.3.1): *Documentation of the activity*

⟶ [1,1] Location (Section D.4.1)

⟶ [0,1] Map Depiction (Section D.4.2)

⟶ [1,–] Actor and Role (Section D.6.1): *Organization/people undertaking the activity*

D.2.2 Designation and protection

⟶ [1,1] Statutory Name

⟶ [1,1] Statutory Name

⟶ [1,1] Entry Type: controlled ('Site' or 'Collection'

⟶ [1,1] Protection Type: controlled (INSCRIPTION [FIS04])

⟶ [0,1] Protection Grade: controlled (National Trust SMR Protection Grade/Status list [FIS04])

⟶ [1,1] Protection Start Date: encoded (dd-MMM-yyyy)

⟶ [0,1] Protection End Date: encoded (dd-MMM-yyyy)

⟶ [1,–] Casework and Consultation (Section D.2.4)

⟶ [0,–] Area (Section D.1.1)

⟶ [0,–] Location (Section D.4.1): *Administrative location of designated area*

⟶ [1,–] Archive and Bibliography (Section D.3.1): *Documentation of designated area*

⟶ [0,1] Management Activity Documentation (Section D.3.3): *Statutory management agreements*

⟶ [1,1] Actor and Role (Section D.6.1): *Authority responsible for designation*

D.2.3 Heritage asset management activity

⟶ [0,1] Entry Type: controlled (by local archive)

⟶ [0,1] Management Activity Method

⟶ [1,–] Management Activity Type: controlled (by local archive)

⟶ [1,1] Work Status

⟶ [0,–] Management Activity Documentation (Section D.3.3)

⟶ [1,–] Date and Period (Section D.5.1)

⟶ [0,–] Actor and Role (Section D.6.1): *Those involved, stakeholders*

An entry must cross-reference the Heritage Assets affected by the activity.

D.2.4 Casework and consultation

⟶ [1,–] Management Proposal Type: controlled (ALGAO Consultation Types list [FIS04])

⟶ [1,1] Notification Date: encoded (dd-MMM-yyyy)

⟶ [1,1] Management Proposal Work Proposed: controlled (ALGAO Work Proposed list [FIS04])

⟶ [1,1] Management Proposal Recommendation: controlled (INSCRIPTION [FIS04])

⟶ [1,1] Case Status: controlled (by local archive)

⟶ [1,1] Management Proposal Outcome: controlled (ALGAO Consultation Outcome List, ALGAO Final Outcome list [FIS04])

⟶ [0,1] Authorization Required: *Type or source of authorization required*

⟶ [0,–] Actors and Role (Section D.6.1): *Applicants, etc.*

⟶ [0,1] Archive and Bibliography (Section D.3.1): *Proposal document*

⟶ [0,–] Investigative Activity (Section D.2.1): *Results of casework*

⟶ [0,–] Designation and Protection (Section D.2.2): *Results of casework*

⟶ [0,–] Heritage Asset Management Activity (Section D.2.3): *Results of casework*

⟶ [0,–] Research and Analysis (Section D.2.5): *Results of casework*

An entry must cross-reference the Heritage Assets affected by the proposal.

D.2.5 Research and analysis

⟶ [0,1] Entry Type: controlled (by local archive)

⟶ [1,–] Activity Type: controlled (ALGAO Event Types [FIS04])

⟶ [0,1] Activity Objective

⟶ [0,1] Work Status

⟶ [1,–] Recovery Method

⟶ [0,–] Potential (Key Item Flag): Boolean

⟶ [1,–] Potential (Note)

⟶ [1,1] Date and Period (Section D.5.1)

⟶ [1,–] Archive and Bibliography (Section D.3.1): *Detailed results, research documentation*

⟶ [1,–] Actor and Role (Section D.6.1): *Researchers*

An entry must cross-reference the Heritage Assets studied by the research.

D.2.6 Historical event

⟶ [1,1] Historical Event Type: controlled (INSCRIPTION [FIS04])

⟶ [1,1] Date and Period (Section D.5.1)

⟶ [0,–] Location (Section D.4.1)

⟶ [0,–] Narrative and Synthesis (Section D.3.2)

⟶ [0,–] Actor and Role (Section D.6.1): *Historical figures involved*

An entry may cross-reference the Heritage Assets affected by the event.

D.3 Information Sources

All information groups within this theme have the following units of information.

- → [1,1] Primary Reference Number
  - → [1,1] Primary Reference Number Type: controlled (MIDAS)
- → [0,–] Information Source Title
- → [1,1] Description
  - → [0,1] Description Type
- → [1,1] Compiler (Organization)
- → [0,1] Compiler (Person)
- → [1,1] Date of Compilation: encoded (dd-MMM-yyyy)
- → [1,1] Date of Last Update: encoded (dd-MMM-yyyy)

D.3.1 Archive and bibliography

- → [0,1] Entry Type: controlled (by local archive)
- → [0,–] External Information System: controlled (by local archive) *Alternative source of documentation*
- → [0,–] External Information System Primary Reference Number
- → [0,1] Statement of Responsibility
- → [1,1] Archive/Source Type: controlled (INSCRIPTION Archaeological Archive Types, National Trust SMR Thesaurus of Resource Description [FIS04])
- → [1,1] Archive/Source Location
- → [0,–] Archive/Source Reference
- → [0,1] Archive Extent
- → [1,1] Archive/Source Format: controlled (National Trust SMR Thesaurus of Resource Description)
- → [0,–] Subject: controlled (UK Archival Thesaurus, etc. [FIS04])
- → [1,1] Date of Origination: encoded (dd-MMM-yyyy)
- → [0,–] Language: controlled (ISO 639-2 [ISO98])
- → [1,1] Right Note
- → [1,1] Right Type: controlled (by local archive)

D.3.2 Narrative and synthesis

- → [0,1] Entry Type: controlled (by local archive)
- → [0,–] External Information System: controlled (by local archive) *Alternative source of documentation*
- → [0,–] External Information System Primary Reference Number
- → [0,1] Statement of Responsibility
- → [1,–] Audience: controlled (by local archive)
- → [0,1] Educational Level: controlled (by Government)

⟶ [1,1] Narrative Text: *The full text itself*

⟶ [1,–] Subject: controlled (UK Archival Thesaurus, etc. [FIS04])

⟶ [0,–] Language: controlled (ISO 639-2 [ISO98])

⟶ [1,1] Right Note

⟶ [1,1] Right Type: controlled (by local archive)

⟶ [1,–] Actor or Role (Section D.6.1): *Authors and contributors*

An entry may cross-reference any related Heritage Asset or Activity.

D.3.3 Management activity documentation

⟶ [0,1] Conservation Plan

⟶ [0,1] Maintenance Plan

⟶ [0,1] Occupancy: controlled (by local archive)

⟶ [0,1] Condition: controlled (REP93 Condition [FIS04])

⟶ [0,1] Condition Statement

⟶ [0,1] Condition Date: encoded (dd-MMM-yyyy)

⟶ [0,–] Vulnerability Level: controlled (by local archive)

⟶ [0,–] Agent of Damage: controlled (by local archive)

⟶ [0,1] Statement of Significance

⟶ [0,–] Value Statement

⟶ [0,–] Value Type: controlled (by local archive)

⟶ [0,1] Characterization Statement

⟶ [1,–] Heritage Asset Management Activity (Section D.2.3)

D.4 Spatial Information

D.4.1 Location

⟶ [1,1] Primary Reference Number

  └⟶ [1,1] Primary Reference Number Type: controlled (MIDAS)

⟶ [1,1] Description

  └⟶ [0,1] Description Type

⟶ [1,–] Administrative Area Name: controlled (English Heritage administrative area lists [FIS04])

  └⟶ [1,–] Administrative Area Type: controlled (INSCRIPTION [FIS04])

⟶ [0,1] Currency: controlled (‘alternate’, ‘former’ or ‘current’)

⟶ [0,–] Locality

⟶ [0,–] Named Location: controlled (by local archive)

⟶ [0,–] Map Sheet

⟶ [0,–] Road or Street Name

⟶ [0,–] Number in Road or Street

⟶ [0,–] Post Code

⟶ [0,–] Language: controlled (ISO 639-2 [ISO98]) *Qualifies area/place names and directions*

⟶ [0,–] Geopolitical Area Type: controlled (FIPS PUB 10-4 [FIP94])

- → [0,–] Geopolitical Area Name: controlled (ISO 3166-1 [ISO])
- → [0,–] Cadastral Reference Value
- → [0,–] Cadastral Reference Source: controlled (by local archive)
- → [0,1] Directions
- → [1,1] Grid Reference
- → [0,1] Buffer Zone Width: encoded (number: width or radius in metres)

D.4.2 Map depiction

- → [1,1] Primary Reference Number
  - → [1,1] Primary Reference Number Type: controlled (MIDAS)
- → [1,1] Compiler (Organization)
- → [0,1] Compiler (Person)
- → [1,1] Date of Compilation: encoded (dd-MMM-yyyy)
- → [1,1] Date of Last Update: encoded (dd-MMM-yyyy)
- → [0,–] External Information System: controlled (by local archive) *Alternative source of documentation*
- → [0,–] External Information System Primary Reference Number
- → [0,1] Data Capture Process
- → [1,1] Positional Accuracy
- → [0,1] Quality
- → [1,1] Spatial Feature Type
- → [1,1] X Coordinate
- → [1,1] Y Coordinate
- → [0,1] Precision
- → [0,1] Buffer Zone Width
- → [0,–] Representation Source
- → [0,1] Data Capture Scale
- → [0,–] Actor and Role (Section D.6.1): *Source of data*

D.5 Temporal Information

D.5.1 Date and period

- → [1,1] Primary Reference Number
  - → [1,1] Primary Reference Number Type: controlled (MIDAS)
- → [0,1] Description
- → [0,1] Entry Type: controlled (by local archive)
- → [1,2] *One or both of:*
  - → *Start and end dates preferred*
    - → [1,1] Start Date: encoded (yyyy, negative for BC)
    - → [1,1] End Date: encoded (yyyy, negative for BC)
  - → Period (Name)

