Demonstration Test Catchments – Data Management Requirements

Bryan Lawrence, Ag Stephens, Andrew Woolf

STFC Centre for Environmental Data Archival

Gwyn Rees

NERC Centre for Ecology and Hydrology

James Doughty

DIASS Ltd

Executive Summary

The demonstration test catchment project needs a data archive which provides both long term persistence of the data products and suitable data access portals. Such portals should consist of a vanilla portal for the archive alongside rich portals with custom functionality to be developed within the research consortia. Both the simple and complex portals should exploit interfaces into a common archive.

This document outlines a standards-based approach to both the data persistence for the archive and the data interfaces needed for the portals. It should prove suitable as a guiding note to both those who need to procure the DTC archive solution, and those who need to tender for its provision.

Five main sections are presented: a two page "overview", an introduction outlining scope etc, and then three sections providing an RM-ODP based view of the major components needed from the enterprise requirements, via the information perspective to the computational viewpoints. No technical nor engineering viewpoints are presented as the details of these would be in the domain of the archive provider.

Table of Contents

1 OVERVIEW

The Demonstration Test Catchments (DTC) project launched by Defra in partnership with other agencies aims for three different research consortia to instrument a catchment each, and investigate how changes in farming practice impact on the environment and farm productivity.

The project requires a common archive allowing participants from all consortia, as well as wider academia and even the general public to access data and some interpretation of the data. The archive will host a variety of data, some of which will need to be restricted, or provided in an "identity-obscured" manner, but in general the aim is to make data access as easy as possible. Data will need to be made accessible as soon as practicable after being obtained, and for the foreseeable future.

This document outlines an approach for developing the archive in such a way that the data will be as accessible as is possible – not just as raw data files (or tables) but also through standards compliant interfaces. The interfaces suggested here are those required by current and expected legislation in the UK and Europe augmented by those already identified by the relevant research communities as the most likely to provide both software and community interoperability. It is expected that this document will be used by Defra and the eventual archive supplier to help define and deliver the archive procurement.

Data longevity and interoperability requires a level of data documentation that is foreign to most scientists, who tend to record only the information that they deem important to their own project goals (and even then, only information that changes rapidly enough that they can't just "remember" it). A key part of ensuring data re-usability is identifying what "meta"data should be kept – that is, what data about data. (It is important to also note that what is metadata to one person may well be fundamental data to another.) A number of classes of metadata are introduced here, of which five are crucial:

- Archive metadata: describing what is measured, where it was measured, and the syntax of the data records along with some of the semantics of the sampling method.
- Browse metadata: describing in more detail how the measurements were made (what instrument or model produced the data), why the data was collected etc. It might include calibration ancillary data if that was relevant. Browse metadata should be enough to discriminate between data which would otherwise appear to be very similar.
- Character metadata: third party and post hoc assessments of the suitability and quality of the data (such as citations and annotations).
- Discovery metadata: information that is shared to national and international catalogues so that the data can be found in the first place. Discovery sits at the start of a usage chain which leads onto selection via inspection of browse and character metadata, then usage by tools which understand the archive metadata).
- Extra metadata: the discipline dependent metadata that cannot be handled by generic systems for the other classes of metadata (including but not limited to academic papers and PDF documents).

Producing systems that can understand and manipulate these sorts of metadata is important: it doesn't take much data to overload human indexing systems. There is an abundant body of material to draw on, much of it based on the Observations and Measurements (O&M) specification which is about to be standardised as ISO19156. O&M provides an integrating paradigm: observations consist of measurement data (results) obtained about features of interest linked with methods (processes). O&M is getting significant uptake in many communities, and is the most obvious protocol to use for DTC metadata.

The main problem for the DTC project will be finding a profile which both supports the requirements of the DTC and is as consistent as possible with the O&M profiles of the (disparate discipline-specific) DTC communities. An example of this issue will be supporting the new WaterML language as it evolves, even as some of the metadata could be described using MOLES and/or GeoSciML (see section 4 for brief introductions to WaterML, MOLES and GeoSciML).

The metadata structures which are necessary need to be accompanied with vocabularies which cover the domains of interest with well defined terms which can be both related to each other and discriminate between the key characteristics of the data and metadata. Some of these vocabularies will be pre-existing, and some will need to be constructed during the project, perhaps with the establishment of community governance procedures so that their relevance and accuracy can continue to be improved.

The DTC project will need to establish a methodology to establish the right structure (profile of O&M) and prioritize the development and maintenance of the relevant vocabularies. This will involve not only an "architectural" task (to be done by the archive supplier), but also a considerable effort by the catchment research consortia.

A key part of the architectural task will be establishing a query model: that is defining the axes along which the data and metadata are most naturally queried. It is these axes which should be indexed, and for which a "web portal" should provide methods of querying. For example, an axis of interest could be to find which river elements exceeded a specific flow rate at a specific time: so a method of indexing that could be constructed, as could a way of a user entering that specific query. (This is a somewhat contrived example, which is why establishing the query model itself is so important, once that is done, a significant part of the archive and interface design has been defined.)

The data within the archive will need to be ingested and stored and made available, and the formats by which all three functions are delivered will have to be limited: if all possible file formats were to be supported (let alone all possible database arrangements), the amount of work would be essentially unlimited. To that end, the project will have to agree on a limited number of formats for ingestion, and a limited number of formats for data download. We recommend the use of formats which include adequate internal metadata to allow informed user support, that is, the project should not only specify the formats, but also how those formats should be used: for example, we recommend the use of the BADC comma-separated value format for spreadsheet data – or something similar – not because the BADC format itself is so special, but because it unambiguously defines what must appear in the spreadsheet to aid reuse. It may well be that the project uses another spreadsheet profile, but whatever is used must admit the incorporation of standardised metadata. Where possible XML formats corresponding to the O&M DTC profile should be preferred. Specific recommendations as to formats appear later in the document.

The archive itself will also need to be constructed so that a variety of portals to the underlying data can be built. We expect that not only will the archive supplier will deliver a "vanilla" portal – with data download and limited visualisation – but the research consortia themselves will want to exploit the archive. To that end, the archive and metadata systems should be constructed to conform with Open Geospatial Consortium (OGC) web service interfaces: in particular, an OGC web map service (WMS) interface to datasets that can be visualized as layers on maps, a Sensor Observation Service (SOS) interface to allow the retrieval of specific sensor observations, and an OGC web feature service (WFS) to allow the extraction of specific features (specific identifiable objects) from within the datasets. The web services should support the query model, so that if remote portals want to subset the data against specific queries, they can do so. Web service interfaces to support the legislative requirements for discovery metadata should also be provided.

The project will have to address some sort of access control, not to stop people accessing general data, but in order to ensure that statistics of usage can be kept (so Defra and/or the archive provider) are in a position to evaluate the importance of the data therein, and to ensure that data with commercial and/or personal privacy implications can have limited access. While there are a plethora of access control protocols available, we recommend the use of OpenID, which has considerable penetration in both the commercial and academic sectors.

2 INTRODUCTION

2.1 The Demonstration Test Catchments Programme

UK land management will need to comply with a range of new and existing EU and domestic legislation all of which aims to improve the quality of water in rivers, lakes, ground water, transitional and coastal waters. In order to meet these EU and domestic targets there is a need to tackle diffuse water pollution from agriculture. Current indications are that the scale of the problem is such that it may be necessary, at least in some catchments, to adapt the ways in which farming can be carried out in the future, and so an evidence-based approach is necessary.

To that end, three catchments have been chosen as target areas for study in the Defra funded "Demonstration Test Catchments" (DTC) project - the Eden, the Hampshire Avon and the Wensum. Three consortia have been awarded contracts to undertake studies in each catchment. The establishment of an 'integrative data infrastructure for collaborative analysis by the wider research community' is a significant objective for each consortium.

Similarly, as each consortium is comprised of many different organisations, Defra require an integrated approach to data management between, within and across the consortia. The ability to collaborate effectively is critical to the success of the programme. It is expected that collected data should be readily available in common formats and use common metadata (conforming to data standards where possible) to facilitate data discovery and data analysis and provide a repository of information for future generations.

To further facilitate the use of the data, and to ensure transparency of scientific results, as much data as possible will be publicly accessible, and all the data infrastructures will be constructed to facilitate public data discovery and reuse. To that end, Defra will procure a data archive solution as part of the programme data management.

2.2 Purpose of this document

This document has been commissioned by the DTC Project Board in advance of the procurement of a data archive solution, to provide guidance

1. For establishing an initial data model to facilitate the initial data handling within the DTC project,
2. Provide guidance for the development of a programme data management plan, and
3. Provide material necessary for both Defra and any potential bidder for the data archival solution to evaluate the complexity of the data archival problem.

In doing so, the document includes recommendations for data policies and data management activities as appropriate. It should be noted that these can only be recommendations, until a data management plan is agreed by the three key stakeholders in the delivery of the archive: the consortia, the data store provider, and the DTC Project Board.

2.3 Document Organisation

Data management in a project this complex will itself be complex. There are a range of stakeholders, from Defra who are funding the project, to the consortia themselves acting both as data producers and data consumers, to those responsible for delivering cross-consortia data management, and those responsible for building the software systems.

Each stakeholder brings their own perspective and requirements in terms of specifying a data management plan, a data model, and a data archive. To that end, the layout of the material presented here loosely follows the first three components of the reference model for open distributed processing (RM-ODP), so that stakeholder readers can concentrate on material from the perspective most relevant to them. RM-ODP suggests the following architectural viewpoints[1]:

- The enterprise viewpoint, which focuses on the purpose, scope and policies for the system.
- The information viewpoint, which focuses on the semantics of the information needed in the system, and the structure and content type of the supporting data.
- The computational viewpoint, which enables distribution through functional decomposition on the system into objects which interact at interfaces. It describes the functionality provided by the system and its functional decomposition.
- The engineering viewpoint, which focuses on the mechanisms and functions required to support distributed interactions between objects in the system.
- The technology viewpoint, which focuses on the choice of technology of the system.

This document does not consider the engineering or technology viewpoints as the eventual supplier of the data management infrastructure would deliver these.

Section 3 provides the Enterprise Viewpoint, Section 4 outlines the Information Viewpoint perspective and presents the shape of an appropriate data model. Section 5 presents a limited computational perspective based on policy requirements from the Enterprise Viewpoint. Each section concludes with a summary, and these are aggregated into a complete set of recommendations in section 6 In the remainder of this introduction we provide an introductory set of definitions to help scope the concepts used hereafter.

2.4 Definitions

Data – within this project we expect both numerical and image data collected from instruments, and subjective data collected from interviews with people. Video data will also be collected. We might think of hard data (numbers and images) and soft data (interpretations).

Raw Data – data which has not been quality controlled, may include gaps, outliers and spurious records which would be removed before a final dataset is made available. Normally this is the data as received from an instrument or analysis.

Specimens & Samples – these refer to physical entities that have been collected for study such as a soil or water sample or a species found at a site. Although there is no intention to deal with the management of these artefacts within this data management plan, such entities will generate metadata, and may generate subsequent numerical data (e.g. when removed to a laboratory and subjected to chemical analysis).

Metadata – this is information that describes data that have been collected. There is a wide range of metadata necessary to underpin requirements that range from that required for legislative compliance, and that required to support data manipulation. The metadata requirements of the DTC are discussed in section 3.

Data Model – Data models are key components of the metadata architecture, describing the structure of data and metadata entities and their relationships. Several data models may be necessary in any given project, covering at the least the semantics and syntax of the data, and the semantics and syntax of the metadata.

Dataset – a collection of data which shares a particular scope and usage. We expect a range of datasets to be defined, many of which overlap. “Archive datasets” will refer to datasets as they are physically organised, “Virtual datasets” will refer to aggregations of data which cross the boundaries of archive datasets. For example, data collected at a single sampling station over the course of a year may be stored together (and is thus an “archive dataset”), but a virtual dataset can be constructed by aggregating across all the data collected within a year at all stations sharing the same instrument.

Feature – Some real world entity which is described and/or measured and simulated within a dataset.

[1] RM-ODP summarised from Wikipedia, accessed 8th March 2010 at: http://en.wikipedia.org/wiki/RM-ODP

Features may be actual objects such as rivers and catchments or sampling objects such as "a profile across a river at specific point".

Derived Dataset – a new dataset formed by amalgamating or post-processing one or more datasets into a new dataset.

Data Archive – a long term storage facility for digital data.

Data Services – broadly speaking this refers to any tools deployed on the web that are used to expose and/or manipulate data. However, we will limit ourselves here to those involved with ***data discovery***, ***metadata browse***, and data ***view*** and ***download***. Data services may be accessed through a web browser but will typically interact with other computational systems (services and portals) rather than human users.

Data Portals – are websites which integrate one or more data services so that a web-browser can present them to human users.

These last terms are best illustrated by an example:

> Someone wishes to know whether there are any data available on water nitrogen levels across the Wensum catchment. A website which provides search facilities and returns a list of datasets is a ***Data Portal*** which utilises a ***Data Discovery Service.***
>
> Before deciding to use a particular dataset one might want to know a little more about it, i.e. more about its context – which instrument captured the information, the sampling regime etc. A website which provides the ability to navigate between datasets and their descriptions is a ***Data Portal*** that will exploit a ***Metadata Browse Service***.
>
> Once satisfied that this is the right dataset one might want to see a sample of the data, a report or a summary view, displayed on a map, a graph, or a time series animation etc. A website which provides visualisation facilities is a particular kind of ***Data Portal***: a ***Visualisation Portal***, which exploits ***View Service***(s).
>
> After all that, one might want a copy of the data itself, or a subset thereof. ***Data Download Services*** provide data subsetting and download, and again, a data portal can provide an interface to such a service.

3 ENTERPRISE VIEWPOINT

3.1 Stakeholders

The key business requirement of the DTC archive activity is to acquire and facilitate the use of data describing activities in a number of catchments. Hence, the first questions to ask are "*for whom are the data being acquired*" and "*how will they use them"?*

The primary input for the requirements presented here have come from representatives of the research consortia, the Environment Agency, NERC, and Defra who attended a workshop in late January, 2010. There is a wider user community that will include future research projects within the consortia, other researcher communities, conservation organisations (such as Natural England) and the wider public. It is clear that there is potential for scope creep as the project progresses.

3.2 Central Archive

The key aim of DTC archival is to provide long term persistence for the data acquired in the DTC project and to facilitate usage of the data during the life-time of the project. While the latter could be achieved by a distributed approach, long term custodianship requires a centralised approach. The expected longevity of the archive will need to be established a priori (given that the sensors are likely to be required to be suitable for ten year deployments, one might anticipate a minimum of ten years). To that end a business model which outlines what will be done with the data when Defra funding ceases will be required (while

Defra may fund the DTC very long term, it is inevitable that a “demonstration” project will eventually finish). The eventual archive specification will need to address the points laid out in Table 1.