- [0,1] Dimension
- [0,1] Dimension Measurement Unit
- [0,1] Dimension Value
- [0,1] Display Date: encoded (dd-MMM-yyyy)
- [0,1] Date Range Qualifier: controlled (‘occasionally’, ‘throughout’, ‘between’, ‘pre’ or ‘post’)
- [0,1] Scientific Date
- [0,1] Scientific Date Method: controlled (ADS Scientific Date Methods [FIS04])

D.6 Actor Information

D.6.1 Actor and role

- [1,1] Primary Reference Number
  - [1,1] Primary Reference Number Type: controlled (MIDAS)
- [0,1] Description
  - [0,1] Description Type
- [1,1] Compiler (Organization)
- [0,1] Compiler (Person)
- [1,1] Date of Compilation: encoded (dd-MMM-yyyy)
- [1,1] Date of Last Update: encoded (dd-MMM-yyyy)
- [0,–] External Information System: controlled (by local archive) *Alternative source of documentation*
- [0,–] External Information System Primary Reference Number
- [0,–] Contact Point
  - [0,–] Contact Point Type: controlled (by local archive)
- [1,–] *One of:*
  - People Name
    - [0,1] Currency: controlled (‘alternate’, ‘former’ or ‘current’)
  - Organization Name
    - [0,1] Currency: controlled (‘alternate’, ‘former’ or ‘current’)
  - Person Name
    - [0,1] Currency: controlled (‘alternate’, ‘former’ or ‘current’)
- [0,1] Occupation
- [1,–] Role: controlled (by local archive)
  - [1,1] Date and Period (Section D.5.1)

Appendix E

MOLES metadata profile

The following corresponds to version 1.3 of the MOLES/NERC DataGrid metadata schema [NER08].

All entities have the following elements:

- [1,1] Metadata ID
  - [1,1] repositoryIdentifier: *Data provider*
  - [1,1] schemeIdentifier: controlled (NDG Identifiers Scheme Vocabulary)
  - [1,1] localIdentifier
- [1,1] Metadata Description
  - [1,1] metadataDescriptionID
  - [1,1] metadataDescriptionLastUpdated: encoded (W3CDTF [WW98])
  - [1,1] abstract
    - [0,1] abstractText
    - [0,–] abstractOnlineReference: encoded (URI)
  - [0,1] descriptionSection: free text or encoded (URI)
- [1,1] name
- [1,1] abbreviation
- [0,1] DataProvenance
  - [1,1] RecordCreation
    - [1,1] CreatedDate: encoded (W3CDTF)
    - [1,1] CreatedBy
  - [0,–] RecordUpdate
    - [1,1] UpdateDate: encoded (W3CDTF)
    - [1,1] UpdatedBy
    - [0,–] UpdateSummary
  - [0,1] RecordReview
    - [1,1] ReviewDate: encoded (W3CDTF)
    - [0,1] ReviewContact
- [0,1] Metadata Security
  - [1,–] Security Condition: complex (Security Condition Type)

E.1 Data Entity

- [1,1] Dataset Type: controlled (‘simulation’, ‘analysis’, or ‘measurement’)
- [1,1] Data Object Type: controlled (see Section E.1.1)
  - [1,–] Input Data Granule ID: *For derived data types only*
- [1,–] Data Granule
  - [1,1] Data Model ID: complex (Metadata ID)
  - [0,–] instance
    - [1,1] URI: encoded (URI)
    - [1,1] format: controlled (by local archive)
  - [0,1] accessControlPolicy
    - [1,1] *One of:*
      - accessControlPolicyURL: encoded (URI)
      - accessControlPolicyText
        - [1,1] accessControlConditions
          - [1,–] accessControlCondition
            - [1,1] effect: controlled (‘allow’ or ‘deny’)
            - [1,1] simpleCondition
              - [1,1] Attribute Authority
              - [1,1] attrauthRole
- [1,1] Data Summary
  - [1,–] Parameter Summary: complex (Parameter Type) *See Section E.7.1*
  - [0,1] Data Coverage: complex (Coverage) *See Section E.7.2*
  - [0,1] Dataset Status
    - [1,1] Dataset Closure: controlled (‘no_data’, ‘complete’, ‘updating’ or ‘incomplete’)
    - [0,1] Update Frequency: complex, controlled
- [0,1] Data Roles
  - [0,1] Data Creator: complex (Role Type)
  - [1,1] Data Curator: complex (Role Type)
  - [0,–] Data Other Roles: complex (Role Type)
- [0,–] RelatedDeployment: complex (Deployment Type)
- [0,–] Related Data Granule ID
  - [1,1] RelatedDataGranuleID: complex (Metadata ID)
  - [1,1] RelationID
- [1,1] Data Production Tool: complex (*see Section E.3*)
- [1,1] Observation Station: complex (*see Section E.4*)

E.1.1 Data Object Types

Basic types:

- sample
- profile
- section
- ensemble

- Langrangian path
- N-dimensional dataset

Derived types:

- climatology
- integration
- timeseries

E.2 Activity

- [0,–] relatedActivity
  - [1,1] activityRelation: complex, controlled
  - [1,1] relatedActivityID: complex (Metadata ID)
- [1,–] *One of:*
  - Activity Data Collection
  - Activity Data Project
  - Activity Data Campaign
  - Activity Data Investigation
    - [1,1] *One of:*
      - Flight
      - Cruise
- [1,1] Activity Role
  - [1,1] Investigator
    - [1,–] Principal Investigator: complex (Role Type)
    - [0,–] CoInvestigator: complex (Role Type)
  - [0,1] Technical Contact: complex (Role Type)
  - [0,1] Project Manager: complex (Role Type)
- [0,–] ActivityDeployment: complex (Deployment Type)
- [0,1] Activity Coverage: complex (Coverage Type)
- [0,1] ActivityDuration
  - [1,1] startDate: complex (W3CDTF)
  - [0,1] endDate: complex (W3CDTF)

E.3 Data Production Tool

- [1,1] contactDetails: complex (Contact Detail Type)
- [1,1] *One of:*
  - Model
  - Instrument: complex, controlled
- [0,1] DPT Roles
  - [0,1] DPT Other Roles: complex (Role Type)
- [0,1] DPT Deployment: complex (Deployment Type)

E.4 Observation Station

- [1,1] contactDetails: complex (Contact Detail Type)
- [1,1] *One of:*
  - Stationary Platform
    - [1,1] position: complex, encoded (numbers representing latitude and longitude)
    - [1,1] *One of:*
      - Land Station
      - Mooring: complex (deployingCruise, dateStart, dateEnd)
      - Station Group: complex
  - Moving Platform
    - [1,1] *One of:*
      - Ship: complex
      - Aircraft: complex
      - Satellite: complex
- [0,–] Observation Station Deployment: complex (Deployment Type)

E.5 Person

- [1,1] name
  - [0,1] title
  - [0,1] knownAs
  - [0,–] personalName
  - [1,1] initials
  - [1,1] familyName
  - [1,1] nameOrder: controlled (‘L2R’ or ‘R2L’)
- [1,1] contactDetails: complex (Contact Detail Type)

E.6 Organization

- [1,1] name
- [1,1] abbreviation
- [1,1] contactDetails: complex (Contact Detail Type)
- [1,1] Organization Role: complex (Role Type)

E.7 Common Types

E.7.1 Parameter Type

- [1,1] isOutput: Boolean
- [1,1] *One of:*
  - Value Data Parameter
    - [1,1] Value

    - [1,1] Standard Unit: complex
    - [0,1] Original Unit: complex
  - Range Data Parameter
    - [1,1] HighValue
    - [1,1] LowValue
    - [1,1] Standard Unit: complex
    - [0,1] Original Unit: complex
  - Enumeration Parameter
    - [1,–] Value
    - [1,1] Standard Unit: complex
    - [0,1] Original Unit: complex
  - Parameter Group
    - [2,–] Component Parameter: complex (Parameter Type)
- [1,–] Standard Parameter Measured: complex, controlled
- [1,1] ParameterName
- [1,1] ParameterAbbreviation
- [0,1] Parameter Level: encoded (integer [0,+∞)) *For grouped parameters*

E.7.2 Coverage Type

- [0,1] Spatial Coverage
  - [0,–] Bounding Box
    - [1,1] LimitNorth: encoded (number [-90, 90] where North is positive)
    - [1,1] LimitSouth: encoded (number [-90, 90] where North is positive)
    - [1,1] LimitEast: encoded (number [-180, 180] where East is positive)
    - [1,1] LimitWest: encoded (number [-180, 180] where East is positive)
  - [0,–] Area: complex, controlled (any URI-citable parameter list)
  - [0,1] Vertical Extent
    - [1,1] Vertical Extent Base Ref: complex, controlled
    - [1,–] *One of:*
      - Vertical Extent Level: complex (number and unit if not metres)
      - Vertical Extent Range: complex (number and unit if not metres)
      - Vertical Extent Text: complex, controlled
  - [0,1] SpatialResolution: complex (keyword, distance value/unit or scale for each dimension)
- [0,1] Temporal Coverage
  - [1,–] *One of:*
    - DateRange
      - [1,1] DateRangeStart: encoded (W3CDTF [WW98])
      - [1,1] DateRangeEnd: encoded (W3CDTF)
    - DateSingle: encoded (W3CDTF)
    - Chronostratigraphic Term: complex, controlled
  - [0,1] TemporalResolution: complex (keyword, time value/unit or scale)

- [0,1] Spatiotemporal Coverage
  - [1,1] Spatiotemporal Range
    - [1,1] Spatial Coverage: *As above*
    - [1,1] Temporal Coverage: *As above*

E.7.3 Deployment Type

- [1,1] DateStart: encoded (W3CDTF)
- [1,1] DateEnd: encoded (W3CDTF)
- [1,1] ActivityID: complex (Metadata ID)
- [1,1] DataProductionToolID: complex (Metadata ID)
- [1,1] ObservationStationID: complex (Metadata ID)
- [0,1] Coverage: complex (Coverage Type)

E.7.4 Role Type

- [1,1] roleName
- [1,1] abbreviation
- [1,–] Role Holder
  - [1,1] *One of:*
    - Organization ID: complex (Metadata ID)
    - Person ID: complex (Metadata ID)
    - Role ID: complex (Metadata ID)
  - [1,1] startDate: encoded (W3CDTF)
  - [0,1] endDate: encoded (W3CDTF)
  - [0,1] localName
- [0,1] contactDetails: complex (Contact Detail Type)

E.7.5 Contact Detail Type

- [0,1] eMail
- [0,1] fax
- [0,1] telephone
- [0,1] address
  - [1,–] addressline
  - [1,1] city
  - [0,1] postcode
  - [1,1] country
- [0,1] URI: encoded (URI)

Appendix F

National Geoscience Data Centre metadata profile

F.1 MD_Metadata

- fileIdentifier
- language
- characterSet: controlled (MD_CharacterSetCode)
- contact.CI_ResponsibleParty: *see Section F.2, role='author'*
- dateStamp: encoded (ISO 8601 [ISO04])
- metadataStandardName
- metadataStandardVersion
- referenceSystemInfo.MD_ReferenceSystem
  - referenceSystemIdentifier.RS_Identifier
    - authority.CI_Citation
      - title
      - date.CI_Date: *see Section F.3*
    - code
- identificationInfo.MD_DataIdentification
  - citation.CI_Citation
    - title
    - alternateTitle
    - date.CI_Date: *see Section F.3*
    - citedResponsibleParty: *see Section F.2*
    - presentationForm: controlled (CI_PresentationFormCode)
    - collectiveTitle
  - abstract
  - purpose
  - status: controlled (MD_ProgressCode)
  - pointOfContact.CI_ResponsibleParty: *see Section F.2, role='pointOfContact'*
  - resourceMaintenance.MD_MaintenanceInformation
    - maintenanceAndUpdateFrequency: controlled (MD_MaintenanceFrequencyCode)

- graphicOverview.MD_BrowseGraphic
  - fileName
  - fileDescription
  - fileType
- resourceFormat.MD_Format
  - name
  - version
- descriptiveKeywords.MD_Keywords
  - keyword
  - thesaurusName.CI_Citation
    - title
    - date.CI_Date: *see Section F.3*
    - citedResponsibleParty: *see Section F.2*
    - collectiveTitle
- resourceConstraints
  - MD_LegalConstraints
    - useLimitation: not used
    - accessConstraints: controlled (MD_RestrictionCode)
    - useConstraints: controlled (MD_RestrictionCode)
    - otherConstraints
  - MD_SecurityConstraints
    - classification: controlled (MD_ClassificationCode)
- aggregationInfo.MD_AggregateInformation
  - aggregateDataSetName.CI_Citation
    - title
    - date.CI_Date: *see Section F.3*
    - identifier.MD_Identifier
      - authority.CI_Citation
      - code
      - series
- spatialRepresentationType: controlled (MD_SpatialRepresentationTypeCode)
- spatialResolution.MD_Resolution
  - equivalentScale
    - MD_RepresentativeFraction
      - denominator
    - distance: complex, controlled (ISO/TS 19103: Distance)
- language
- characterSet: controlled (MD_CharacterSetCode)
- topicCategory: controlled (MD_TopicCategoryCode)
- extent.EX_Extent
  - geographicElement
    - EX_GeographicBoundingBox

```
        extentTypeCode
        westBoundLongitude
        eastBoundLongitude
        southBoundLatitude
        northBoundLatitude
        extentTypeCode
      EX_GeographicDescription
        geographicIdentifier
          MD_Identifier
            authority.CI_Citation: Name of gazetteer
            code
    temporalElement.EX_TemporalExtent
      extent: complex, controlled (ISO 19108: TM_Primitive)
    verticalElement.EX_VerticalExtent
      minimumValue
      maximumValue
      verticalCRS: complex, controlled (gml: VerticalCRS)
  supplementalInformation
distributionInfo.MD_Distribution
  distributionFormat.MD_Format
    name
    version
  transferOptions.MD_DigitalTransferOptions
    onLine.CI_OnlineResource
      linkage
      name
      description
dataQualityInfo.DQ_DataQuality
  scope
    DQ_Scope
      level: controlled (MD_ScopeCode ) ‘dataset’
  report
    result.DQ_QuantitativeResult
      valueUnit: complex, controlled (UnitOfMeasure)
      value: complex, controlled (Record)
    nameOfMeasure
    measureDescription
  lineage.LI_Lineage
    statement
    processStep.LI_ProcessStep
      description
      rationale
```

- dateTime: (ISO 8601)
- source.LI_Source
  - description
  - sourceCitation.CI_Citation
    - title
    - date.CI_Date: *see Section F.3*
    - identifier.MD_Identifier
      - authority
      - code
    - citedResponsibleParty.CI_ResponsibleParty: *see Section F.2*

F.2 CI_ResponsibleParty

- individualName
- organisationName
- positionName
- contactInfo.CI_Contact
  - phone.CI_Telephone
    - voice
    - facsimile
  - address.CI_Address
    - deliveryPoint
    - city
    - administrativeArea
    - postalCode
    - country
    - electronicMailAddress: encoded (e-mail address)
- role: controlled (CI_RoleCode)

F.3 CI_Date

- date: encoded (ISO 8601 preferred)
- dateType: controlled (CI_DateTypeCode)

Appendix G

UK GEMINI metadata profile

The following corresponds to version 1.0 of the UK GEMINI Standard [EGU04].

- ⟶ [1,1] Title
- ⟶ [0,–] Alternative title
- ⟶ [1,–] Dataset language: *ISO 639-2 [ISO98] encoding recommended*
- ⟶ [1,1] Abstract
- ⟶ [1,–] Topic category: controlled (MD_TopicCategoryCode [ISO03])
- ⟶ [1,–] Subject: *May be drawn from Government Categories List*
- ⟶ [1,1] Date: encoded (ISO 8601 [ISO04]) *Data capture period*
- ⟶ [1,1] Dataset reference date: encoded (ISO 8601) *Date of 'publication'*
- ⟶ [0,–] Originator
- ⟶ [0,1] Lineage: *Events/source data used to construct dataset*
- ⟶ [1,1] West bounding coordinate: encoded (number [-180, 180] where East is positive)
- ⟶ [1,1] East bounding coordinate: encoded (number [-180, 180] where East is positive)
- ⟶ [1,1] North bounding coordinate: encoded (number [-90, 90] where North is positive)
- ⟶ [1,1] South bounding coordinate: encoded (number [-90, 90] where North is positive)
- ⟶ [1,–] Extent: controlled (ISO 3166 [ISO]) *Named areas covered by dataset*
- ⟶ [0,–] Vertical extent information
  - ⟶ [1,1] minimum value: encoded (number)
  - ⟶ [1,1] maximum value: encoded (number)
  - ⟶ [1,1] unit of measure: *Qualifies minimum value and maximum value*
  - ⟶ [1,1] vertical datum
- ⟶ [1,1] Spatial reference system: controlled (SpatialReferenceSystem, *see Section G.2*)
- ⟶ [0,1] Spatial resolution: encoded (number: granularity of data in metres)
- ⟶ [0,–] Spatial representation type: controlled (MD_SpatialRepresentationTypeCode [ISO03])
- ⟶ [0,–] Presentation type: controlled (CI_PresentationFormCode [ISO03])
- ⟶ [1,–] Data format: *Possible dissemination format*
- ⟶ [0,–] Supply media: controlled (MD_MediumNameCode [ISO03])
- ⟶ [1,–] Distributor
  - ⟶ [1,1] Distributor Contact title: *Role or position*
  - ⟶ [1,1] Name of distributor
  - ⟶ [0,1] Postal address of distributor

  - → [0,1] Telephone number of distributor
  - → [0,1] Facsimile number of distributor
  - → [0,1] Email address of distributor: encoded (e-mail address)
  - → [0,1] Web address of distributor: encoded (URI)
- → [1,1] Frequency of update: controlled (MD_MaintenanceFrequencyCode [ISO03])
- → [0,–] Access constraint: controlled (MD_RestrictionCode [ISO03])
- → [0,–] Use constraints: controlled (MD_RestrictionCode)
- → [0,1] Additional information source
- → [0,–] Online resource: *From which the dataset can be obtained*
- → [0,–] Browse graphic: *For illustrating the data*
- → [1,1] Date of update of metadata: encoded (ISO 8601)
- → [0,1] Metadata Standard Name
- → [0,1] Metadata Standard Version

G.1 Extent List

Code	Area
CHA	Channel Islands
ENG	England
IOM	Isle of Man
NIR	Northern Ireland
SCT	Scotland
WLS	Wales
EAW	England and Wales
GBN	Great Britain
UKM	United Kingdom

G.2 SpatialReferenceSystem

Domain code	Definition
001	National Grid of Great Britain
002	Irish Grid
003	Irish Transverse Mercator
004	WGS84
011	postcode
012	parish
013	ward
014	electoral constituency
015	census area

016	local authority (county/unitary/district/borough)
017	region
018	country
019	Health Authority area
020	travel-to-work area
021	other area type

Appendix H

SeaDataNet EDMED metadata profile

The following is taken from SEA-SEARCH/EDMED Information Note 2 [SS00]. The apparent duplication in metadata elements reflects the fact that a core set is filled out by the data depositor, and a collation centre uses this information to fill out more structured, encoded or controlled metadata elements.