Service Level	The DTC project should establish and agree service and support expectations between the Data Archive and the data producers and consumers at the outset (including, for example whether support is available 24x7x365 or only during UK office hours). This should be done via a Service Level Agreement (SLA) forming part of the contract with the archive supplier. We recommend, following the experience with UKCP, that such a SLA recommends that data users cannot get help with data after download unless they can provide the metadata that would have accompanied the download (thus obviating a considerable amount of work).
Standards Compliance	Interfaces to the archive website should conform to HTML and CSS standards. Interfaces to the data should include INSPIRE compliant discovery, view and download services, and data models should conform to Open Geospatial Consortium (OGC) and data.gov.uk linked data best practices.
Legislative Compliance	The data archive should have interfaces that conform with INSPIRE legislation. Security and download should not preclude valid FOI or Environmental Information Regulation requests (but may require requesters under such legislation to register).
Central Dataset Storage	The DTC project will not want to store all third party datasets in the Data Archive, nor is it practical to do so. It is recommended that an assessment mechanism is created to ensure that the right datasets get into the Data Archive. Those extra datasets may well be stored in local archives. In such cases the appropriate discovery metadata should be provided to data discovery services. Facilities to allow subsetting of data, and the construction of appropriate virtual datasets should be supported (in particular, via aggregation and/or subsetting to specific properties and/or along spatio/temporal axes).
Metadata Storage	All data should be documented to a level (to be defined in the agreed data models) which allows all programme participants to understand the data syntax and semantics, and the characteristics of the protocols or instruments used to obtain and manipulate the data.
Consortium Data Storage	There may be a requirement to have separate consortium data archives to support extra portal functionality as well as interim and third party data. This may require the provision of simple tools to the consortia to expedite the provision of discovery metadata for such data.
Data Ingestion Tools	Tooling will be required to ingest all the key data streams (section3.3), and if necessary convert to/from the common data model (section 3.6) and/or expose data to interfaces which comply to that model.
Data Syntax	Download data formats should include the syntax decided on for encoding the data model and at least one suitable ASCII (plain text) format along with embedded URIs and metadata.
Access Control:	As discussed in section 3.8, access control will be required for most data and metadata even though the aim of the project is to make the data as open and easily accessible as possible. Hence a security layer between data discovery and data and/or metadata view or download will be required.
Registration	The expectation is that for much data, registration will be required in order to meet access control requirements. However, it is known that registration can be a barrier to data access, even where the user would be granted access simply by registering with simple contact details. To that end, if possible, community standards for cross-site user authentication should be supported (see section 5.6)

Table 1: Archive Enterprise Level Requirements

3.3 Key Data Streams

The requirements workshop identified the following key data streams to be archived.

Data Direct from Instruments in the Field (or Space): The DTC project will receive data from a variety of instruments deployed in each catchment, some are already in place and some will be deployed directly by the programme. In the majority of cases this is a straightforward data flow – the instrument is calibrated and deployed, contextual metadata (including calibration data) is captured and recorded, the instrument takes measurements, the measurements are quality assured and the dataset is made available electronically. Some earth observation data may also be required.

Data Generated from Laboratory Analysis: The DTC project will collect specimens and samples in the field that require further analysis either in the field (soil pH) or back in the laboratory (that may be provided by one or more third parties). This process needs careful design as it is likely that the analysis will be undertaken by more than one laboratory, and information about which laboratory did what will be important.

Data Generated by Post-Processing and Simulation: Some derived datasets will be generated by analysis codes, and some primary datasets may be simulated from models. In both cases, the codes themselves may be important metadata, and metadata which describes the codes will be important (in most cases the code itself is not adequate documentation)**.**

Data from Surveys: Social survey/questionnaire data are likely to be required and one needs to be able to either discover and link to these data or in circumstances where the data are not stored securely store them. There are legislative privacy implications: see section 3.8.

Anecdotal Data: Data extracted from farm management documents will also be needed, as will copies of the underlying documents in some cases. There are business privacy implications: see section 3.8.

Other "Soft" data: It is understood that the project will produce and collect a significant amount of other "soft" data. This data will require special attention, both in terms of the data model, and in terms of access control (see section 3.8). Archive procedures outlined in this document should be considered for all data products involved in the project. Soft data examples identified in the workshop were:

- Farm practice data: raw data transcript needs to be kept
- Web cam: still images and real-time footage, perhaps tied to specific events
- Crop yields
- Agronomic data
- Videos of workshops

Legacy and Third Party Data: The DTC project will require access to datasets collected in the past and it will not be able to dictate the format and structure of these legacy datasets. In many cases, the requirements on the data archive will be to point at such data, but in some cases, such data will need to be acquired and archived in the main DTC archive. Clear criteria will need to be established as to whether to "point to" or to acquire. (There is the potential for duplicated effort across the DTC projects as well as the risk, where payment is involved, of multiple purchases of the same data.)

Raw data will flow direct to the data as well as into the research consortia for quality control (Figure 1). Both sorts of data will be available, but the raw data should be made more difficult to obtain, so that by default most users only see quality controlled data. (However, experience shows that the raw data are as valuable as the quality assured data since quality control algorithms can be improved, leading to better data products with time.)

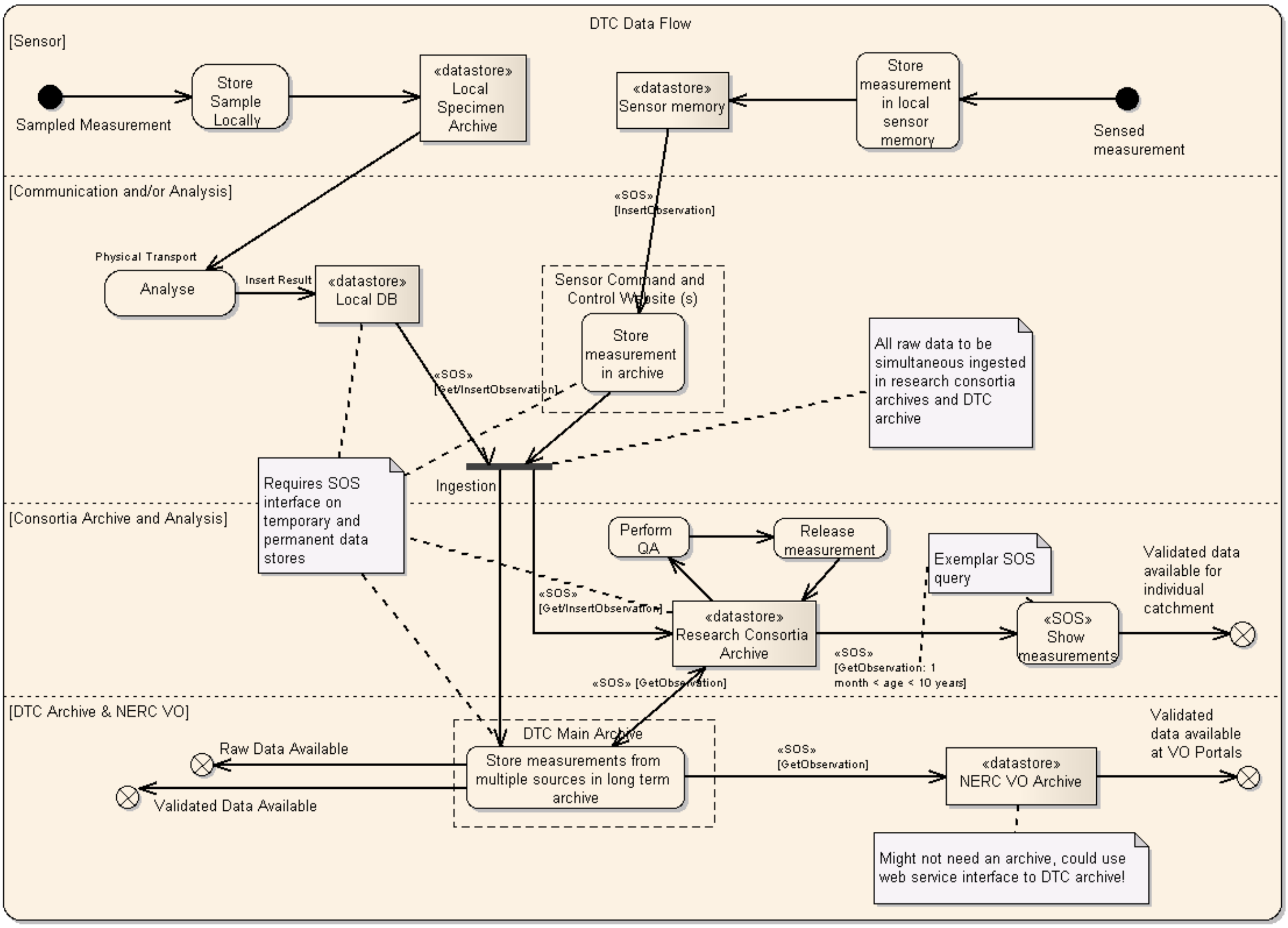
DTC Data Flow
[Sensor]
Sampled Measurement
Store Sample Locally
«datastore» Local Specimen Archive
«datastore» Sensor memory
Store measurement in local sensor memory
Sensed measurement
[Communication and/or Analysis]
«SOS» [InsertObservation]
Physical Transport
Analyse
Insert Result
«datastore» Local DB
Sensor Command and Control Website (s)
Store measurement in archive
«SOS» [Get/InsertObservation]
All raw data to be simultaneous ingested in research consortia archives and DTC archive
Requires SOS interface on temporary and permanent data stores
Ingestion
[Consortia Archive and Analysis]
Perform QA
Release measurement
Exemplar SOS query
Validated data available for individual catchment
«SOS» [Get/InsertObservation]
«datastore» Research Consortia Archive
«SOS» Show measurements
«SOS» [GetObservation: 1 month &lt; age &lt; 10 years]
«SOS» [GetObservation]
[DTC Archive &amp; NERC VO]
DTC Main Archive
Raw Data Available
Store measurements from multiple sources in long term archive
Validated Data Available
«SOS» [GetObservation]
«datastore» NERC VO Archive
Validated data available at VO Portals
Might not need an archive, could use web service interface to DTC archive!

Figure 1: Key Data Flow characteristics: data originates as samples or sensed measurements. Samples are analysed (somewhere) and both streams of data are expected to be ingested into both consortia archives and the DTC archive. QA processed data is also ingested into the DTC archive as soon as is practicable. Note that this diagram indicates the types of data flow and does not represent a finalised design of the system.

3.4 Interoperability Requirements

The key requirement of the data archive activity is to ensure that data is available now and for the foreseeable future to the communities of interest, which will include both the members of the research consortia, and the general public (such as farmers within the relevant catchments).

Within each consortium, across the consortia, and into the public, there is a wide range of maturity and expertise associated with data handling and interpretation, so beyond mere data persistence, a major function of the archival system will be to enhance the accessibility of data. Such enhancement is best achieved by portals and interfaces customised for particular user communities which exploit common service interfaces to the underlying data. Such an approach can be consistent with both the European INSPIRE regulations and the data.gov.uk linked data initiatives, provided appropriate interfaces and data models are used.

The relationship between services, data and metadata is well described by ISO19101, as summarised in Figure 2, from Lawrence *et al*, 2009[2]. Other appropriate standards and interfaces are discussed in section 4, but engineering solutions which exploit these standards to provide an interoperable archive will depend on establishing appropriate data models for the project. That is, a key requirement of the project is both to

[2] B.N. Lawrence,R. Lowry, P. Miller, H. Snaith, and A. Woolf (2009): Information in environmental data grids. Phil. Trans. R. Soc. A, 367, 1003 - 1014. doi:10.1098/rsta.2008.0237.

develop appropriate data models, and for the project participants to exchange data either in formats which directly conform to the data model, or via software interfaces which enforce semantic compliance[3]. However, experience tells us that data models can never be specified completely ab initio, they evolve as new properties of the data are found to be important or new data entities are added to the mix. Thus, an important criterion of any data model is extensibility.

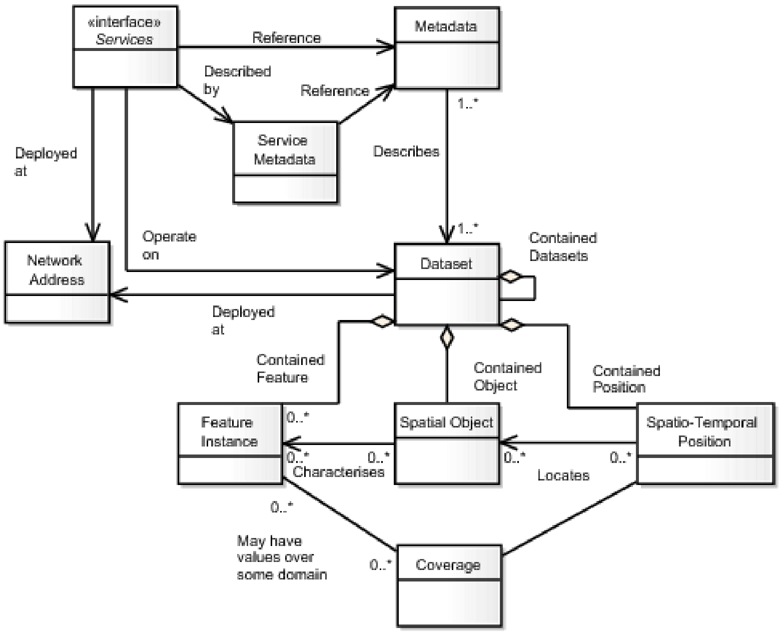

Figure 2: Key entities (extended from ISO19101): data and services are described by metadata. Services deployed on the network operate on data, which consists of features themselves often characterised by spatio temporal characteristics and possibly numerical values expressed over some domain.

Another important criterion for the successful use of such data models is the engagement of the data producers and consumers in the development and specification of the data models – ensuring that the key data properties are captured.

The potential extensibility of the data model has implications for the software solutions chosen. In many cases such extensibility precludes the use of commercial off the shelf software (COTS) for many of the data services required and practical implementations are usually thought to be better off relying on open source software in such cases. (Note that there are data services, in particular web map services, which may not have significant dependencies on data models for some data types, and for those, COTS may well be appropriate.)

3.5 Adoption of Sensor-Web standards

Interoperability concerns the ability to exchange information between parties using a shared syntax and semantics – a goal best achieved by adopting best practice standards. In this case, the Sensor Web Enablement (SWE) suite of specifications from OGC is ideally suited. Benefits of their adoption include: increased interoperability with other sensor-web initiatives (and spatial data infrastructures more generally), enhanced data sharing opportunities (since COTS client support for these standards is developing), and the opportunity to benefit from significant open source software development activity internationally.

The SWE standards cover a range of the information technologies required – conceptual models (e.g. ISO 19156 'Observations and Measurements'), encoding standards (SensorML, TransducerML), and web

[3] Semantic compliance? Two objects which comply to the same semantics have the same information content, even if their format and/or syntax differs.

service interfaces (Sensor Observation Service, Sensor Planning Service). See Figure 3[4] for a schematic outlining the relationship between these technologies. Examples of working prototypes adopting the SWE standards include the US OpenIOOS consortium[5] and the Tasmanian South Esk Hydrological Sensor Web[6].