H.1 Data Holding Centre

⟶ [1,1] Centre-ID: controlled (by Collating Centres)
⟶ [1,1] Collate-ID: controlled (EDMED Table 1 [SS00, p. 10])
⟶ [1,1] Centre-name: *English name and acronym of data centre*
⟶ [0,1] Centre-host: *Native language name and acronym of data centre*
⟶ [1,1] Visit-address
⟶ [1,1] Country: Controlled (EDMED Table 2 [SS00, p. 10], profile of ISO 3166 [ISO])
⟶ [1,1] Centre-website: encoded (URI)
⟶ [1,1] Description: *Status, role, activities, access policies*
⟶ [1,1] Currency-date: encoded (yyyy-mm-dd) *Date of last accuracy check*
⟶ [1,1] Revision-date: encoded (yyyy-mm-dd)
⟶ [1,1] Modify-date: encoded (yyyy-mm-dd) *Date of last collation*

H.2 Data Contact

⟶ [1,1] Contact-ID: controlled (by Collating Centres)
⟶ [1,1] Centre-ID: controlled (by Collating Centres)
⟶ [1,1] Collate-ID: controlled (EDMED Table 1)
⟶ [0,1] Contact-name
⟶ [0,1] Contact-title: *Post title*
⟶ [1,1] Post-address
⟶ [1,1] Phone: encoded (international format phone number)
⟶ [1,1] Fax: encoded (international format phone number)
⟶ [1,1] Email: encoded (e-mail address)
⟶ [1,1] Currency-date: encoded (yyyy-mm-dd) *Date of last accuracy check*
⟶ [1,1] Revision-date: encoded (yyyy-mm-dd)
⟶ [1,1] Modify-date: encoded (yyyy-mm-dd) *Date of last collation*

H.3 Data Set

⟶ [1,1] Dataset-ID: controlled (by Collating Centres)

⟶ [1,1] Centre-ID: controlled (by Collating Centres)

⟶ [1,1] Collate-ID: controlled (EDMED Table 1)

⟶ [1,1] Dataset-name

⟶ [1,1] Time-period: *Date of earliest data, date of most recent data, whether ongoing*

⟶ [0,1] Begin-year: encoded (yyyy) *Earliest data in set*

⟶ [0,1] End-year: encoded (yyyy) *Latest data in set, or year when metadata last reviewed*

⟶ [0,1] Ongoing: controlled ('Yes', or empty or omitted)

⟶ [1,1] Geographic coverage

⟶ [0,1] South: encoded (integer [-90, 90] where North is positive)

⟶ [0,1] North: encoded (integer [-90, 90] where North is positive)

⟶ [0,1] West: encoded (integer (-180, 180] where East is positive)

⟶ [0,1] East: encoded (integer (-180, 180] where East is positive)

⟶ [0,1] Sea-areas: controlled (EDMED Table 5 [SS00, pp. 12–13]) or *Limits of Oceans and Seas* [Ihb]

⟶ [0,1] Area-type: controlled ('S'/'offshore', 'C'/'coastal', 'B'/'offshore and coastal')

⟶ [0,1] Coastal-zone: controlled (EDMED Table 4 [SS00, p. 12], comma-separated)

⟶ [1,–] Project: *Name and acronym*

⟶ [1,1] Project-acronym: *Comma-separated acronyms*

⟶ [1,1] Parameters: *Comma-separated keywords*

⟶ [1,1] Instruments: *Comma-separated keywords*

⟶ [1,1] Data-themes: controlled (EDMED Table 3 [SS00, p. 11], comma-separated)

⟶ [1,1] Summary

⟶ [0,1] Reference: *Citations of further descriptions*

⟶ [0,1] Data-website: encoded (URI)

⟶ [1,1] Originator

⟶ [1,1] Centre: *Title and acronym of centre holding data*

⟶ [1,–] Storage-medium: *Quantity, medium; if digital, size preferably in MB*

⟶ [1,1] Availability

⟶ [0,1] Supply-details: *Form, format, media of dissemination; standard products; availability online*

⟶ [1,1] Contact: *Name, position, contact details*

⟶ [1,1] Contact-ID1: controlled (Centre ID, comma, Contact ID) *Primary contact*

⟶ [0,1] Contact-ID2: controlled (Centre ID, comma, Contact ID) *Secondary contact*

⟶ [1,1] Completed-by: *Name and number/e-mail address*

⟶ [1,1] Currency-date: encoded (yyyy-mm-dd) *Date of last accuracy check*

⟶ [1,1] Revision-date: encoded (yyyy-mm-dd)

⟶ [1,1] Collated-by: *Name of person, name of centre, date of collation*

⟶ [1,1] Modify-date: encoded (yyyy-mm-dd) *Date of last collation*

Appendix I

Cruise Summary Report metadata standard

The following is taken from Cruise Summary Report form [IOC].

- ⟶ [1,1] Ship
  - ⟶ [1,1] Name
  - ⟶ [1,1] Call sign
  - ⟶ [1,1] Type of ship
- ⟶ [1,1] Cruise no./name
- ⟶ [1,1] Cruise period
  - ⟶ [1,1] Start: encoded (dd mm yyyy) *Date set sail*
  - ⟶ [1,1] End: encoded (dd mm yyyy) *Date return to port*
- ⟶ [1,1] Port of departure: *Name and country*
- ⟶ [1,1] Port of return: *Name and country*
- ⟶ [1,1] Responsible laboratory
  - ⟶ [1,1] Name
  - ⟶ [1,1] Address
  - ⟶ [1,1] Country
- ⟶ [1,–] Chief scientist: *Name and laboratory*
- ⟶ [1,1] Objectives and brief narrative of cruise
- ⟶ [0,1] Project
  - ⟶ [1,1] Project name
  - ⟶ [1,1] Coordinating body
- ⟶ [1,–] Principal investigator: *Name and address*
- ⟶ [1,1] Moorings, bottom mounted gear and drifting systems: tabular data (position, instruments, parameters measured, dates
- ⟶ [1,1] Summary of measurements and samples taken: tabular data (amount of data, instruments, parameters measured)
- ⟶ [1,1] Track chart: Boolean
- ⟶ [1,1] General ocean areas: controlled (*Limits of Oceans and Seas* [Ihb])
- ⟶ [0,1] Specific areas
- ⟶ [1,–] Geographic coverage: encoded (integer in range $[1,288] \cup [300,587] \cup [901,936]$: grid square)

Appendix J

NERC Environmental Bioinformatics Centre metadata profile

The following is based on the EnvBase Guidelines for Editing Dataset Records [NEB03].

- [1,1] dataset_name
- [1,1] summary
- [0,1] citation: *Free text or URL*
- [0,1] grant_number
- [1,1] startdate: encoded (dd/mm/yyyy) *Start of research/grant*
- [1,1] enddate: encoded (dd/mm/yyyy) *End of research/grant*
- [1,1] thematic: controlled (NERC directed mode programmes [NEB])
- [1,1] originator
  - [0,1] id: controlled (NEBC contacts database ID)
  - [1,1] name
  - [1,1] contact_details
- [0,1] contributors
  - [1,–] person
    - [0,1] id: controlled (NEBC contacts database ID)
    - [1,1] name
    - [1,1] contact_details
- [0,–] Species
  - [0,1] species_name: controlled (NCBI taxonomy [Whe+00], or ‘Multiple’)
  - [0,1] strain
- [0,–] software_used
  - [1,1] program_name: *Packages used in production or analysis of data*
- [1,1] make_public: Boolean
- [1,1] datacentre: controlled (EnvBase terms [NEB])
- [1,1] contact: E-mail address (contact for whole dataset)
- [1,1] entry_date: dd/mm/yyyy (date of metadata creation)
- [1,1] modified_date: dd/mm/yyyy (date of metadata load)
- [1,1] availability: controlled (EnvBase terms)

- [1,–] holding
  - [1,1] type: controlled (EnvBase terms)
  - [1,–] subtype: controlled (EnvBase terms) *For types ‘Raw -omic’, ‘Physical Holding’ or ‘Publication’*
  - [0,–] subtype: *For other types, EnvBase terms preferred*
  - [1,1] description: *What the data constitutes, how it may be used*
  - [0,1] estimated_size: encoded (number: size in megabytes)
  - [0,1] online_query: encoded (URI) *Third party or personal copy of data*
  - [0,–] storage_location
    - [0,1] file_format
    - [n,1] repository: controlled (EnvBase terms)
    - [0,1] accession: *ID or reference number*
  - [0,1] ipr: *Copyright statement*
  - [1,1] curator
    - [0,1] id: controlled (NEBC contacts database ID)
    - [1,1] name
    - [1,1] contact_details
  - [0,1] how_to_obtain
  - [1,1] submission_status: controlled (‘Proposed’, ‘Exists’, ‘Submitted’, ‘Verified’ or ‘Removed’)
  - [1,1] submission_date: encoded (dd/mm/yyyy)
  - [0,1] expiry_date: encoded (dd/mm/yyyy) *Date when physical holdings will be disposed of*

Appendix K

UK Data Archive metadata profile

- [1] Study Number (Identifier)
- [1] Title
  - [1] Main Title
  - [0] Sub-title/Note field
  - [0] Alternative Title
  - [0] Series Title
  - [0] Project Number
- [1] Distribution
  - [1] Distributor
  - [1] Availability
  - [1] Contact
  - [1] Date of Deposit
  - [1] Date of Distribution
- [1] Subject Category
- [1] Names
  - [1] Depositor
  - [1] Principal Investigator
  - [0] Data Collector
  - [0] Original Data Producer
  - [0] Sponsor
  - [0] Grant Number
  - [0] Other Acknowledgements
  - [0] Research Initiator
- [1] *One or both of:*
  - Abstract
  - Main Topics
- [1] Coverage
  - [1] *One or both of:*
    - Time Period Covered
    - Dates of Fieldwork

    - [1] Country
    - [0] Town/Village
    - [0] Region/County
    - [1] Spatial Unit
    - [0] Observation Unit
    - [0] Kind of Data
- [1] Universe sampled
    - [1] Locations of Units of Observation
    - [1] Population
- [1] Methodology
    - [1] Time Dimensions
    - [1] Sampling Procedures
    - [0] Number of Units
    - [1] Method of Data Collection
    - [1] Weighting
- [1] Level of Processing
- [0] Data Sources
- [1] Language of Written Material
- [0] References to Related Datasets
- [0] References and Publications by PI
- [0] References and Publications by Others
- [0] Administrative Notes
- [1] Access
    - [1] Access Conditions
    - [1] Access Code
    - [0] Recommendations
    - [0] External Note
    - [0] Originating Archive
- [1] Date of Edition
- [1] Copyright Statement
- [1] Hasset keywords: controlled (HASSET [UKD08])

Appendix L

CCLRC Scientific Metadata Model

The following is taken from Sufi and Matthews [SM04].

L.1 Scientific Study Metadata

- [1,1] Topic: *See Section L.2*
- [1,1] Study: *See Section L.3*
- [1,1] Access Conditions: *Ownership, access control, etc.*
- [0,–] Related Material
  - [0,–] Publications: *Reports of study results*
  - [0,–] References: *Standards, etc.*
  - [0,–] Community: *Others working in the same field*
- [0,–] Legal Note: *Copyright Permissions, Patent Licences, etc.*
- [1,1] Metadata Source: *Name of metadata archive*
- [1,1] Metadata ID
- [1,1] Metadata ID Scheme: controlled (by local archive)
- [1,1] Metadata Conformance: controlled (integer [1,5]) *Higher numbers indicate greater conformance to the model*
- [1,1] Metadata Schema

L.2 Topic

- [0,–] Keywords
  - [1,1] Discipline: *Academic area from which keywords drawn*
  - [1,1] Keyword Source: *Pointer to where keywords defined*
  - [1,–] Keyword: controlled
- [1,–] Subjects
  - [1,1] Discipline: *Academic area providing context for subject*
  - [1,1] Subject Source: *Pointer to where subject hierarchy defined*
  - [1,–] Subject: controlled *Further Subject elements nested within to represent path through subject hierarchy*

L.3 Study

- [1,1] Study Name
- [1,1] Study Id
- [0,–] Study Institution
  - [1,1] Name
  - [1,1] Role: controlled (by local archive, e.g. ‘Data Manager’)
  - [1,1] Type: controlled (by local archive, e.g. ‘Academic’)
  - [1,1] Id
- [1,–] Investigator
  - [1,1] Name
  - [1,1] Institution Id
  - [1,1] Contact Details
    - [1,1] Address
    - [1,–] Phone Number
      - [1,1] Number
      - [1,1] Type Of Number: controlled (by local archive, e.g. ‘Fax’)
    - [1,–] E-mail Address: encoded (e-mail address)
    - [1,–] Web Page: encoded (URI)
  - [1,1] Role In Study: controlled (by local archive, e.g. ‘Principal Investigator’)
  - [1,1] Role In Institution
- [1,1] Study Information
  - [1,–] Funding Source
  - [1,–] Time Line: *May also contain elements describing other milestones*
    - [0,1] Start Date: encoded (ISO 8601 [ISO04])
    - [0,1] End Date: encoded (ISO 8601)
  - [1,1] Purpose
    - [1,1] Abstract
    - [1,1] Study Type: controlled (by local archive)
  - [1,1] Status of Study: controlled (by local archive, e.g. ‘Complete’)
  - [1,–] Resources: *E.g. the facility used*
- [0,1] Notes
- [1,–] Investigation: *See Section L.4*

L.4 Investigation

- [1,1] Name
- [1,1] Investigation Type: semi-controlled (‘Experiment’, ‘Simulation’, ‘Measurement’ or free text)
- [1,1] Abstract: *Type and purpose of investigation*
- [1,–] Resources: *E.g. the facility used*
- [1,1] Data Holding: *See Section L.5*

L.5 Data Holding

- [1,1] Data Description
  - [1,1] Data Name
  - [1,1] Type of Data: controlled (‘Collection’ or MIME Type [IAN07])
  - [1,1] Status: controlled (by local archive, e.g. ‘Complete’)
  - [0,–] Data Quality: *Quality rating, link to certification*
  - [1,1] Data Topic: *See Section L.2*
  - [1,1] Logical Description
    - [0,–] Parameter
      - [1,1] Param Name
      - [1,1] Param Id: *For internal cross-referencing*
      - [0,–] Param Ref: *Related parameters*
      - [1,1] Param Class: controlled (‘Qualitative’ or ‘Quantitative’)
      - [1,1] Derivation: controlled (‘Fixed’, ‘Measured’ or ‘Computed’)
      - [0,1] Units
        - [0,1] Unit Name
        - [0,1] Unit Acronym
        - [0,1] Unit System
      - [0,1] Param Value
      - [1,1] Facility Used: *E.g. instrument used*
      - [0,1] Range: controlled/encoded (whether ‘Upper’ or ‘Lower’ limit, margin of error)
      - [0,–] Parameter: *Nesting implies parameter aggregation*
    - [0,1] Time Period: *Creation and last modification of data*
    - [0,1] Start Time: encoded (ISO 8601)
    - [0,1] End Time: encoded (ISO 8601)
    - [0,1] Description
    - [0,1] Facility Used
  - [0,1] Software: *Documentation of software used to create/process the data*
- [0,–] Data Collection
  - [1,1] Data Description: *See Data Holding Data Description*
  - [0,–] Atomic Data Object: *See Data Holding Atomic Data Object*
  - [0,–] Data Collection: *Allows nesting*
  - [0,–] Related Reference: *See Data Holding Related Reference*
  - [0,–] Data Collection Locator: *See Data Holding Locator*
- [0,–] Atomic Data Object: *E.g. file, bitstream*
  - [1,1] Name
  - [1,1] Data Description: *See Data Holding Data Description*
  - [0,–] Related Reference: *See Data Holding Related Reference*
  - [1,–] ADO Locator: *More specific metadata for particular ADO types may be used*
    - [1,1] URI: encoded (URI)

            - [1,1] Size: encoded (integer: size in bytes)
            - [0,1] Offset: encoded (integer: size in bytes)
- [0,–] Related Reference
    - [1,1] Type: *May be controlled*
    - [0,1] Direction: controlled (‘From’, ‘To’ or ‘Peer’)
    - [1,1] Referred To Item: controlled (‘Study’, ‘Investigation’, ‘Data Collection’ or ‘ADO’)
    - [1,1] Method
    - [1,–] Reference Location: *Server, port, etc.*
- [0,1] Data Holding Locator: encoded (URI)
    - [1,1] Data Name
    - [0,1] Locator: *Absolute or relative*

Appendix M

AGMAP metadata profile

The following is taken from Version 1.0 of the UK Academic Geospatial Metadata Application Profile [Go-08].

- [1,1] Citation
  - [1,1] Dataset Title
  - [0,–] Alternative Dataset Title
  - [1,–] Creator
  - [0,1] Identifier
  - [1,1] Edition
  - [1,1] Dataset Date Code: controlled (CI_DateTypeCode [ISO03]: ‘Creation’, ‘Publication’, ‘Revision’ or ‘Deletion’)
  - [1,1] Dataset Event Date: encoded (W3CDTF [WW98])
  - [1,1] Dataset Update Frequency: controlled (MD_MaintenanceFrequencyCode [ISO03])
- [1,1] Identification Information
  - [1,–] Dataset Language: controlled (ISO 639-2 [ISO98]: ‘eng’, ‘cor’, ‘gla’, ‘gle’, ‘wel’ or ‘cym’)
  - [1,–] Dataset Topic: controlled (MD_TopicCategoryCode [ISO03])
  - [1,1] Controlled Vocabulary
  - [1,–] Controlled Keywords: controlled (*see Controlled Vocabulary*)
  - [0,–] Other Keywords
  - [1,1] Description
  - [1,1] Abstract
  - [0,–] Spatial Representation Type: controlled (MD_SpatialRepresentationTypeCode [ISO03])
  - [0,–] Presentation Type: controlled (CI_PresentationFormCode [ISO03])
  - [0,–] Sample: *Graphic for illustrating the data*
  - [0,–] Further Information
  - [0,–] Related Datasets
  - [1,1] Spatial Reference System of the Dataset
    - [1,1] Spatial Reference System: controlled (UK GEMINI Spatial Reference System, *see Section G.2*)
- [0,1] Data Quality Information
  - [0,1] Data Hierarchical Level: controlled (MD_ScopeCode [ISO03], default ‘dataset’)
  - [0,1] Data Process Steps

    - [0,1] Status of the Data Creation Process: controlled (MD_ProgressCode [ISO03])
    - [0,1] Start Date of Data Capture Period: encoded (yyyy-mm-dd, negative for BC)
    - [0,1] End Date of Data Capture Period: encoded (yyyy-mm-dd, negative for BC)
    - [0,1] Dataset Lineage: *Events/source data used to construct dataset*
    - [0,1] Description of Dataset Process Steps: *How the data were generated/derived*
  - [0,1] Data Quality Statement
    - [0,1] Logical Consistency
    - [0,1] Completeness
    - [0,1] Positional Accuracy
    - [0,1] Attribute Accuracy
  - [0,1] Level of Spatial Detail
    - [0,1] Source Scale Denominator: encoded (integer)
    - [0,1] Ground Scale Distance: encoded (number and unit)
    - [0,1] Imagery or Grid x-Dimension Name: controlled (MD_DimensionNameTypeCode [ISO03], default ‘Column’)
    - [0,1] Imagery or Raster Cell/Pixel Size (x- Value): encoded (number and unit)
    - [0,1] Imagery or Grid y-Dimension Name: controlled (MD_DimensionNameTypeCode [ISO03], default ‘Row’)
    - [0,1] Imagery or Raster Cell/Pixel Size (y- Value): encoded (number and unit)
    - [0,1] Smallest Administrative Unit
- [1,1] Extents of Dataset
  - [1,1] Spatial Reference System used for the Bounding Rectangle/Bounding Polygon: controlled (‘British National Grid’, ‘Irish National Grid’ or ‘Latitude and Longitude’)
  - [1,–] Extents of a Dataset Based on the Co-ordinates of a Bounding Rectangle
    - [1,1] West Bounding Co-ordinate: encoded (Grid: integer, in kilometres; Longitude: number [-180, 180] to two decimal places where East is positive)
    - [1,1] East Bounding Co-ordinate: encoded (Grid: integer, in kilometres; Longitude: number [-180, 180] to two decimal places where East is positive)
    - [1,1] North Bounding Co-ordinate: encoded (Grid: integer, in kilometres; Latitude: number [-90, 90] to two decimal places where North is positive)
    - [1,1] South Bounding Co-ordinate: encoded (Grid: integer, in kilometres; Latitude: number [-90, 90] to two decimal places where North is positive)
  - [0,–] Extents of a Dataset Based on the Co-ordinates of a Bounding Polygon
    - [0,–] Co-ordinates of a Bounding Polygon: encoded (comma delimited pair of numbers, giving easting and northing of co-ordinate respectively)
  - [1,1] Extents of a Dataset Based on Geographic Identifiers
    - [1,–] Nations: controlled (UK GEMINI Extent List, *see Section G.1*)
    - [0,–] Administrative Areas: controlled (ISO 3166 [ISO])
    - [0,–] Postcode Districts: controlled (UK postcodes)
    - [0,1] Controlled Place Name Vocabulary
    - [0,–] Controlled Place Name Keywords: controlled (*see Controlled Place Name Vocabulary*)
  - [0,–] Vertical Extents of a Dataset