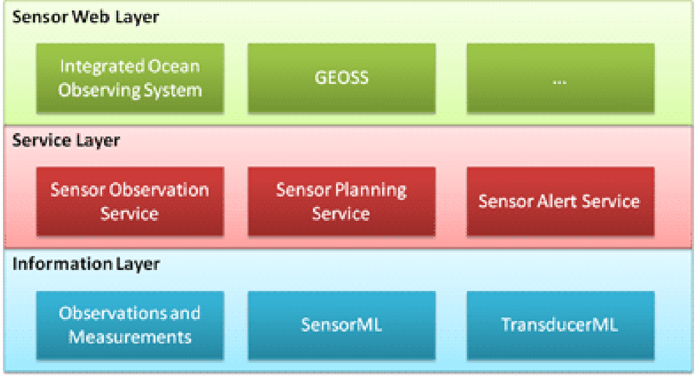

Figure 3: The 'Sensor Web' standards stack (from Woolf, 2009)

A core functional requirement for a sensor web is the ability to aggregate or 'cascade' individual services into a single query and access point. This is addressed by the Sensor Observation Service (see Figure 4), but an important design decision is how best to factor the sensor observations into separate 'observation offerings', each characterised by:

- one or more specific sensor system(s)
- time period(s) for available observations
- sensed phenomena
- locations that are the subject of observation (i.e. 'features of interest')

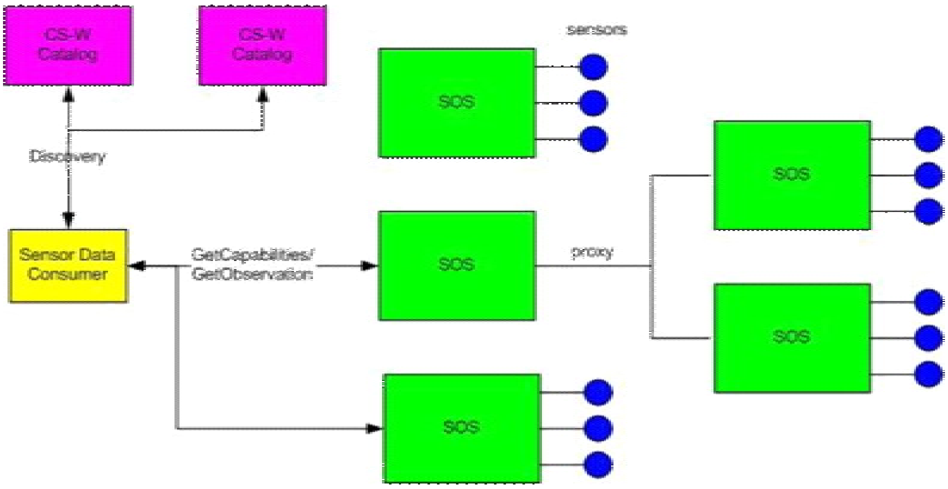

Figure 4: Sensor Observation Service concept of operations (from OGC 06-009r6 'Sensor observation Service')

The answer requires an analysis of the specific sensor web querying requirements – the configured offerings should contain data that are 'dense' in the expected query parameters (i.e. unlikely to result in null results).

[4] Woolf (2009): Building the Sensor Web – Standard by Standard. ERCIM News, 76, 24-25.

[5] US OpenIOOS consortium: http://www.openioos.org

[6] The Tasmanian South Esk Hydrological Sensor Web: http://wron.net.au/au.csiro.OgcThinClient/OgcThinClient.html

3.6 Requirements for Common Data Models

In this section we list the broad semantic requirements of DTC data models, methodologies for constructing data models are discussed in section 3.7 and detailed semantic concepts appropriate for the DTC project in section 4.

Project participants have identified the need for the use and documentation of a number of common protocols such as the Laboratory Calibration Standards (EA and ADAS have these), along with standard sampling methods, field logs, and baseline farm surveys.

Clearly the actual data itself (as listed in section 3.3) needs documenting, as do the details of the equipment used to make measurements, and the codes used to make simulations.

Common data models should use standard vocabularies where possible, and if external vocabularies do not exist, the archival activity should support the development, deployment, and governance of DTC common vocabularies.

All quality control processes should be documented, and the provenance of all datasets should be traceable through prior versions and any processing.

The National Grid Reference datum should be used where possible for geospatial coordinates (possibly in addition to other coordinates where appropriate).

3.7 Methodology for Establishing Common Data Models

Data models themselves should be developed in ways which conform to standards which enhance their applicability, document-ability and potential for re-use. Because of the requirements for standards compliance, this means following the methodology outlined in ISO 19101.

In brief, the ISO19101 methodology outlined in figure 5 begins by defining a “universe of discourse”: the subset of the real world which one is “modelling”, and using a “conceptual” formalism – the Unified Modelling Language (UML) – to describe a conceptual model of the universe of discourse (that is, to construct a UML description of real world feature types in terms of their properties and relationships). In practice one uses an overarching “metamodel” describing how to use UML. for modelling the world in such a way that user communities can construct “application” schema constraining “instances” which describe actual subsets of features (i.e. data records and datasets).

The time taken to construct such data models depends on the complexity of the entities and their relationships, and the size of the communities which have to agree on the properties and vocabularies necessary to describe them. Given the domain of interest to the DTC, if the project were to try to construct ab initio data models, a number of years could be consumed, and it is possible that the data models which resulted would not be suitable for exchanging data with a wider community, hence it is important to exploit community data models which are either already in advanced stages of development or in use. However, such exploitation needs to be tensioned against the necessity to ensure the properties of interest to the DTC community are described (which may involve extensions to community models), the necessity to ensure that the data exchanged can conform (which in practice comes down to a requirement for data producers to be able to generate the required information from the tools they have available), and potential conflicts as to which community models to use.

It will be seen in section 4.3 that the upcoming ISO19156 standard for observations and measurements provides an integrating data model which can be specialised for most of the requirements of the DTC, but it will still be necessary to go through a process early in the project to refine the DTC data model and to identify and populate suitable vocabularies. This process will need to be revisited as the project matures and more data is acquired.

The process envisaged is:

1. The data modelling philosophy for the DTC is outlined (in this document), followed by

2. The identification of the key properties required in the DTC data model(s) (also in this document), followed by
3. The elaboration of the DTC data model with support for the protocol and instrument descriptions required (, and for exploiting WaterML (section 4.5) and CSML (section 4.6). This work will need to be done early in the delivery of the archival solution by a partnership of those delivering the archive and key individuals within the consortia, who will need to devote time to evaluating the models and their data description requirements. Within this activity we envisage the development of example data records which conform to the prototypical schema and a number of cycles of review and refinement.

An outcome of this work is likely to be the establishment of the necessity for interface tools which can convert DTC data instances into vanilla WaterML and CSML documents for data types which can be described by those data models. (A priori, we know that neither schema is suitable for all data types identified in section 3.3, nor does the union of these two schema cover all the required data types, yet both schema will provide interoperability into DTC user communities.)

Additionally:

4. Further work on vocabulary and entity relationships will be required, probably involving the construction of visualisations of vocabulary and entity relationships in data producer workshops. (For example, the EC funded Metafor project[7] has found the construction of mindmaps a very useful activity for constructing vocabularies and data model details).

It is clear that the DTC project will also need to leverage the existing work within CE, where work building on past projects such as LOIS and LOCAR to refine vocabularies for water quality monitoring is being carried out. This work includes dictionaries of chemical determinands but also of lab machines and processes, sample, filtration and storage methods, units of measurement, etc. The identification of these vocabularies, working with the scientists involved, has proved complicated but very beneficial, in particular the separation of measured determinand and processes. The CEH vocabularies are likely to be loaded into the NERC vocabulary server for future use. Although these have been project specific, the expertise of this group will be crucial for the DTC activities, and ideally the DTC vocabularies will augment this work.

[7] The METAFOR project aims to develop a Common Information Model (CIM) to describe climate data and the models that produce it: http://metaforclimate.eu/

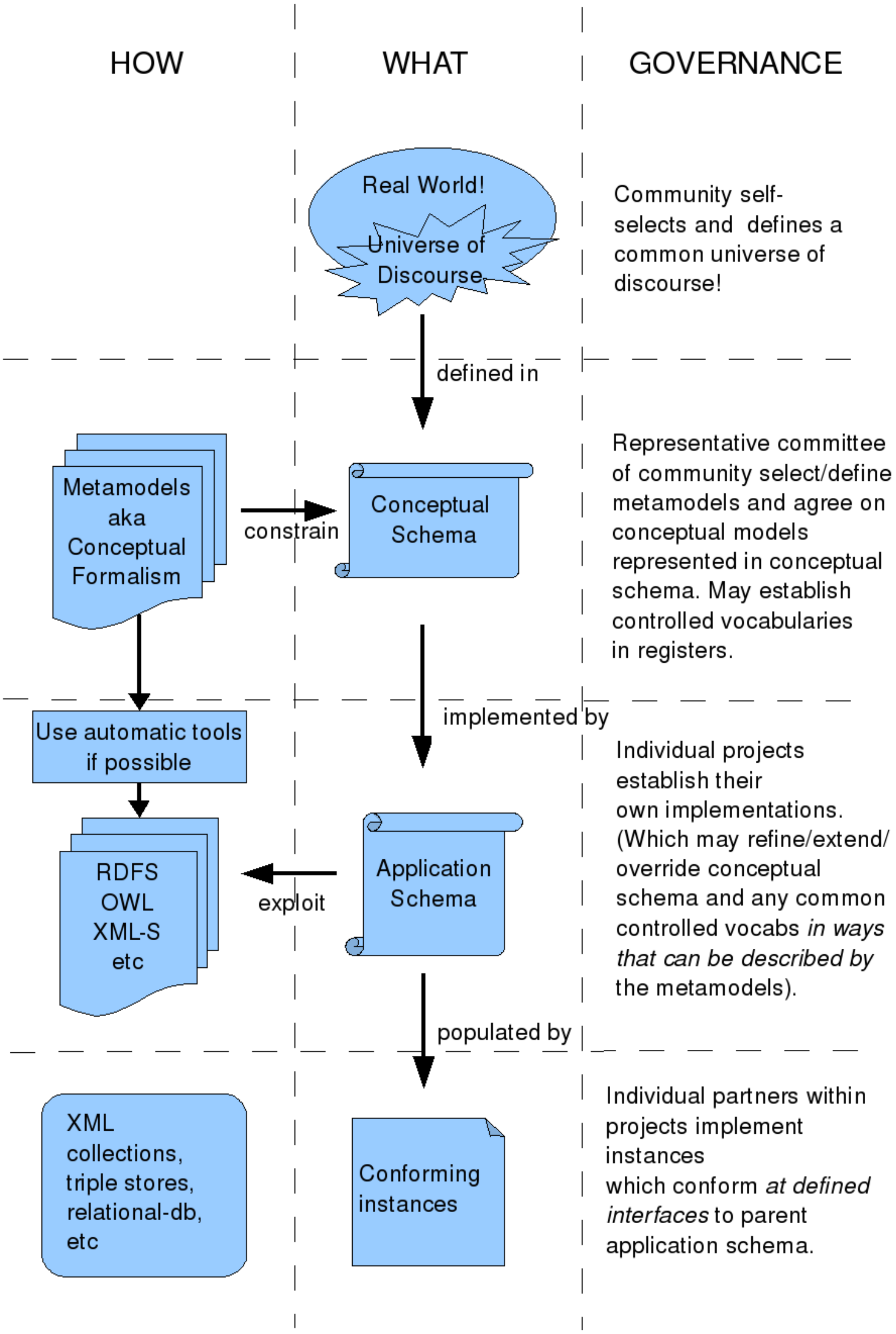
HOW
WHAT
GOVERNANCE
Real World!
Universe of Discourse
Community self-selects and defines a common universe of discourse!
defined in
Metamodels aka Conceptual Formalism
constrain
Conceptual Schema
Representative committee of community select/define metamodels and agree on conceptual models represented in conceptual schema. May establish controlled vocabularies in registers.
implemented by
Use automatic tools if possible
RDFS OWL XML-S etc
exploit
Application Schema
Individual projects establish their own implementations. (Which may refine/extend/ override conceptual schema and any common controlled vocabs *in ways that can be described by* the metamodels).
populated by
XML collections, triple stores, relational-db, etc
Conforming instances
Individual partners within projects implement instances which conform *at defined interfaces* to parent application schema.

Figure 5: The ISO19101 data modelling formalism: proceeding from a formal description of the universe of discourse to instances of feature type descriptions.

3.8 Data Access: Rights, Licenses and Security

There are three key aspects of data access to consider: what data is available, who can access the data, and what systems are in place to enforce access control.

While we believe it is a requirement for the data to be made public, there are some constraints on publication:

1. Personal data cannot be made public: survey results must be anonymous and individual names and addresses must be restricted.
2. Business data cannot be made public: where necessary, and agreed by the data owner, data which is of public interest *can* be extracted from private documents and made available to research teams, and more widely.
3. High volume datasets need access control to protect the archive systems from a "denial of service" attack which attempts to overload the service capacity. Such access control may apply both to download and view services. To this end registration can be used to prioritise access (as oppose to restricting it).
4. Derived datasets are the intellectual properties of those who derive them. However, the expectation of the DTC project is that these should also enter the archive after a suitable embargo period to allow exploitation and academic publication.
5. Similarly, third party datasets may be acquired and archived, which have their own licensing and access control regimes, which will need to be respected.

In addition, there is a clear expectation by the scientific community involved in the research consortia that:

6. Where aspects of the data collection might yield information to research competitors about *experimental design,* access should be temporarily restricted.
7. While access to raw data should be possible, access to quality controlled data should be preferred in systems providing data to users: that is, the systems should clearly indicate to prospective users as to what data is recommended (and if data is not recommended, why not).

Where data access would otherwise be prevented because of privacy of business concerns:

8. Systems should be put in place to anonymise data before download and visualisation as required (such systems may extend to random scattering in time and/or space so that a dataset has the required statistical properties, but the individual data points are no longer valid).

Apart from the situations identified above:

9. No embargo (temporary or protracted) on data acquired by routine systems and/or analysis should be supported.

Data ownership and accompanying intellectual property rights can be difficult to ascertain if appropriate provenance is not available, to that end, the ownership of all data should be established as it is collected. (Note that the legal situation as to who owns the IPR embodied in databases is difficult to establish, so clearly distinguishing dataset IPR makes life much easier).

Regardless of how data access occurs, data users will acquire data, and Defra (or the data holders) may wish to assert ownership of the data (and some re-use criteria) and they will have some legal liability as to the fitness of purpose of the data. Even if there are no re-use restrictions, Defra should waive liability for use of the data to the fullest extent permitted by law. Note that such waivers can only be demonstrated to have been entered into if it can be shown that any data portals did not allow data access without the user being aware of such a waiver.

The DTC project, via its Data policy, should provide clarity on who owns each archival dataset, and who has the right to use the data and under what circumstances. It is assumed that the default position is that Defra will own the IPR of any data created by the DTC project and that this will be reflected in the

Programme's Data Policy.

3.9 Management and Governance: the Data Management Plan

The expectations of data producers, Defra and the providers of the data archive should be documented in a formal data management plan – to be a living document update throughout the lifetime of the project. It should detail how and when data will be acquired and processed, and where and when it should be archived. It should be created in advance of data collection, and be approved by all parties. It should codify the access control expectations outlined above.

The data archive of the DTC project cannot be simply contracted out to a service provider: both the DTC Project Board and the individual research consortia will have to remain engaged in data management. Someone needs to be responsible for ensuring this occurs.

The initial development of controlled vocabularies is something that would need to be done as part of the refinement of the data models, however, ongoing management of appropriate vocabularies is necessary. This is something that might fall within the remit of one of the NERC centre surveys (for example, the management of the international climate-forecast conventions is managed by the NERC National Centre for Atmospheric Sciences).

3.10 Key External Relationships

In this section we list assumptions of the relationship with other projects and stakeholders which will or could influence or inform either the data archival or the data model(s) or both, and which are not mentioned elsewhere.

The NERC Virtual Observatory Project (VO) – is expected to deliver (on a two year timescale) visionary prototypes of informatics systems which support the systematic study of the complex interactions that make up the soil-water system. The timescales of these projects are rather different, and so one might expect the main relationships to be data transfers and possibly the accommodation of interface requirements from the VO upon the archive at some later date.

Defra SPIRE: The internal DEFRA spatial data infrastructure is currently not accessible outside of Defra, and does not take real time data. However, clearly there may be some level of engagement with SPIRE which would need to be evaluated.

ERFF (EOF), The Environmental Research Funders Forum and the UK Environmental Observation Framework: The DTC project will need to inform EOF as to what is being measured where, and feed into data discovery. This should be handled as a minor perturbation on the requirements of the metadata models.

SEIS: The European Shared Environmental Information System: The Shared Environmental Information System (SEIS) is a collaborative initiative of the European Commission and the European Environment Agency (EEA) to establish together with the Member States an integrated and shared EU-wide environmental information system. This system would tie in better all existing data gathering and information flows related to EU environmental policies and legislation. It will be based on technologies such as the internet and satellite systems and thus make environmental information more readily available, transparent, and easier to understand to policy makers and the public.

According to the SEIS concept, environmentally-related data and information will be stored in electronic databases throughout the European Union. These databases would be interconnected virtually and be compatible with each other. The proposed SEIS is a decentralised but integrated web-enabled information system based on a network of public information providers sharing environmental data and information. It will be built upon existing e-infrastructure, systems and services in Member States and EU institutions.

data.gov.uk: The UK government seeks to make as much data as possible open for reuse. It is likely that during the DTC project there will be a requirement to make some DTC data available via data.gov.uk.

This should be relatively easy provided the data model procedures outlined elsewhere in this document are followed, but resourcing for such an effort would need to be quantified after the issuing of the main data archival contract.

The National River Flow Archive (NRFA, at the Centre of Ecology and Hydrology):

The NRFA at the Centre for Ecology & Hydrology (CEH) is the UK's focal point for hydrometric data, providing stewardship of, and access to, over 50,000 years' daily and monthly flow data for some 1500 gauging stations nationally. Maintenance of the NRFA involves routine collation, quality control, and archiving of river flow data from UK measuring authorities, namely, the Environment Agency (EA), in England and Wales, the Scottish Environment Protection Agency (SEPA), and the Rivers Agency in Northern Ireland. A comprehensive information delivery infrastructure has been developed that provides, in addition to quality controlled flow data, easy web-based access to a range of summary and derived information. The NRFA also hosts the National Hydrological Monitoring Programme (NHMP) in collaboration with the British Geological Survey. The NHMP capitalises on the NRFA and the national groundwater archives to provide authoritative commentaries on prevailing hydrological and water resources conditions and issues. Outputs from the programme, including monthly Hydrological Summaries for the UK, are extensively exploited by policy makers, planners and the research community. The NRFA fulfils CEH's formal obligations to Defra and the devolved administrations (National Assembly of Wales, Scottish Parliament, Northern Ireland Assembly) relating to the supply of data and information on UK water resources.

3.11 Summary of Enterprise Level Recommendations

This section provides a list of enterprise level recommendations for the DTC archiving activity.

1. When issuing an eventual specification for data archival, Defra should make the scope of intended users clear, so as to both manage expectation and maximize benefit for the desired stakeholders.
2. Any provider should include in their bid details of their exit plan for the data as to what would happen when (not if) Defra withdraw funding for the DTC archive.
3. Defra should include all points listed in Table 1 (section 3.2) in the eventual specification of the requirements of the data archive.
4. The DTC project should identify the most important third party datasets and then start negotiation for access as soon as possible. It may well be that the target datasets already conform to international standards, have managed vocabularies and fit within agreed data models. However this is unlikely and it is recommended that the programme sets aside some contingency funds to help with establishing access to specific datasets.
5. Both raw data and quality assured data should be preserved in perpetuity in the Data Archive.
6. The DTC data models are developed conform to both INSPIRE specifications and data.gov.uk requirements.
7. Defra requires project participants to engage with data model specification efforts throughout the project.
8. Open source software is used wherever dependency on data model extensibility is expected.
9. That a data policy document be drawn up based on the points listed above, and agreed by the DTC Project Board (and thus all DTC project participants) which should include clear guidelines as to when embargo periods are suitable, and for how long.
10. That the archive uses a registration system to control access to data, and potentially to some (but perhaps not all) data visualisation functionality.
11. That the DTC project construct a suitable data license, and ensure that all data users sign up to the terms and conditions of the license before access to the data is provided. (This last implies that all data portals implement some method of ensuring that a license agreement has been entered into).

12. The National Grid Reference datum should be used where possible for geospatial coordinates (possibly in addition to other coordinates where appropriate).
13. The DTC project should appoint a Programme Data Manager who will be responsible for coordinating data management activities. This person should be responsible for liaising with those responsible for local data archives (should they exist), liaising with the central Data Archive, establishing, implementing and monitoring programme data policy and establishing mechanisms to support scientists in getting the most out of their data.
14. Each Consortium involved in the DTC should appoint an individual Data Officer who has responsibility over how that Consortium interacts with the wider data management activities within DTC.
15. Defra should negotiate with the Centre for Ecology and Hydrology to ensure that any specific DTC vocabularies are incorporated in the vocabulary management activities at CEH.

4 INFORMATION VIEWPOINT

The information viewpoint addresses the semantics of the information and the information processing performed. It describes the interfaces and logical elements that need to be received, stored, ingested and manipulated within user portals and user services. In the computational viewpoint we further consider the syntax of the key information objects (such as file formats).

A key requirement from the enterprise viewpoint (section 3) is that the DTC project conforms to the Sensor Web Enablement (SWE) standards. The SWE standards of interest are summarised in figure 6.

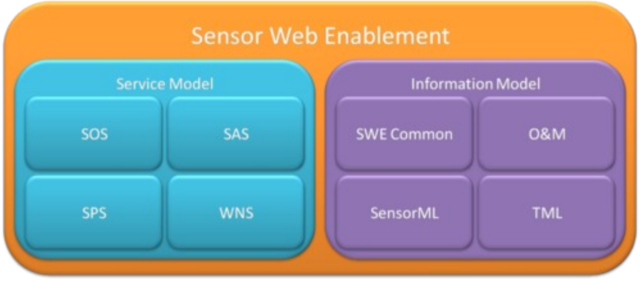

Figure 6: Key components of the Sensor Web Enablement: SOS (Sensor Observation Service), SAS (Sensor Alert Service), SPS (Sensor Processing Service) and WNS (Web Notification Service) along with SWE Common (basic encodings), O&M (the Observations and Measurements model, soon to be ISO19156), SensorML (the Sensor Markup Language) and TML (Transducer Markup Language).

Of particular importance to the DTC project are the SOS (Sensor Observation Service, which provides methods to get observations across the web from sensors); O&M (Observations and Measurements, which provides a structure for describing observations and simulations), SWE Common (which provides encoding rules for O&M observations), and SensorML (for describing the characteristics of particular instruments). Because of the wide range of activities within the DTC project, these standards are not in and of themselves a sufficient collection to construct a data model for the DTC, but they are a good place to start.

In the remainder of the description of the information viewpoint for DTC data handling, we begin by introducing some key concepts necessary to construct a useful data model for the DTC, introduce the key relevant standards, and use these to identify a suitable framework for the semantics of data objects in the DTC project. The procedure to finalise this framework was discussed in section 3.7 of the enterprise viewpoint.

4.1 A taxonomy of DTC data types

The key data streams introduced in section 3.3 will result in the following data *types*:

1. *Feature* descriptions: describing some observable *property* of a catchment, location or region within it.
2. *Activity* descriptions: describing something that was *done* as opposed to something that *is.*
3. *Property* descriptions: via feature descriptions, or numeric values, or selected (individually or in combinations) from *vocabularies* (also known as categories), or free text.
4. The names of some features, activities and properties and are likely to be important in themselves and form part of *vocabularies.*
5. Numeric values consist of essentially two types: those which are single-valued properties of features, and those which consist of arrays of numbers which have some functional relationship a

spatial and/or temporal distribution. The latter are known as *coverages* which have a *range* of values over some *domain.*

Understanding Coverages

Two example of the coverages would be a) time series of values at a point, and b) maps of rainfall on a spatial grid.

Note that a coverage may be constructed ab initio from a single multi-valued observation or simulation, or *aggregated* from a number of single valued properties (in which case *decomposing* it back into feature properties is trivial) or it may be constructed via some sort of computation applied to single valued properties (e.g. interpolation onto an equally spaced grid), in which case decomposition is not possible. In the latter case, the original data and the computation technique are metadata associated with the coverage data, in the former case, the original data forms part of the coverage data. Clearly the DTC project will need to consider both cases and manage the provenance of data appropriately.

ISO19123[8] provides a formalism for describing coverages, but in practice, most communities handle coverages by writing coverage data into files formatted in particular ways which map onto the semantics of the coverage of interest – for example, the climate and forecasting community make heavy use of the NetCDF format to store and describe the structure and properties of four-dimensional space-time grids. Distinguishing between the semantics of coverages and the format with which they are stored will be a crucial part of the DTC information model – and this is discussed in 5.1.

4.2 Taxonomy of Metadata

The discussion of coverages in the previous section introduces some of the complexity that needs to be dealt with in metadata, describing provenance and formats. Additional complexity arises because for some applications, what one group call metadata (because it is extraneous to their own analysis), another group will consider to be an integral part of the data (because it is incorporated directly in their results).

In this section we introduce a broad taxonomy of the types of metadata which need to be encapsulated within the data models required, and briefly describe the consequences for engineering systems which will support the construction, maintenance, and interrogation of the relevant information systems. Details of how to deliver those systems would of course be part of any eventual tender to deliver the DTC archive.

Figure 7 introduces five major classes of metadata: from A to E[9]:

1. Tools which manipulate data need to exploit metadata describing the layout and structure of the data, phenomenon names etc. (this metadata appears as A-type metadata in Figure 7)
2. There is an intermediate level of detail which is of use to potential users of data, providing enough detail and context for someone else to use the data without contacting the originator. This sort of metadata is both enough to enable the choice between similar datasets, and to provide adequate provenance. Some of this metadata could be constructed using common cross-disciplinary standards (this metadata appears as B-type metadata in Figure 7), and some requires more specific detail encoded using discipline specific standards (termed as E-type metadata in Figure 7).
3. There is a high level "catalogue" level of detail, which is useful for organizing information and providing data "discovery" (usually via some sort of directed search exposing key parameters, services or characteristics). This sort of metadata is usually only enough to advertise the presence and potential usefulness of data. (This metadata appears as D-type metadata in Figure 7). INSPIRE metadata is "D-type" metadata.
4. Scientists and data users often classify data, using annotations and citation, and capturing such

[8] ISO19123: Geographic Information – Schema for coverage geometry and functions.

[9] An extended discussion of this material appears in Lawrence *et. al.*, 2009 in Phil. Trans. Roy. Soc. A, 367, 1003 - 1014. doi:10.1098/rsta.2008.0237.

“ranking” material to define the “character” and “fitness” of the data is an important activity. (This metadata appears as C-type metadata in Figure 7).

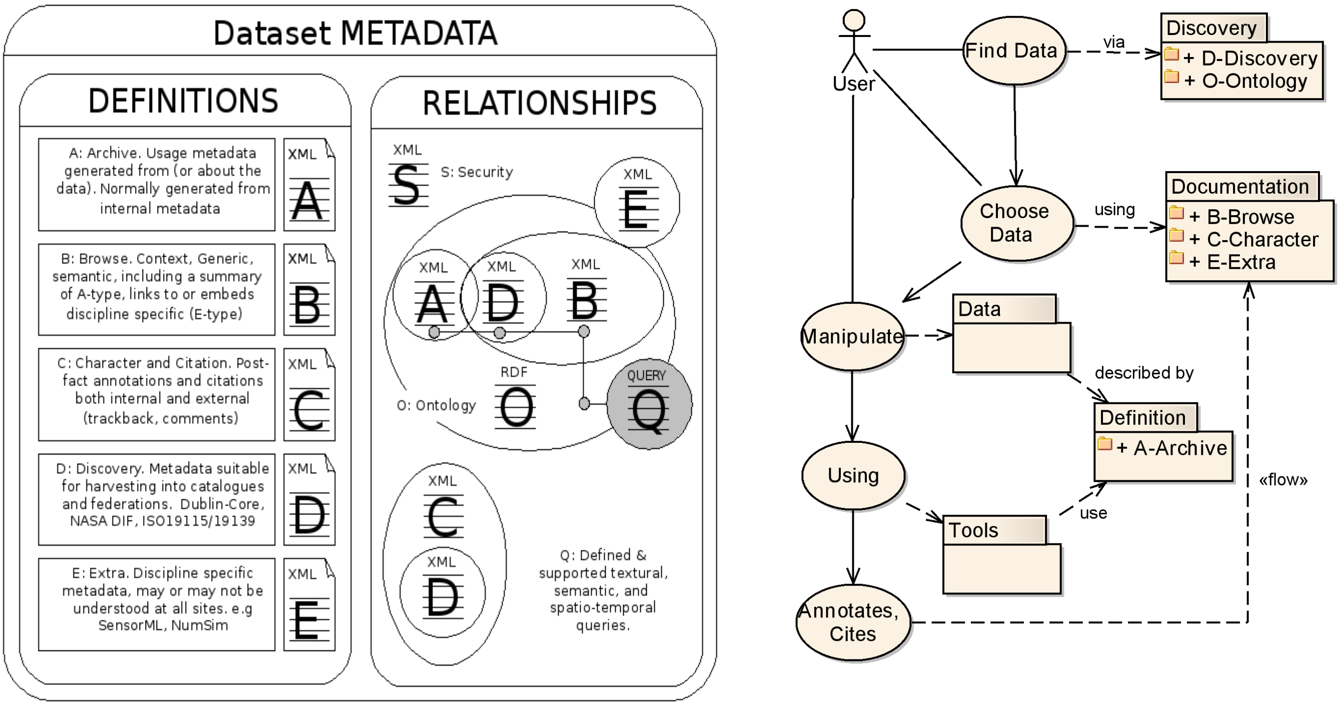
Dataset METADATA
DEFINITIONS
A: Archive. Usage metadata generated from (or about the data). Normally generated from internal metadata
B: Browse. Context, Generic, semantic, including a summary of A-type, links to or embeds discipline specific (E-type)
C: Character and Citation. Post-fact annotations and citations both internal and external (trackback, comments)
D: Discovery. Metadata suitable for harvesting into catalogues and federations. Dublin-Core, NASA DIF, ISO19115/19139
E: Extra. Discipline specific metadata, may or may not be understood at all sites. e.g SensorML, NumSim
XML A
XML B
XML C
XML D
XML E
RELATIONSHIPS
XML S
S: Security
XML E
XML A
XML D
XML B
RDF O
O: Ontology
QUERY Q
XML C
XML D
Q: Defined & supported textural, semantic, and spatio-temporal queries.
User
Find Data
via
Discovery
+ D-Discovery
+ O-Ontology
Choose Data
using
Documentation
+ B-Browse
+ C-Character
+ E-Extra
Manipulate
Data
described by
Definition
+ A-Archive
Using
Tools
use
«flow»
Annotates, Cites