    - [1,1] Minimum Value: encoded (number)
    - [1,1] Maximum Value: encoded (number)
    - [1,1] Unit of Measure: *Qualifies Minimum Value and Maximum Value*
    - [1,1] Vertical Datum
  - [0,–] Temporal Extents of a Dataset
    - [0,1] Start Date for Time Period Covered by Dataset: encoded (yyyy-mm-dd, negative for BC)
    - [0,1] End Date for Time Period Covered by Dataset: encoded (yyyy-mm-dd, negative for BC)
- [1,1] Custodian
  - [1,1] Name of Custodian
  - [1,1] Postal Street Address of Custodian
  - [1,1] Postal City of Custodian
  - [0,1] Postal County of Custodian
  - [1,1] Postal Code of Custodian
  - [1,1] Postal Country of Custodian
  - [0,1] Telephone Number of Custodian: encoded (international format phone number)
  - [0,1] Facsimile Number of Custodian: encoded (international format phone number)
  - [0,1] Email Address of Custodian: encoded (e-mail address)
  - [0,1] Web Address of Custodian: encoded (URI)
- [1,–] Distributor
  - [1,1] Name of Distributor
  - [0,1] Distributor Contact Title
  - [1,1] Postal Street Address of Distributor
  - [1,1] Postal City of Distributor
  - [1,1] Postal Code of Distributor
  - [1,1] Postal Country of Distributor
  - [0,1] Telephone Number of Distributor: encoded (international format phone number)
  - [0,1] Facsimile Number of Distributor: encoded (international format phone number)
  - [0,1] Email Address of Distributor: encoded (e-mail address)
  - [0,1] Web Address of Distributor: encoded (URI)
  - [0,–] Supply Media: controlled (MD_MediumNameCode [ISO03])
  - [0,1] On-line Linkage: encoded (URI) *Site/online service holding the dataset*
  - [0,1] Dataset File Size: encoded (number: size in megabytes)
- [1,–] Dataset Name and Format
  - [1,1] Dataset Format Name
  - [1,1] Dataset Format Version
- [0,1] Access and Use Constraints
  - [0,–] Access Constraints: controlled (MD_RestrictionCode [ISO03])
  - [0,–] Use Constraints: controlled (MD_RestrictionCode)
  - [0,–] Use Constraints Details
  - [0,–] Other Constraint Details

- [1,1] Metadata Creator
  - [1,1] Name of Metadata Creator
  - [1,1] Postal Street Address of Metadata Creator
  - [1,1] Postal City of Metadata Creator
  - [1,1] Postal Code of Metadata Creator
  - [1,1] Postal Country of Metadata Creator
  - [0,1] Telephone Number of Metadata Creator: encoded (international format phone number)
  - [0,1] Facsimile Number of Metadata Creator: encoded (international format phone number)
  - [0,1] Email Address of Metadata Creator: encoded (e-mail address)
  - [0,1] Web Address of Metadata Creator: encoded (URI)
  - [0,1] Metadata Record Identifier
  - [0,–] Parent Metadata Record Identifier
  - [1,1] Metadata Last Updated: encoded (W3CDTF)
  - [0,1] Metadata Standard Name
  - [0,1] Metadata Standard Version

Appendix N

eBank UK metadata profile

N.1 Version 4

The following is taken from Duke and Heery [DH04].

⟶ [0,1] Data Name

⟶ [0,1] EPrint_type: ‘Crystal structure’

⟶ [0,1] Authors

⟶ [0,1] Affiliations

⟶ [0,1] Formula_empirical: encoded (atom symbols with total count subscript)

⟶ [0,1] Compound_name: *IUPAC Chemical name*

⟶ [0,1] CCDC_Code: *Cambridge Structural Database identifier*

⟶ [0,1] Compound_class

⟶ [0,1] Available_data: Boolean

⟶ [0,1] Related_publications

⟶ [0,1] Publication_date

⟶ [0,1] Last_revised_date

⟶ [0,1] Keywords

⟶ [0,1] Scheme: encoded (SMILES)

⟶ [0,1] InChI: encoded (InChI)

N.2 Version dated 2 November 2005

The following is taken from Koch, Duke and Coles [KDC05].

⟶ [1,1] Title

⟶ [1,–] Creator: encoded (last name, first name, initials)

⟶ [0,–] Subject and Keywords: controlled (eBank profile of IUCr World Directory of Crystallographers)

⟶ [1,1] Subject and Keywords (InCHi): encoded (InChi)

⟶ [1,1] Subject and Keywords (Chemical Formula): encoded (atom symbols with total count subscript)

⟶ [1,1] Subject and Keywords (Compound Class): controlled (eBank Compound Classes)

⟶ [1,1] Publisher

- [1,1] Date modified: encoded (W3CDTF [WW98])
- [1,1] Date created: encoded (W3CDTF)
- [1,1] Resource type: ‘Crystal structure data holding’
- [1,1] Resource Identifier: encoded (URI) *Crystal Structure Report URL*
- [1,1] Resource Identifier: encoded (DOI)
- [0,–] Relation
  - [0,–] Is Referenced By: *Article, etc. that cites the resource*
  - [1,–] Has Part: *Data files making up the resource*
- [0,1] Rights Management

N.3 SPECTRa version

The following is taken from Tonge and Morgan [TM07].

- [1,1] Title
- [1,–] Creator: encoded (last name, first name, initials) *Data owner*
- [1,–] Contributor: encoded (last name, first name, initials) *Spectroscopist/crystallographer*
- [0,–] Subject and Keywords (Chemists Reference): controlled (SPECTRa list of Chemist References)
- [1,1] Subject and Keywords (Experiment Reference)
- [1,1] Subject and Keywords (Chemical Formula): encoded (atom symbols with total count subscript)
- [1,1] Subject and Keywords (Compound Class): controlled (eBank Compound Classes)
- [1,1] Subject and Keywords (Systematic Name): encoded (IUPAC Chemical Nomenclature [LFM98])
- [1,1] Publisher
- [1,1] Experiment Date: encoded (W3CDTF)
- [1,1] Resource type: controlled
- [1,1] Resource Identifier: encoded (InChi)
- [0,–] Relation
  - [0,–] Is Referenced By: *Article, etc. that cites the resource*
  - [1,–] Has Part: *Data files making up the resource*
- [0,1] Rights Management
- [0,1] License
- [0,1] Embargo
  - [1,1] Embargo License
    - [1,1] URL: encoded (URI)
    - [1,1] Description
    - [1,1] Machine Readable: encoded (RDF) or empty
  - [1,1] Post Embargo License
    - [1,1] URL: encoded (URI)
    - [1,1] Description
    - [1,1] Machine Readable: encoded (RDF)

├→ [1,1] Period: encoded (number)

├→ [1,1] Release: controlled (e.g. ‘automatic’)

└→ [1,1] Start: encoded (W3CDTF)

Appendix O

Edinburgh DataShare metadata profile

The following is taken from Rice, Macdonald and Hamilton [RMH08].

⟶ [0,1] Depositor: *Name*

⟶ [1,–] Data Creator: *Name, personal or corporate*

⟶ [1,1] Title

⟶ [0,–] Alternative Title

⟶ [0,1] Dataset Description (abstract)

⟶ [1,–] Type: controlled (*see Section O.1*)

⟶ [0,1] Subject Classification: controlled (JACS [HES06])

⟶ [0,–] Subject Keywords

⟶ [0,1] Funder: controlled

⟶ [0,1] Data Publisher

⟶ [0,–] Spatial Coverage: controlled (country/place Geoname)

⟶ [0,1] Time Period: encoded (W3CDTF [WW98] preferred)

⟶ [0,1] Language: controlled (profile of ISO 639-2 [ISO98])

⟶ [0,–] Source: *Primary data sources, URLs*

⟶ [0,1] Dataset Description (TOC): *filenames, descriptions*

⟶ [0,1] Relation (Is Version Of): encoded (URI)

⟶ [0,–] Relation (Is Referenced By): encoded (URI)

⟶ [0,1] Supercedes: encoded (URI)

⟶ [0,1] Rights: *copyright statement*

⟶ [1,1] Date Accessioned: encoded (repository default)

O.1 Type vocabulary

- Collection
- Dataset
- Image
- Moving image
- Sound
- Text

Appendix P

Data Audit Framework metadata profile

These elements are taken from Audit Form 3B: Data Asset Management (Optional extended element set) as set out in version 1.6 of the Data Audit Framework Methodology [JRR08].