Figure 7: Categories of metadata, and usage sequence. Users begin by finding data in discovery services, perhaps exploiting an ontology to locate things via a traversal from their vocabulary to that used to store the data, then they choose data by browsing between datasets and examining contextual and other detailed metadata. Having obtained datasets, they manipulate data using tools (or home-grown software) which are informed by the metadata describing the layout and meaning of the data objects themselves. Ideally, having analysed the data, they cite the data in publications and/or annotate the data collections directly.

The various types of metadata required to manage data are normally created by different individuals, using different tools at different times. Key tools that need to exist include:

1. Graphical user interfaces (GUIs) which are designed so that humans can enter both selections from controlled vocabularies describing the data and free text.
2. Automatic metadata production tools which either:
   a) run over data products produced by instruments or software, run with or without human intervention (possibly, in the former case, including the addition of human-generated metadata),

   or

   b) are part of the software inherent in the instrument or software so that the original human operator has configured the instrument or software to produce the appropriate metadata. (Inevitably, with time, this class of metadata needs to be supplemented by metadata created by one of the previous methods, as metadata requirements generally evolve faster than the internal capability of instruments and production software.)
3. Vocabulary services exposing controlled vocabularies which can be used by the GUIs to provide vocabulary selections, and by all metadata tools to validate metadata entry which is expected to be from controlled vocabularies.
4. Tools to create controlled vocabularies, which in practice means more GUIs, along with tools that can mine vocabularies from free text.

5. Tools which manage vocabularies and expose appropriate interfaces (generally Web Service[10] interfaces unless these tools are to be in tightly coupled software systems, and not intended to expose the vocabularies to wider communities).
6. Tools which manage metadata (in various schema) and expose appropriate interfaces (generally Web Service interfaces unless these tools are to be in tightly coupled software systems, and not intended to expose the metadata to wider communities).

4.3 Relevant Information Standards: Observations and Measurements

A key part of the data modelling problem for DTC is establishing a relationship between the data created by observations and the procedures associated with those observations. To that end, the upcoming ISO19156 observations and measurements (hereafter O&M) standard provides a good framework, but it will be seen that O&M alone is not enough, it needs a number of extensions to be usefully applied within the DTC, some of which will be DTC specific, and some of which will conform with international best practice in DTC relevant fields. In addition, O&M doesn't yet provide a normative specification of how to encode and share instances of observations, although the OGC is about to release a canonical XML encoding specification.

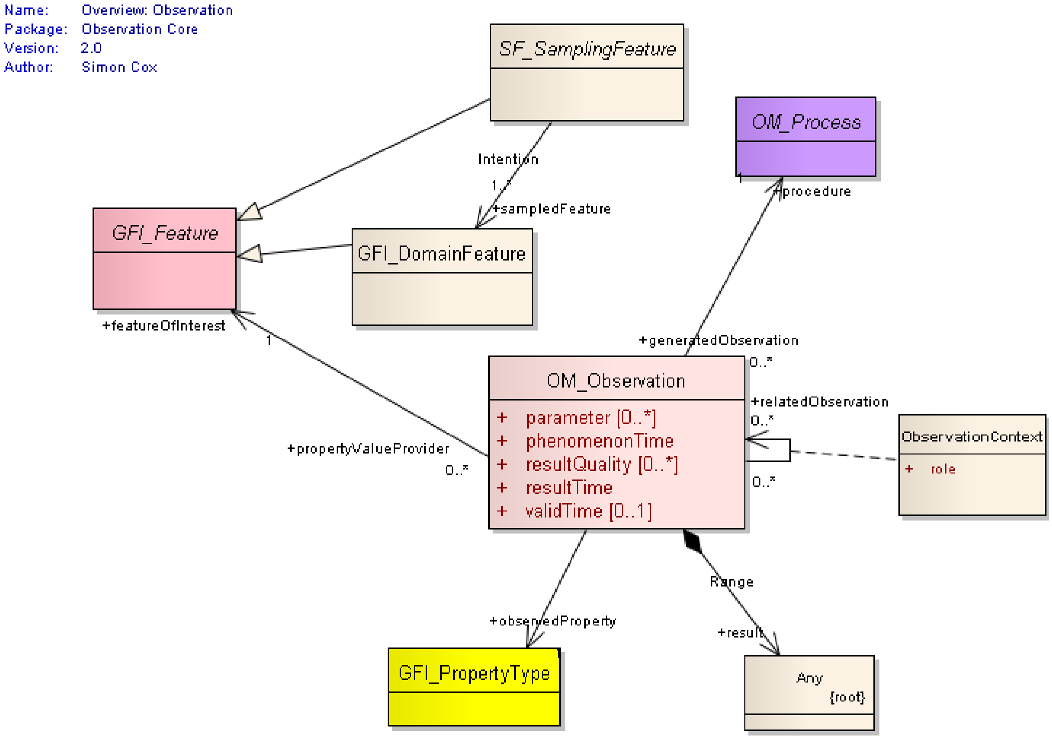

Figure 8: The (upcoming) ISO19156 observation: provides a definition of an observation in terms of what is measured (the featureOfInterest and the observed property), how it is measured (the procedure) and the result.

The basic concept of O&M (see figure 8) is that an an *observation* of a *property* of a *feature of interest*, is made using a *process* which produces a *result* .

Along with the overarching concepts of measurement, O&M also addresses issues associated with the distinction between directly measuring a specific domain feature of interest, and taking (or using) a sample of the feature of interest. In most cases of environmental interest, we engage with a "Sampling Feature", that is a feature which itself represents the real feature of interest in the environment. For example,

- When measuring river flow rate, we can't decant the entire river through a flow meter, so we

[10] The term *Web Service* describes a service exposes a clearly defined programmatic interface across the network that allows multiple clients to connect to in order to access its functionality.

sample it at various points in a profile across the river, or

- When measuring the quantities of a particular pollutant in the atmosphere, we might take a physical sample and take it back to a laboratory for detailed measurement.

In both cases, a sampling feature (the profile and the physical sample) have stood as proxies for the real feature of interest (the river and the atmosphere) – see Figure 9.

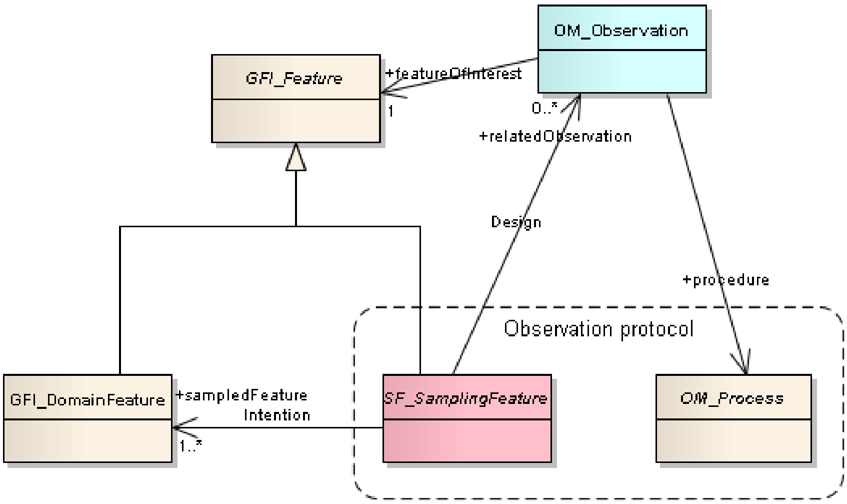

Figure 9: Observations use a procedure to estimate properties of either the feature of interest itself, or of a sampling feature standing in for the feature of interest.

One of the most important types of sampling feature is the specimen, because the data provenance need to describe what has happened (and where) to a specimen before the results are obtained. In principle, the specimen which is analysed to deliver the final estimate of the relevant property of the feature of interest can be one which has been manipulated several times. Keeping details of those manipulations is a crucial part of having faith in the results obtained by the analysis procedure (see Figure 10).

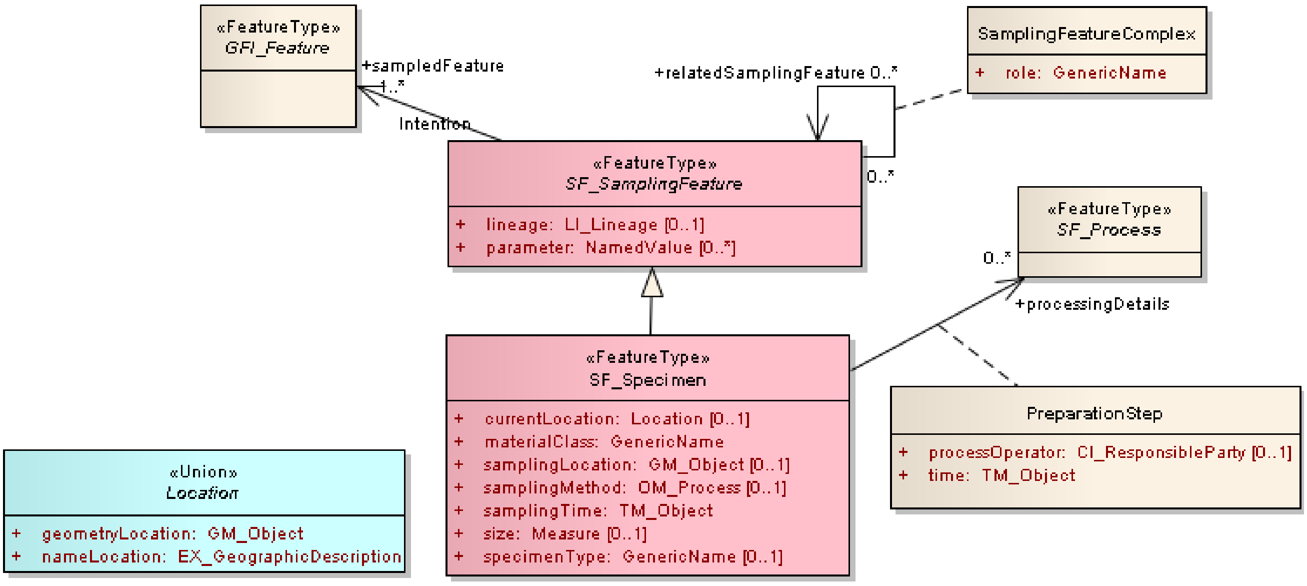
«FeatureType»
GFI_Feature
+sampledFeature
1..*
Intention
+relatedSamplingFeature 0..*
SamplingFeatureComplex
+ role: GenericName
«FeatureType»
SF_SamplingFeature
+ lineage: LI_Lineage [0..1]
+ parameter: NamedValue [0..*]
0..*
«FeatureType»
SF_Process
0..*
+processingDetails
«FeatureType»
SF_Specimen
+ currentLocation: Location [0..1]
+ materialClass: GenericName
+ samplingLocation: GM_Object [0..1]
+ samplingMethod: OM_Process [0..1]
+ samplingTime: TM_Object
+ size: Measure [0..1]
+ specimenType: GenericName [0..1]
PreparationStep
+ processOperator: CI_ResponsibleParty [0..1]
+ time: TM_Object
«Union»
Location
+ geometryLocation: GM_Object
+ nameLocation: EX_GeographicDescription

Figure 10: The specimen in observations and measurements: a specific sampling feature, which may itself be the product of manipulations of previous specimens.

4.4 Relevant information Models: Existing Profiles of O&M

There are a number of existing profiles of O&M in various stages of maturity from inception to deployment which have been (or are) growing from specific discipline based communities.

Of particular relevance to the DTC, in no particular order, are:

1. WaterML2: The Water Markup Language was initially developed by the US Consortium of Universities for the Advancement of Hydrological sciences Inc (CUASHI), and is now being migrated to a standards compliant V2 by an international community. See section 4.5.
2. The Geosciences Markup Language (GeoSciML) has been developed by a consortium of geological surveys to meet an objective to develop and implement a data model for interoperability of geoscientific information between (particularly) national geological surveys. GeoSciML covers geological units, earth materials (lithologies), geological structures, boreholes, and in the latest version is addressing geological samples and measurements (exploiting O&M). GeoSc9ML is not further discussed here, but is in many ways the most mature exposition of the domain modelling philosophy espoused here, and it is obvious that some of the measurements and quantities of interest to the DTC should be handled as GeoSciML documents.
3. The Climate Sciences Modelling Language (CSML) was initially developed by the NERC DataGrid project, and provides a number of sampling features suitable for use in describing measurements and simulations of timeseries, trajectories and grids. CSML finds use mainly as “A” type metadata. See section 4.6.
4. MOLES: The Metadata Objects for Linking Environmental Sciences (MOLES) were also developed by the NERC DataGrid project, aimed at filling a gap in the “B” type metadata spectrum. It will be seen in section 4.7 that MOLES is significantly more immature than the other profiles of O&M discussed here, but it provides a model and objective which is very similar to that of the DTC project, in that it was explicitly aimed at providing a common vernacular for data described in more detailed discipline specific metadata schema.

The problem for the DTC data model is that they are all likely to have relevance to the DTC activities, but even in aggregate they will not cover all the attributes required, so the DTC archive solution will need to specialize O&M to cover all the attributes need by the DTC community, AND, support export of data using these community standards.

4.5 Relevant Information Models: WaterML2.0

WaterML2.0 is currently under heavy development, being an attempt to upgrade the existing WaterML language developed primarily in the CUASI community to a standards compliant activity which is inclusive of a much wider international community.

There is currently little available outside the WaterML development group describing the shape of WaterML, but Taylor et.al. 2010[11] provided a description at the May 2010 European Geosciences Union General Assembly. WaterML2.0 exploits the sampling features package of O&M to deal with sampling parts of a water body, taking samples to laboratories etc, and so matches well to the aims of the DTC archive activity. However, the project (at least in the standards compliant version) is relatively new, and a finalised standard is unlikely to be available for the DTC project. Nonetheless, it should be possible to track the evolution of WaterML, and exploit it within the DTC archiving activity.

[11] Peter Taylor, Gavin Walker, David Valentine and Simon Cox WaterML2.0: Harmonising standards for water observations data. EGU2010 Poster ID-7680

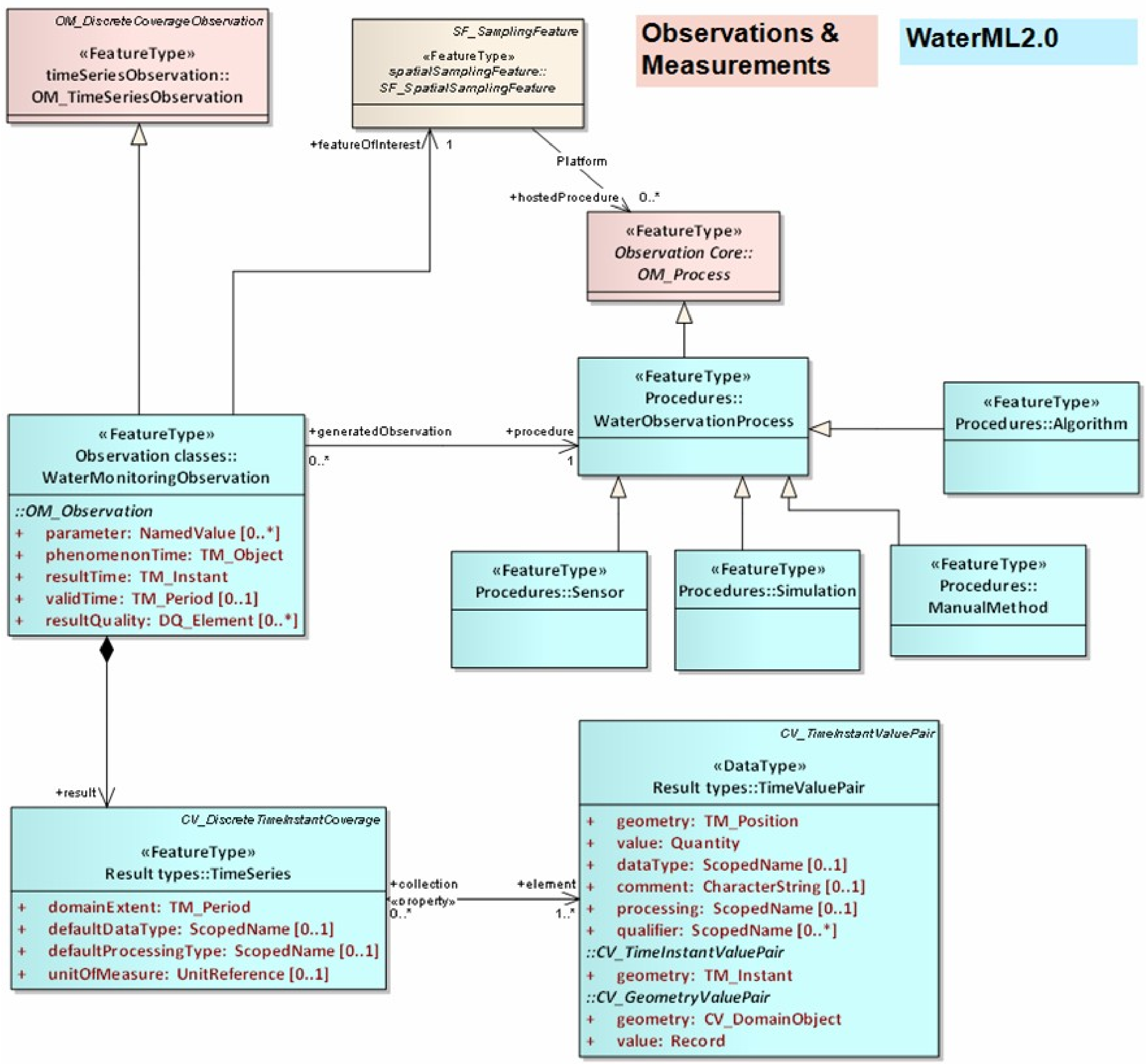

Figure 11: Some aspects of the upcoming WaterML2.0 model. The top level classes (shown in red) are those defined by Observations and Measurements, where the blue classes indicate specialisations in WaterML2.0. The figure shows the relationship between the observation, the process used in making the observation and the structure of the results of the observation (a time series).(This figure is from Taylor et.al. Op Cit, and was kindly provided by the lead author).

4.6 Relevant Information Models: CSML

The Climate Science Modelling Language (CSML) captures a pattern that occurs very often in environmental applications – an observation on a 'spatial sampling feature' (e.g. point, curve, surface) produces a 'coverage' result. Moreover, temporal aspects of such sampling are not fully captured by the spatial sampling geometry – e.g. a sampling curve is used for both a single trajectory and a time-series of repeat soundings. CSML therefore defines a number of specialised sampling features with associated coverage geometries for a set of sampling patterns that occur widely in the atmospheric and oceanographic (i.e. 'climate') sciences.

The current version of CSML is very nearly O&M compliant, and a completely O&M compliant version (CSML V3) is about to be released.

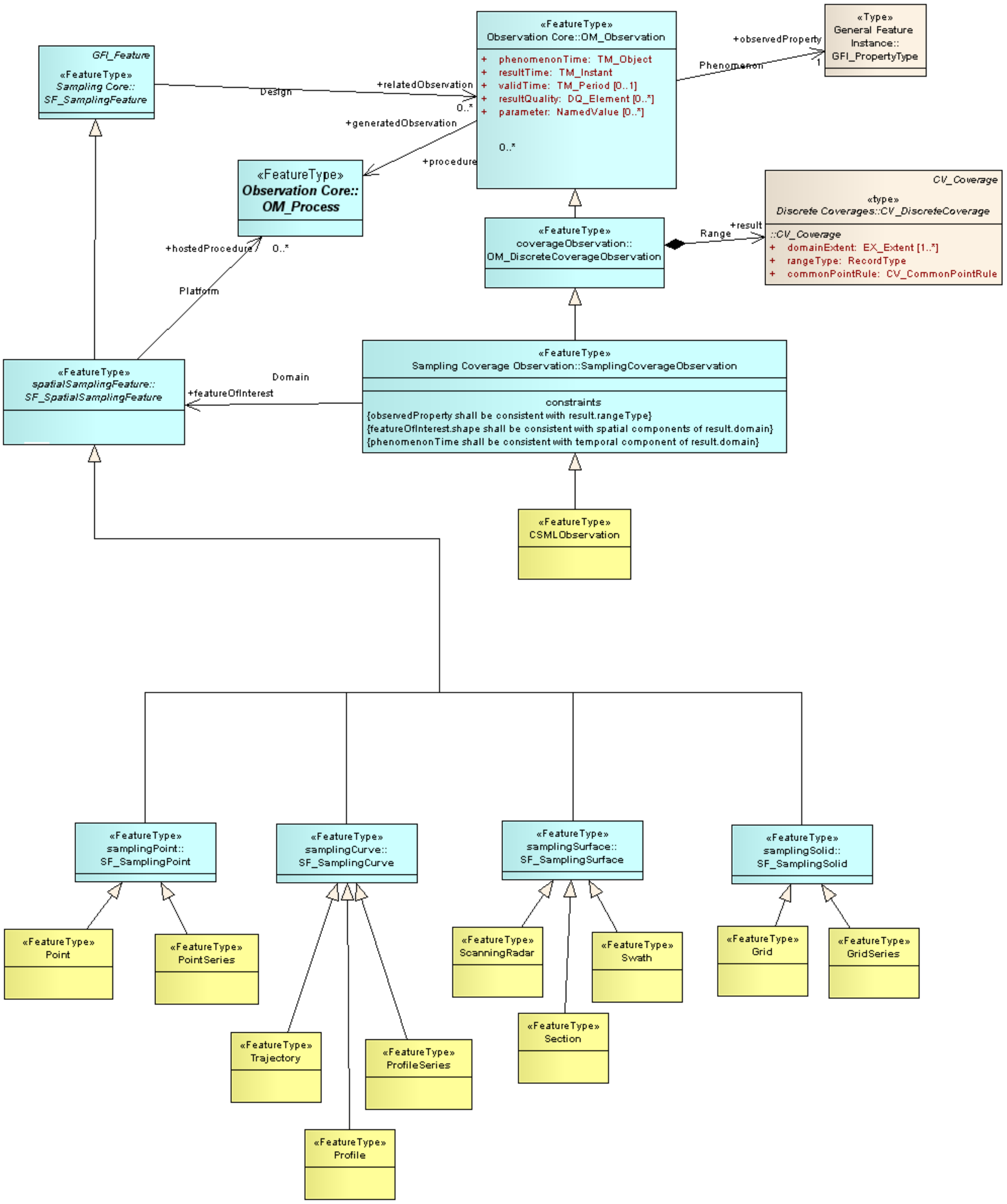

Figure 12: Key aspects of CSML, showing specialisations of the sampling features and the place of a CSML observation specialisaiton.

4.7 Relevant Information Models, MOLES: Metadata Objects for Linking Environmental Sciences

Following the taxonomy outlined earlier, we see that the process of obtaining the right information generally involves a discovery step using “D” metadata, followed by a “navigation” or browse step exploiting “B” metadata, before the use of “A” metadata in manipulation. The NERC datagrid project (http://ndg.nerc.ac.uk) has developed a discipline independent browse profile of O&M which exploits ISO19115-part2 (metadata for imagery and gridded datasets) to address data provenance and provide hooks into the discipline specific metadata. This profile is the third version of the Metadata Objects for Linking Environmental Sciences (MOLES) – a previous version is operational within CEDA, but not suitable for wider use. A prototype of version 3 was delivered in 2009, and work is currently under-way by NERC to make it more fit for purpose – particularly in the areas of interest to the DTC.

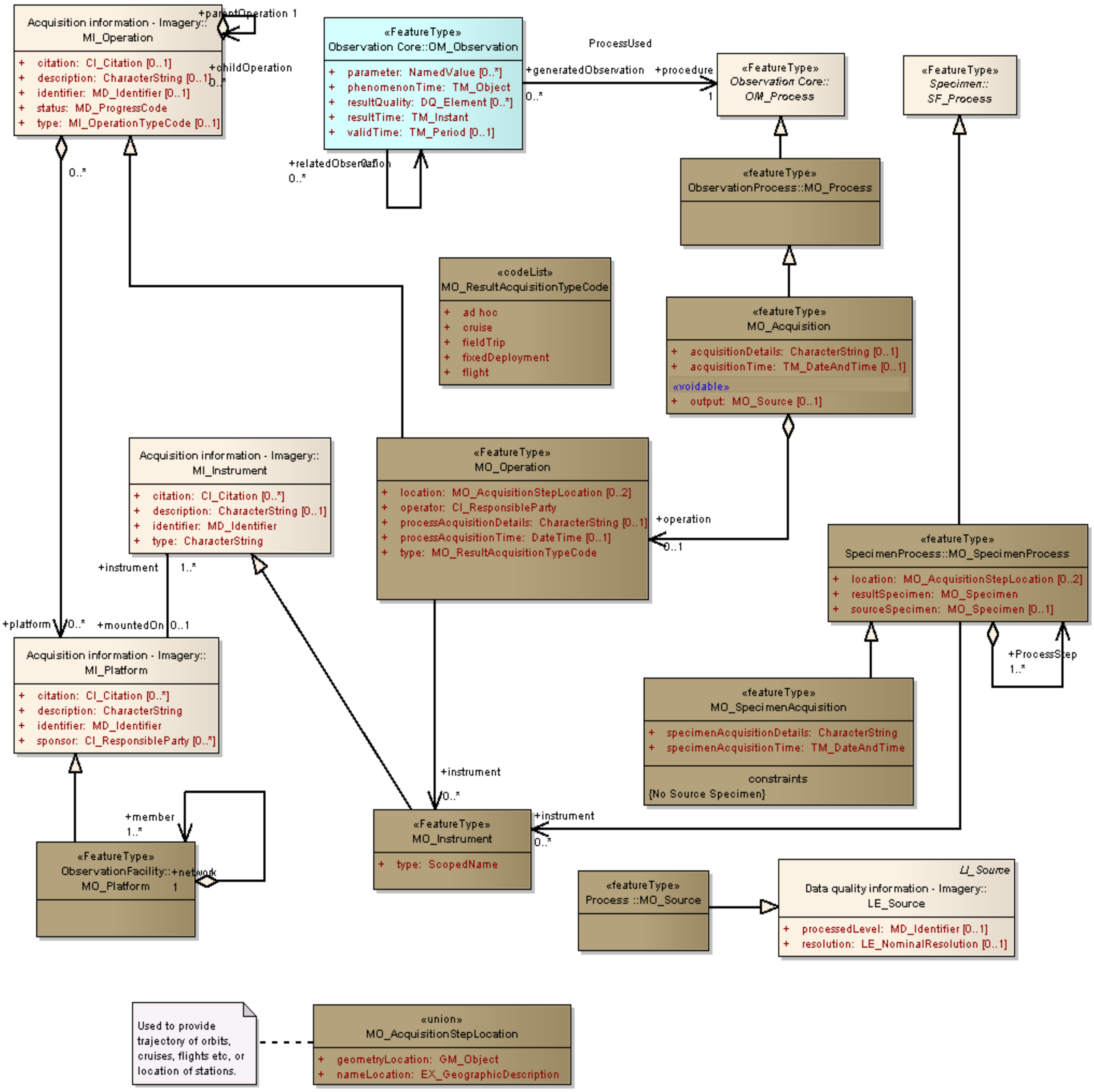
Acquisition information - Imagery::
MI_Operation
+ citation: CI_Citation [0..1]
+ description: CharacterString [0..1]
+ identifier: MD_Identifier [0..1]
+ status: MD_ProgressCode
+ type: MI_OperationTypeCode [0..1]
+parentOperation 1
+childOperation
0..*
0..*
«FeatureType»
Observation Core::OM_Observation
+ parameter: NamedValue [0..*]
+ phenomenonTime: TM_Object
+ resultQuality: DQ_Element [0..*]
+ resultTime: TM_Instant
+ validTime: TM_Period [0..1]
+relatedObservation
0..*
ProcessUsed
+generatedObservation
0..*
+procedure
1
«FeatureType»
Observation Core::
OM_Process
«FeatureType»
Specimen::
SF_Process
«featureType»
ObservationProcess::MO_Process
«codeList»
MO_ResultAcquisitionTypeCode
+ ad hoc
+ cruise
+ fieldTrip
+ fixedDeployment
+ flight
«featureType»
MO_Acquisition
+ acquisitionDetails: CharacterString [0..1]
+ acquisitionTime: TM_DateAndTime [0..1]
«voidable»
+ output: MO_Source [0..1]
Acquisition information - Imagery::
MI_Instrument
+ citation: CI_Citation [0..*]
+ description: CharacterString [0..1]
+ identifier: MD_Identifier
+ type: CharacterString
«FeatureType»
MO_Operation
+ location: MO_AcquisitionStepLocation [0..2]
+ operator: CI_ResponsibleParty
+ processAcquisitionDetails: CharacterString [0..1]
+ processAcquisitionTime: DateTime [0..1]
+ type: MO_ResultAcquisitionTypeCode
+operation
0..1
«featureType»
SpecimenProcess::MO_SpecimenProcess
+ location: MO_AcquisitionStepLocation [0..2]
+ resultSpecimen: MO_Specimen
+ sourceSpecimen: MO_Specimen [0..1]
+ProcessStep
1..*
+instrument
1..*
+platform
0..*
+mountedOn
0..1
Acquisition information - Imagery::
MI_Platform
+ citation: CI_Citation [0..*]
+ description: CharacterString
+ identifier: MD_Identifier
+ sponsor: CI_ResponsibleParty [0..*]
«featureType»
MO_SpecimenAcquisition
+ specimenAcquisitionDetails: CharacterString
+ specimenAcquisitionTime: TM_DateAndTime
constraints
{No Source Specimen}
+member
1..*
«FeatureType»
ObservationFacility::
MO_Platform
+network
1
+instrument
0..*
«FeatureType»
MO_Instrument
+ type: ScopedName
+instrument
0..*
«featureType»
Process ::MO_Source
LI_Source
Data quality information - Imagery::
LE_Source
+ processedLevel: MD_Identifier [0..1]
+ resolution: LE_NominalResolution [0..1]
Used to provide trajectory of orbits, cruises, flights etc, or location of stations.
«union»
MO_AcquisitionStepLocation
+ geometryLocation: GM_Object
+ nameLocation: EX_GeographicDescription