⟶ [1,1] ID

⟶ [1,1] Title

⟶ [1,1] Type: *Examples include ‘database’, ‘image collection’, ‘text corpus’*

⟶ [1,1] Owner(s)

⟶ [1,1] Subject

⟶ [1,1] Language

⟶ [0,1] Variant name

⟶ [0,1] Level: *Granularity of description*

⟶ [0,1] Abstract

⟶ [0,1] Keywords

⟶ [1,1] Original purpose

⟶ [1,1] Description

⟶ [1,1] Start date

⟶ [1,1] Usage frequency: *Also required speed of retrieval if known*

⟶ [1,1] Description of context

⟶ [1,1] Source: *Data collection methods, ancestor datasets*

⟶ [0,1] Completion date

⟶ [0,1] Date last modified

⟶ [0,1] Management to date: *History of maintenance, integrity checks*

⟶ [0,1] Curation to date: *History of curation/preservation*

⟶ [1,1] Data creator(s)

⟶ [1,1] Asset manager(s)

⟶ [1,1] Rights: *Users’ rights regarding viewing, copying, redistributing and republishing*

⟶ [1,1] Usage constraints

⟶ [0,1] Former asset manager(s)

⟶ [0,1] Other acknowledgements: *Contact details of other contributors*

⟶ [0,1] FoI, DP, personal privacy issues

⟶ [0,1] Potential Re-uses

⟶ [1,1] Current Location

⟶ [1,1] Coverage: *Intellectual domain, spatiotemporal coverage*

⟶ [1,1] Relation

⟶ [0,1] Version

⟶ [0,1] Responsibility for the asset in the long term: *Retention policy, etc.*

⟶ [0,1] Can/should it be handed to a service provider for curation?

⟶ [1,1] Long term value

⟶ [1,1] Back-up and archiving policy: *Number of copies, frequency of backup*

⟶ [1,1] Disaster recovery measures

⟶ [0,1] Retention period

⟶ [0,1] Preservation policy

⟶ [1,1] File format(s)

⟶ [1,1] Structure of the data asset

⟶ [1,1] Documentation available: *What documentation is available, where it can be found*

⟶ [1,1] Audit trail and fixity: *Measures taken*

⟶ [1,1] Current cost: *Annual maintenance cost*

⟶ [1,1] Funding basis

⟶ [0,1] Original cost of creating the asset

⟶ [0,1] Planned costs for maintenance

⟶ [0,1] Size: encoded (number and byte-related unit)

⟶ [0,1] Hard- and Software Requirements

Bibliography

[ANS95] ANSI/NISO Z39.50 (1995). *Information Retrieval: Application Service Definition and Protocol Specification*. American National Standards Institute. URL: `http://www.loc.gov/z3950/agency/markup/markup.html` (2008-12-12).

[AP08] J Allinson & A Powell (2008). *Scholarly Works Application Profile*. 2008-09-23. URL: `http://www.ukoln.ac.uk/repositories/digirep/index/SWAP` (2008-11-26).

[Bal09] A Ball (2009). *Beyond basic metadata in OAI-PMH*. 2009-05-06. URL: `http://homes.ukoln.ac.uk/~ab318/beyond-oai_dc/` (2009-05-06).

[CG04] R Clayphan & R Guenther (2004). *Library Application Profile*. DCMI Working Draft. Dublin Core Metadata Initiative. URL: `http://dublincore.org/documents/library-application-profile/` (2008-12-23).

[Cat08] W Cathro (2008). 'Collaboration in Building a Sustainable Repository Environment: A National Library's Role'. In: *Open Repositories 2008*. Southampton. URL: `http://pubs.or08.ecs.soton.ac.uk/7/1/OR08_cathro.pdf` (2008-12-23).

[Cha09] T Chaudhri (2009). 'Assessing FRBR in Dublin Core Application Profiles'. *Ariadne* 58 (Jan.). ISSN: 1361-3200. URL: `http://www.ariadne.ac.uk/issue58/chaudhri/` (2009-03-09).

[Cro+09] N Crofts et al. (eds.) (2009). *Definition of the CIDOC Conceptual Reference Model*. Version 5.0.1. ICOM/CIDOC CRM Special Interest Group. URL: `http://cidoc.ics.forth.gr/official_release_cidoc.html` (2009-04-03).

[DAD07] DADDI Project (2007). *Solutions Use Case: Coasts Ocean Arctic 1a*. Output of the DADDI Community Workshop and Seminar held on 16–17 November 2006 at Lamont Doherty Earth Observatory, Columbia University, Palisades, NY. URL: `http://wiki.esipfed.org/index.php/SolutionsUseCase_CoastsOcean_Arctic_1a` (2008-12-23).

[DCM08] DCMI Usage Board (2008). *DCMI Metadata Terms*. DCMI Recommendation. Dublin Core Metadata Initiative. URL: `http://dublincore.org/documents/dcmi-terms/` (2008-12-23).

[DCM09] Dublin Core Metadata Initiative (2009). *DCMI Science and Metadata Community*. 2009-02-10. URL: `http://dublincore.org/groups/sam/` (2009-02-20).

[DDI08] Data Documentation Initiative (2008). *Schema Documentation for DDI Version 3.0*. 2008-04-28. URL: `http://www.icpsr.umich.edu/DDI/documentation/ddi3.0/` (2008-12-12).

[DH04] M Duke & R Heery (2004). *Elements in the eBank Schema*. Version 4. UKOLN : Bath. URL: `http://www.ukoln.ac.uk/projects/ebank-uk/private/ebank-schema.doc` (2008-12-12).

[DSp09] DSpace Foundation (2009). *DSpace 1.6*. 2009-04-13. URL: `http://wiki.dspace.org/DSpace_1.6` (2009-05-21).

[DSpF] DSpace Foundation. *FAQ: What Metadata Standards Does DSpace Support?* URL: `http://www.dspace.org/faqs/index.html#standards` (2008-12-23).

[DSpM] DSpace Foundation. *Metadata*. URL: `http://www.dspace.org/index.php/Architecture/technology/metadata.html` (2008-12-23).

[EGU04] e-Government Unit (2004). *UK GEMINI Standard: A Geo-spatial Metadata Interoperability Initiative*. Version 1.0. Cabinet Office : London.

[EPr07] EPrints Development Team (2007). *EPrints 3 Reference: Metadata*. 2007-01-11. URL: `http://wiki.eprints.org/w/Metadata` (2008-12-23).

[FIP94] *Countries, Dependencies, Areas of Special Sovereignty, and Their Principal Administrative Divisions* (1994). Federal Information Processing Standard 10-4.

[FIS04] Forum on Information Standards in Heritage (2004). *INSCRIPTION List of Wordlists*. URL: `http://www.fish-forum.info/i_lists.htm` (2008-12-12).

[FIS07] Forum on Information Standards in Heritage (2007). *Midas Heritage: A Data Standard for the Historic Environment*. Forum on Information Standards in Heritage. URL: `http://www.english-heritage.org.uk/server/show/nav.18140` (2008-12-08).

[GCM08] Global Change Master Directory (2008). *Directory Interchange Format (DIF) Writer's Guide*. National Aeronautics & Space Administration. URL: `http://gcmd.nasa.gov/User/difguide/difman.html` (2008-11-26).

[GCM08] Global Change Master Directory (2008). *ISO Topic Categories*. URL: `http://gcmd.nasa.gov/User/difguide/iso_topics.html` (2008-12-11).

[GCM08] Global Change Master Directory (2008). *Suggested Keywords for Use in Distribution_Format in the Distribution Field*. URL: `http://gcmd.nasa.gov/User/difguide/distbn_format_sugval.html` (2008-12-11).

[GCM08] Global Change Master Directory (2008). *Suggested Keywords for Use in Distribution_Media in the Distribution Field*. URL: `http://gcmd.nasa.gov/User/difguide/distbn_media_sugval.html` (2008-12-11).

[Gib09] H Gibbs (2009). *DataShare Metadata Schema for ePrints Soton (ePrints 3.1)*. DataShare Project. Data Information Specialists Committee – UK. URL: `http://www.disc-uk.org/docs/sPrints_Soton_Metadata.pdf` (2009-05-05).

[Go-08] Go-Geo! Project (2008). *UK Academic Geospatial Metadata Application Profile*. Version 1.0. EDINA, University of Edinburgh : Edinburgh. URL: `http://www.gogeo.ac.uk/files/UK%20AGMAP%20with%20Intro.pdf` (2008-12-22).

[Gre07] A Gregory (2007). *DDI Lifecycle View and Use Cases*. Presented at the DDI 3 Workshop. GESIS-ZUMA Mannheim. 2007-10-24. URL: `http://db.zuma-mannheim.de/DDI/Workshop/2007-10-24/DDI%203%20Workshop/Slides/Day1-2-LifeCycle.ppt` (2008-12-23).

[HES06] Higher Education Standards Agency (2006). *Joint Academic Coding System (JACS)*. Version 2. URL: `http://www.hesa.ac.uk/jacs2` (2008-12-11).

[HP00] R Heery & M Patel (2000). 'Application Profiles: Mixing and Matching Metadata Schemas'. *Ariadne* 25 (Sept.). ISSN: 1361-3200. URL: `http://www.ariadne.ac.uk/issue25/app-profiles/` (2008-11-26).

[IAN07] Internet Assigned Numbers Authority (2007). *MIME Media Types*. Internet Corporation for Assigned Names & Numbers. 2007-03-06. URL: `http://www.iana.org/assignments/media-types/` (2008-12-12).

[IFL98] IFLA Study Group on the Functional Requirements for Bibliographic Records (1998). *Functional Requirements for Bibliographic Records: Final Report.* UBCIM Publications — New Series 19. Saur : Munich. URL: http://www.ifla.org/VII/s13/frbr/frbr.pdf (2008-12-23).

[IOC] Intergovernmental Oceanographic Commission. *Cruise Summary Report: ROSCOP (3rd Edition).* URL: http://www.bodc.ac.uk/data/information_and_inventories/cruise_inventory/documents/new_csr_form.pdf (2008-10-08).

[ISO03] ISO 19115 (2003). *Geographic information – Metadata.* 1st ed. International Organization for Standardization.

[ISO04] ISO 8601 (2004). *Data elements and interchange formats – Information interchange – Representation of dates and times.* 3rd ed. International Organization for Standardization.

[ISO05] ISO 2146 (2005). *Information and documentation – Registry Services for Libraries and Related Organisations.* 3rd ed. Draft International Standard. International Organization for Standardization.