Figure 13: Draft MOLES 3.4 structure of data acquisition steps. Included here to show how MOLES exploits both O&M and ISO19115-2. While the details are expected to change in V3.4 final, the basic concept of splitting processes between acquiring specimens and results, and the use of MOLES specialisations of MI_Platform, MI_Instrument and MI_Operation, are robust.

4.8 Relevant Standards: INSPIRE and Gemini

INSPIRE is a directive of the European Parliament which has been incorporated into British law which aims to establish an "Infrastructure for Spatial Information within Europe" which will help support Community environmental policies and activities. Data interoperability and data sharing are prime objectives for INSPIRE and these are underpinned by a specification for metadata used for Data Discovery within INSPIRE (INSPIRE mandates other services beyond discovery too, but Discovery is the most mature specification).

The INSPIRE metadata regulation[12] requires a fairly minimal set of discovery metadata elements based on ISO 19115. The UK implementation is being overseen by the cross-government Location Council, which has adopted an enhanced set of ISO19115 metadata: GEMINI2[13]. GEMINI2 has been developed to be consistent with INSPIRE, as well the earlier GEMINI1 and the UK e-Government Metadata standard (e-GMS)[14]. Thus we can think of ISO19115 as a superset of GEMINI2 itself as superset of the INSPIRE requirements. All DTC datasets will require GEMINI2 compliant discovery records.

There is no easily and publicly available schema or UML for GEMINI2, so we can only provide the list of 35 metadata codes (see table 2) which are defined in GEMINI2. It can be seen that these are very high level discovery information, and would not provide very little useful discrimination between DTC datasets, although they will make DTC data discoverable to a much wider community.

4.9 Relevant Standards: Annotation using Atom and GeoRSS

The Atom Syndication Format [15](RFC4287) provides a schema for organising "feeds" of information "entries". Originally developed for the "blogging" community, Atom is now finding wide use because not only does it organise the simple information associated with html content (authorship, keywords, links etc), it can have content which is binary (e.g. podcasts) or out of band XML (e.g. it's use by the EC Metafor project to provide links to complex XML descriptions of simulations[16]). Atom entries can also be decorated with geographic information by use of GeoRSS[17] attributes. Because the Atom syndication format is intended to be consumed by simple HTTP GET statements, and because of the simplicity of its design, it has been widely implemented. It is now replacing other technologies (such as the Open Archives Initiative Protocol for Metadata Harvesting[18]) as a harvesting tool for metadata.

Atom is widely used for annotating information held elsewhere using the atom:link attributes to identify the target. In conjunction with one or more of the Resource Description Framework (RDF), the trackback protocol[19], and OpenSearch[20], and efficient web searches, it can be used to provide a framework for annotating remote resources. In the context of the DTC it could be used as an information model to capture third party documentation of data quality, as well as a method for exposing metadata in any of the other formats. Because of it's ubiquity, Atom information can be created anywhere, and Google will exploit GeoRSS to match it to specific features. For these reasons Atom may have a role to play in the DTC project either directly as "C" metadata or to expose "C" metadata in other formats.

[12] EC regulation No.1205/2008,
http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32008R1205:EN:NOT

[13] http://location.defra.gov.uk/wp-content/uploads/2009/12/UKLC-Meeting-Summary-4-Dec-20091.pdf

[14] http://www.gigateway.org.uk/metadata/pdf/GEMINI2.pdf

[15] Atom Syndication Format: RFC4287, see http://www.ietf.org/rfc/rfc4287.txt

[16] See the Metafor service report at http://home.badc.rl.ac.uk/lawrence/static/2010/03/24/MetaforWP4Services.Report.pdf

[17] GeoRSS, see http://www.georss.org/Main_Page

[18] OAI/PMH: See http://www.openarchives.org/pmh/

[19] Trackback protocol specification: http://www.sixapart.com/pronet/docs/trackback_spec

[20] Opensearch: See http://www.opensearch.org/Home

A	*B*	*C*	*Element Name*	*Type*	*Relationsihp with ISO (Mainly 19115, some 19119)*
1	M	1	Title	String	MD_DataIdentificaiton.citation.>CI_Citation.title
2	O	N	Alternative Title	String	MD_DataIdentification.citatoin.>CI_Citation.alternateTitle
3	C	N	Dataset Language	String	MD_DataIdentification.language)
4	M	1	Abstract	String	MD_DataIdentificatoin.abstract
5	C	N	Topic Category	CodeList ^1	MD_DataIdentification.topicCategory
6	M	N	Keyword	String **	MD_Identification>MD_Keywords.keyword
7	M	1	Temporal Extent	date or two dates ISO8601	EX_Extent>EX_TemporalExtent..extent
8	M	1	Dataset Reference Date	date IS08601	MD_Identification.citation > CI_Citation.date
10	M	1	Lineage	String	DQ_DataQuality.lineage > LI_Lineage.statement
11	M	1	West Bounding	longitude	MD_DataIdentification.extent > EX_Extent >EX_GeographicExtent > EX_GeographicBoundingBox.
12	M	1	East Bounding	longitude	
13	M	1	North Bounding	latitude	
13	M	1	South Bounding	latitude	
15	O	N	Extent	Codelist^ 2	MD_DataIdentification.extent > EX_Extent >EX_GeographicExtent > EX_GeographicDescription.geographicIdentifier
16	O	1	Vertical Extent	Class^1 (min, max, coord. sys)	MD_DataIdentification.extent > EX_Extent > EX_VerticalExtent
17	M	1	Spatial Reference System	String **	MD_ReferenceSystem.referenceSystemIdentifier >RS_Identifier.code
18	M	1	Spatial Resolution	Real (m)	MD_DataIdentification.spatialResolution >MD_Resolution.distance
19	C	N	Resource Locator	URL	MD_Distribution >MD_DigitalTransferOptions.online > CI_OnlineResource.linkage
21	O	N	Data format	String **	MD_Distribution > MD_Format.name
23	M	N	Responsible Organization	Class^2	MD_Identification.pointOfContact
24	M	1	Frequency of Update	Codelist^3	MD_MaintenanceInformation..maintenanceAndUpdateFrequency
25	M	N	Limitations on public access	Codelist^4	MD_Identification>MD_Constraints>MD_LegalConstraints.accessConstraints
26	M	N	Use Constraints	String	MD_Identification >MD_Constraints.useLimiitation
27	O	1	Additional information source	String	MD_Identification >MD_Constraints.useLimiitation
30	M	1	Metadata date	Date ISO8601	MD_Metadata.dateStamp
33	C	1	Metadata Language	String**	MD_Metadata.language
35	M	N	Metadata point of contact	String ?	MD_Metadata.contact > CI_ResponsibleParty
36	M	1	URI	URI	MD_Identification.citation >CI_Citation.identifier
37	C	1	Spatial Data Service Type	Codelist^5	ISO19119 serviceType
38	C	N	Coupled Resource	URI/URL	ISO19119 operatesOn
39	M	1	Resource Type	Codellist^6	MD_Metadata.hierarchyLevel (extended)
40	C	1	Originating Controlled Vocab	String	MD_Identification >MD_Keywords.thesaurusName
41	C	1	Conformity (to INSPIRE product spec)	Boolean	DQ_DataQuality > DQ_Element.result >DQ_ConformanceResult.pass
42	C	1	Specification	String	DQ_DataQuality>DQ_Element.result>DQ_ConformanceResult.specification

*Table 2: GEMINI2 content: Column A is the numeric identifier for each element., B indicates whether Mandatory, Optional or Conditional, C indicates how many times an element may appear. Note the use of multiple code lists, and some strings marked with ** which ought to use controlled vocabularies. ISO standard equivalents are included. Elements specific to service metadata are shaded. Some string elements ought to be constrained*

4.10 A DTC profile of Observations and Measurements

As outlined above, O&M provides a suitable framework for the DTC data model, but at the same time, the DTC data model needs to embrace existing profiles of O&M both to exploit the work already done and to ensure interoperability into the wider community. To that end there is a tension between the DTC interoperating between the sub-communities within it, and with the extensions of those sub-communities outside. The situation is depicted in Figure 14: some of the properties to be described in the DTC data model will be describable in more than one of the community data models, some will be in one of them, and some will be in none, and need describing within a DTC specific profile.

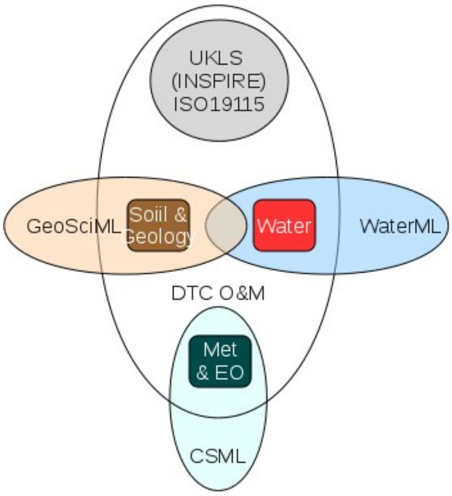

Figure 14: Relationship between DTC information requirements and those already encapsulated in relevant communities

To that end it is important to identify what are the key specializations that will be needed by the DTC, before investigating how to exploit the existing community models. Figure 15 shows the key areas where the DTC needs will extend the O&M model: in describing the methods used where there are a number of specializations required, and in the description of the data itself, where the DTC project includes a wide range of data types.

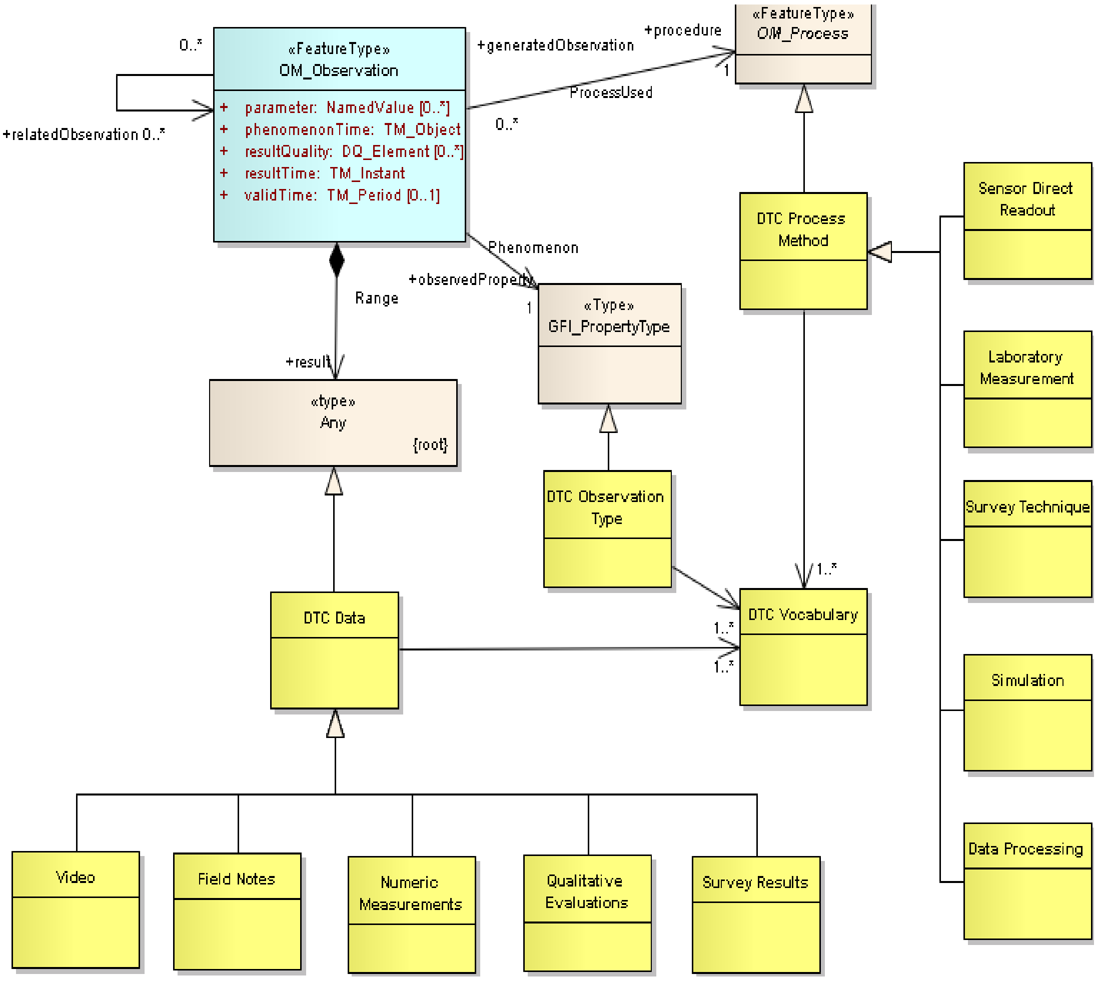

Figure 15: Characterization of some specializations that will be needed to accommodate the DTC data in an observations and measurements framework: a number of specializations associated with the various types of observation results, a number of specializations associated with process, and many vocabularies will be needed.