[ISO07] ISO/TS 19139 (2007). *Geographic information – Metadata – XML schema implementation.* 1st ed. International Organization for Standardization.

[ISO] ISO 3166. *Codes for the representation of names of countries and their subdivisions.* Multipart standard. International Organization for Standardization.

[ISO98] ISO 639-2 (1998). *Codes for the representation of names of languages – Part 2: Alpha-3 code.* 1st ed. International Organization for Standardization.

[Ihb] *Limits of Oceans and Seas* (1953). 3rd ed. Special Publication 23. International Hydrographic Bureau : Monaco.

[Int] Intute Repository Search Project. *UK Institutional Repository Search.* MIMAS. URL: http://www.intute.ac.uk/irs/ (2009-06-01).

[JRR08] S Jones, S Ross & R Ruusalepp (2008). *Data Audit Framework Methodology.* Version 1.6. HATII, University of Glasgow. URL: http://www.data-audit.eu/DAF_Methodology.pdf (2008-12-12).

[Jon+07] A Jones et al. (2007). *FuGE: Functional Genomics Experiment model specification.* Final Recommendation. Version 1. FuGE Working Group. URL: http://fuge.sourceforge.net/dev/V1Final/FuGE-v1-SpecDoc.doc (2009-04-17).

[KDC05] T Koch, M Duke & S Coles (2005). *Metadata Application Profile: eBank UK project.* UKOLN, University of Bath. URL: http://www.ukoln.ac.uk/projects/ebank-uk/schemas/profile/2005/11/02/ (2008-12-12).

[LFM98] GJ Leigh, HA Favre & WV Metanomski (1998). *Principles of Chemical Nomenclature: A Guide to IUPAC Recommendations.* Ed. by GJ Leigh. Blackwell Science : Oxford. ISBN: 0-86542-685-6. URL: http://old.iupac.org/publications/books/principles/principles_of_nomenclature.pdf (2008-12-12).

[Lat+09] SE Latham et al. (2009). 'The NERC DataGrid Services'. *Philosophical Transactions of the Royal Society A* 367 (Mar.). Discussion Meeting on the Environmental eScience Revolution, London, 7th–8th April, 2008. 1015–1019. ISSN: 1364-503X. DOI: {10.1098/rsta.2008.0238}.

[Law+09] BN Lawrence et al. (2009). 'Information in Environmental Data Grids'. *Philosophical Transactions of the Royal Society A* 367. 1003–1014. ISSN: 1364-503X. DOI: 10.1098/rsta.2008.0237.

[Lyo07] L Lyon (2007). *Dealing with Data: Roles, Rights, Responsibilities and Relationships.* Consultancy Report. JISC. URL: http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf (2008-11-25).

[Mor+08] L Moreau et al. (2008). *The Open Provenance Model.* Version 1.01. Second Provenance Challenge. URL: http://eprints.ecs.soton.ac.uk/16148/ (2009-04-17).

[NBJ08] M Nilsson, T Baker & P Johnston (2008). *The Singapore Framework for Dublin Core Application Profiles.* Recommended Resource. Dublin Core Metadata Initiative. URL: http://dublincore.org/documents/singapore-framework/ (2008-12-24).

[NEB03] NERC Environmental Bioinformatics Centre (2003). *Guidelines for Editing Dataset Records.* 2003-08-15.

[NEB] NERC Environmental Bioinformatics Centre. *NEBC EnvBase Catalogue Data Types.* URL: http://envgen.nox.ac.uk/datatypes.html (2008-12-12).

[NER08] NERC DataGrid Project (2008). *Metadata Objects for Linking Environmental Sciences.* 2008-12-15. URL: http://proj.badc.rl.ac.uk/ndg/wiki/MOLES (2008-12-15).

[NGD00] National Geospatial Data Framework (2000). *Discovery Metadata Guidelines.* Version 1.2. National Geospatial Data Framework Management Board : Southampton, UK. URL: http://www.ngdf.org.uk/Metadata/metguide/metaguide12.pdf (2008-11-26).

[OAI08] Open Archives Initiative (2008). *The Open Archives Initiative Protocol for Metadata Harvesting.* Ed. by C Lagoze & H Van de Sompel. Version 2.0. Open Archives Initiative. URL: http://www.openarchives.org/OAI/openarchivesprotocol.html (2008-12-23).

[OE07] M Osborne & P Eastwood (2007). *Report on Use Cases for Assessing MDIP Benefits.* Report of the Mapping and Applications Working Group. UK Marine Data & Information Partnership. URL: http://www.oceannet.org/mdip/documents/MDIP_MAWG_use_cases_report.pdf (2008-12-23).

[Ols+07] L Olsen et al. (2007). *NASA/Global Change Master Directory (GCMD) Earth Science Keywords.* Version 6.0.0.0.0. Global Change Master Directory. National Aeronautics & Space Administration. URL: http://gcmd.nasa.gov/Resources/valids/archives/keyword_list.html (2008-12-11).

[Ope09] OpenDOAR (2009). *OpenDOAR Chart: Usage of Open Access Repository Software – Worldwide.* University of Nottingham. 2009-02-24. URL: http://www.opendoar.org/onechart.php?groupby=r.rSoftWareName&orderby=Tally%20DESC&charttype=pie&width=600&height=300&caption=Usage%20of%20Open%20Access%20Repository%20Software%20-%20Worldwide (2009-02-24).

[RMH08] R Rice, S Macdonald & G Hamilton (2008). *DSpace Metadata Schema for Edinburgh DataShare.* Version 1. DataShare Project. Data Information Specialists Committee – UK. URL: http://www.disc-uk.org/docs/Edinburgh_DataShare_DC-schema1.pdf (2009-05-05).

[Rya07] T Ryan (2007). *Next Generation Discovery and Delivery at the University of California.* Presentation given to the University Library Council Task Group on Discovery and Metadata. Harvard University. Mar. 2007. URL: http://isites.harvard.edu/fs/docs/icb.topic117213.files/Terry_Ryan_March_2007.ppt (2008-12-23).

[SDS08] C Smith, D Davis & T Staples (2008). *Fedora Commons Community Registry*. Version 19. 2008-12-18. URL: https://fedora-commons.org/confluence/display/FCCommReg/Fedora+Commons+Community+Registry (2009-02-24).

[SM04] S Sufi & B Matthews (2004). *CCLRC Scientific Metadata Model: Version 2*. Technical Report DL-TR-2004-001. CCLRC Daresbury Laboratory : Warrington. URL: http://epubs.cclrc.ac.uk/work-details?w=30324 (2008-12-10).

[SS00] Sea-Search (2000). *Instructions For Collating EDMED Entries*. European Directory of Marine Environmental Datasets Information Note 2. URL: http://www.bodc.ac.uk/data/information_and_inventories/edmed/documents/edmed2.pdf (2008-11-26).

[Spa01] D Spanner (2001). 'Border Crossings: Understanding the Cultural and Informational Dilemmas of Interdisciplinary Scholars'. *Journal of Academic Librarianship* 27:5. 352–360. ISSN: 0099-1333. DOI: 10.1016/S0099-1333(01)00220-8.

[TM07] A Tonge & P Morgan (2007). *Project SPECTRa: Submission, Preservation and Exposure of Chemistry Teaching and Research Data*. JISC Final Report. University of Cambridge & Imperial College London.

[TW08] A Treloar & R Wilkinson (2008). 'Access to Data for eResearch: Designing the Australian National Data Service Discovery Services'. *International Journal of Digital Curation* 3:2. ISSN: 1746-8256. URL: http://www.ijdc.net/index.php/ijdc/article/viewFile/95/ (2008-12-23).

[UKD08] UK Data Archive (2008). *Humanities and Social Science Electronic Thesaurus (HASSET)*. Ed. by L Ageer. Version 3.0. 2008-07-21. URL: http://www.data-archive.ac.uk/search/hassetAbout.asp (2008-12-12).

[W3C07] W3C (2007). *XQuery 1.0: An XML Query Language*. Ed. by S Boag et al. Recommendation. World Wide Web Consortium. URL: http://www.w3.org/TR/xquery/ (2008-11-23).

[W3C09] W3C (2009). *SKOS Simple Knowledge Organization System Primer*. Ed. by A Isaac & E Summers. Working Draft. World Wide Web Consortium. URL: http://www.w3.org/TR/2009/WD-skos-primer-20090317/ (2009-04-15).

[WW98] M Wolf & C Wicksteed (1998). *Date and Time Formats*. Note. World Wide Web Consortium. URL: http://www.w3.org/TR/NOTE-datetime (2008-12-11).

[Whe+00] DL Wheeler et al. (2000). 'Database resources of the National Center for Biotechnology Information'. *Nucleic Acids Research* 28:1 (2000-01-01). 10–14. ISSN: 0305-1048. DOI: 10.1093/nar/28.1.10. The NCBI Taxonomy Browser may be accessed from: http://www.ncbi.nlm.nih.gov/Taxonomy/.

[Whi+08] HC White et al. (2008). 'The Dryad Data Repository: A Singapore Framework Metadata Architecture in a DSpace Environment'. In: *Proceedings of the International Conference on Dublin Core and Metadata Applications*. Humboldt-Universität zu Berlin. Ed. by J Greenberg & W Klas. Dublin Core Metadata Initiative & Universitätsverlag Göttingen, 157–162. ISBN: 978-3-940344-49-6. URL: http://edoc.hu-berlin.de/docviews/abstract.php?lang=eng&id=29145 (2009-02-20).

[Woo+06] A Woolf et al. (2006). 'Data Integration with the Climate Science Modelling Language'. *Advances in Geosciences* 8:1. 83–90. ISSN: 1680-7340. URL: http://www.adv-geosci.net/8/83/2006/ (2008-12-23).

[Woo+07] A Woolf et al. (2007). 'Enterprise Specification of the NERC DataGrid'. In: *Proceedings of the UK e-Science All Hands Meeting 2007*. Ed. by SJ Cox. National e-Science Centre. ISBN: 978-0-9553988-3-4. URL: `http://epubs.cclrc.ac.uk/bitstream/493/128.pdf` (2008-12-23).