Some of these DTC required specializations could be supported within the O&M model itself, for example, specific types of single valued observations might utilize the O&M specializations shown in Figure 16, which constrain what the result type will be.

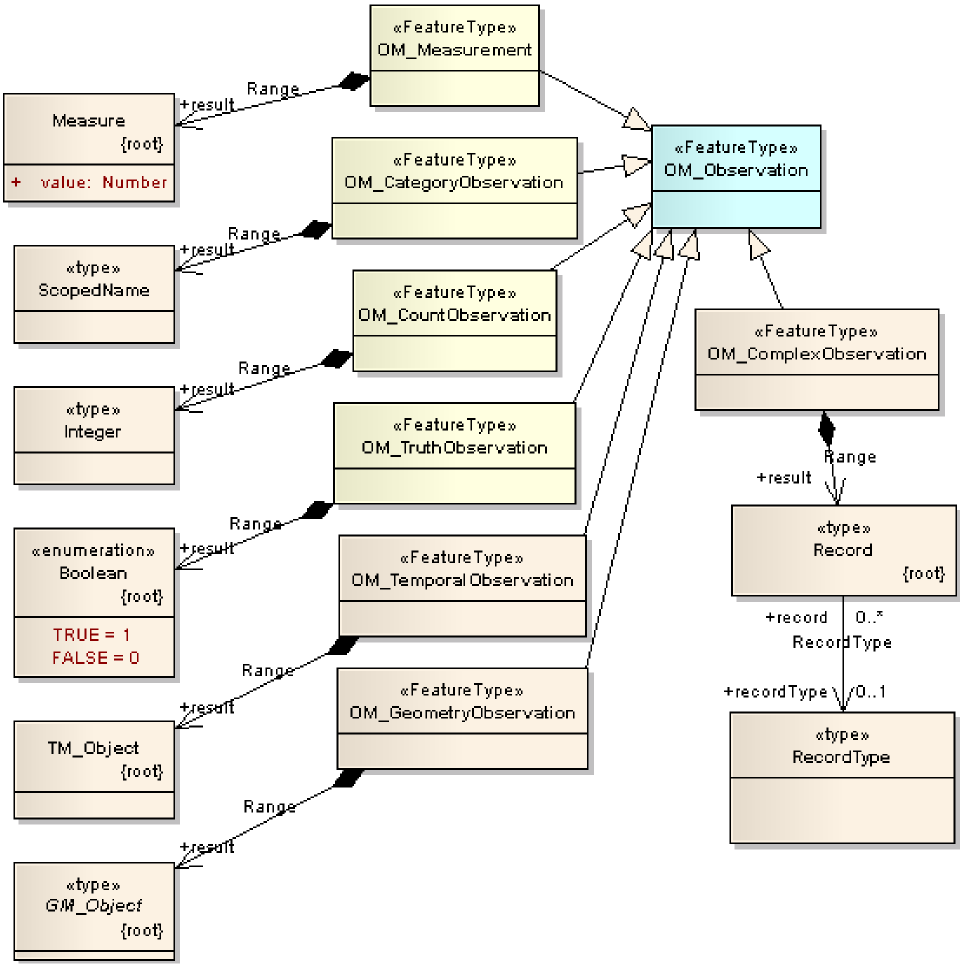

Figure 16: O&M specialization of observations into specific types.

However, many of the DTC data types will either require extension (for example, the construction of scoped name vocabularies – see section 4.11), or the use of specific sampling features to support coverage observations. The nature of such specializations is covered within O&M, in the section on sampling features, but specific implementation of those specializations is in the domain of profiles of O&M. Similarly, specialisations of coverages in sampling manifolds is covered in O&M (Figure 17), but in this case, the most complete profile of O&M addressing these features is CSML, which identifies twelve coverage type sampling features (in addition to one single-valued sampling feature), as shown in Table 3.

When considering extension, the key thing to keep in mind will be the necessity to extend by exploiting the modularity that exists in existing domain models, particularly GeoSciML, which already exposes a number of modular profiles which communities can mix and match. Thus, we would expect a DTC O&M profile to extend only in areas where other domain models cannot provide the structure needed.

It was not possible, nor practical, within this work, to define a complete DTC data model, nor indeed to propose a formal skeleton. This issue is discussed in detail elsewhere in this report.

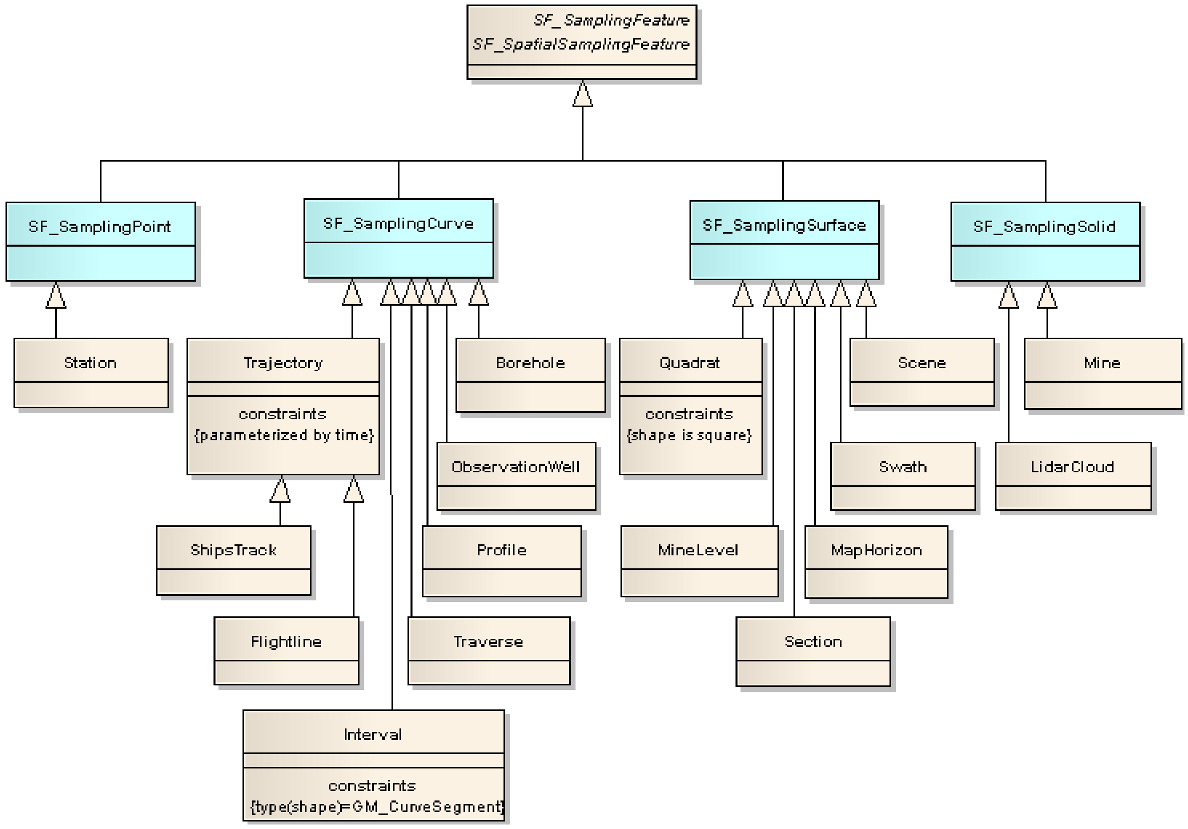

Figure 17: Specializations of sampling features discussed in Observations and Measurements.

Feature Type	Description	Example
PointFeature	Single point measurement.	raingauge measurement
PointSeriesFeature	Time-series of single datum measurements at a fixed location in space.	tidegauge, rainfall timeseries
TrajectoryFeature	Measurement along a discrete path in time and space.	surface salinity along a ship's cruise track; volcanic dust along an aircraft's flight path
PointCollectionFeature	Collection of distributed single datam measurements at a specific time.	2m temperatures measured at weather stations across the UK at 0600Z.
ProfileFeature	Single 'profile' of some parameter along a vertical line in space.	Wind sounding, radiosonde
ProfileSeriesFeature	Time-series of profiles on fixed vertical levels at a fixed location.	Vertical radar timeseries, thermistor chain timeseries
RaggedProfileSeriesFeature	Time-series of unequal-length profiles, but on fixed-vertical levels at a fixed location.	Repeat daily balloon soundings of atmospheric temperature from the same location
SectionFeature	Series of profiles from positions along a trajectory in time and space.	Acoustic Doppler Current Profiler
RaggedSectionFeature	Series of profiles of unequal length along a trajectory in time and space.	marine CTD measurements along a ship's cruise track
ScanningRadarFeature	Backscatter profiles along a look direction at fixed elevation but rotating in azimuth.	weather radar
GridFeature	Single-time snapshot of a gridded field.	gridded analysis field
GridSeriesFeature	Time-series of gridded parameter fields.	numerical weather prediction model, ocean general circulation model
SwathFeature	Two-dimensional grid of data along a satellite ground-path	AVHRR satellite imagery

Table 3: CSML (v2) feature types: note the clear correspondence with the sampling feature specializations in Figure 17.

4.11 Vocabularies

In order to share data, data producers and data consumers need to be sure that they are talking about the same thing and how these things relate to other things.

> For example scientist one describes a catchment as two-dimensional (area) whereas scientist two describes a catchment as three-dimensional (taking account of the underlying geology). Scientist one calls it 'CATCHMENT' in his database, scientist two calls it 'CTCHMNT'. They are both talking about essentially the same thing but have described it in slightly different ways. A human can spot that this has occurred, a computer cannot.

Hence, even with a well constructed data model it is crucial to construct and manage vocabularies of the terms used as property values. O&M refers to the use of such controlled vocabularies as ScopedNames – that is *terms*, selected from a specific scope.

The vocabularies also need to address procedures, and this will be one of the important areas of extension from O&M, where both procedure structure will be needed, as will vocabularies. The processes through which data are measured are also key to the understanding, reuse, and longevity of the data. Without these measurement metadata, encapsulated in vocabularies, the understanding of the data by third parties and interoperability with other datasets is limited, and in the long term the lack of information on provenance can render data useless. For the example of lab-derived water quality data, hydro-chemical analysis techniques are essential to full understanding of measurements. These can include sampling methods, sample storage methods, pre-filtration levels, sample preservation methods, and, for a given determinand, analysis process, pre-concentration, lab machines used, Standard Operating Procedures, and the subsequent limits of detection and quotation accuracies of the resulting data values.

It is important to establish a common understanding of these concepts, to extract in detail the variety of process stages which are used across the project, and to develop a data model that allows for the preservation of this data.

The following existing vocabularies have already been identified as important within the DTC project:

- Classification of Farm Types (Defra 2010[21])
- Classification of Farmer Attitudes (Defra, see Pike 2008[22])
- Standard chemical definitions (IUPAC[23])
- Climate & Forecasts (CF) Standard Names[24]
- Standard Operating Procedures (SOP)

Such vocabularies will need to be converted to standardised formats, and inevitably other vocabularies will be identified, and either constructed or imported from other governance domain as the project progresses. Such vocabulary requirements will include defining which units are to be used, for example, the use of molality, molarity, and moles/litre as units of concentration. Such definitions will also be required for

- How Nitrogen is measured (Nitrate, Total-N etc.,)
- How Phosphorus is measured (Phosphate, Total-P etc.,)

and these definitions will impact on the process descriptions required within the O&M model.

[21] Defra (2010): Definitions of Terms used in Farm Business Management, seehttp://www.defra.gov.uk/foodfarm/farmmanage/advice/documents/communisis-a4.pdf

[22] Pike (2008): Understanding behaviours in a farming context. Defra discussion paper, see http://www.defra.gov.uk/evidence/statistics/foodfarm/enviro/observatory/research/documents/ACEO%20Behaviours%20Discussion%20Paper%20%28new%20links%29.pdf

[23] E.g. see the IUPAC gold book at http://goldbook.iupac.org/index.html

[24] NetCDF Climate and Forecast (CF) Metadata Conventions: http://cfconventions.org

Vocabulary management is addressed in section 3.9, but the vocabularies too need to conform to data models of their own.

The most important data model for vocabularies, is the Simple Knowledge Organization System (SKOS[25]), which provides a structured mechanism for linking thesauri, taxonomies, classification schemes and subject lists. SKOS exploits the Resource Description Framework[26] to provide a machine-readable mechanism for establishing distributed vocabularies which also allows a level of a machine understanding of the relationships between terms in the various vocabulary entities.

4.12 Query Model

One of the key constraints on the data model is understanding the usage patterns of the data, and in particular, what queries are likely to be be commonly carried out by data and/or portal users. It would not be cost effective (or useful) to ensure that all data is indexed against a parameter which was only of use to one small constituency of users – of course that parameter ought to be obtained and stored, but not as a "first-order" entity in the data model.

Typically then, we can consider three levels of information:

1. Important named objects become "feature types" and we make sure the systems are built around storing and querying these objects.
2. Second-order "properties" of those features are those obtained and used as indexes along which one can navigate through the data efficiently, and
3. Third-order properties are those which we obtain and store, but do not support as important axes of navigation.

In terms of the information model, the difference between second and third-order properties may appear unimportant, but they become important when building information systems to collect these data, and so it is useful to try and establish the importance of properties during the information modelling phase.

However, the importance for querying only becomes apparent during the prototype evaluation phase, and so no data model can be complete until this phase has been completed. For example, one might need to extract measurements based on *sampling frequency* as a search parameter. This might be considered as an inherent characteristic of the data by some data providers, and not of great interest, but data consumers might consider this an important axis of interrogation.

In this context, the user workshop has already identified that users typically want to query the data by the following fields:

- Time
- Determinand
- Geo-spatial location
- Determinand threshold

However, it is likely that more query axes will be established during the course of the project.

[25] http://www.w3.org/2004/02/skos/

[26] Resource Description Framework (RDF), see http://www.w3.org/RDF/

4.13 Next steps in information modelling

The material presented in the enterprise viewpoint has outlined the information requirements of the project as they have been understood based on the input from one workshop as well as the domain expertise of the report authors. The material presented in the information viewpoint has outlined the key information entities which need to exist and the relevant standards landscape. The material to be presented in the computational viewpoint will identify the major computational entities which are expected to exist, along with the interfaces required (and the relevant standards for those interfaces).

However, this document will not provide a complete specification of the information entities or the computational entities. It is not practicable to construct a proper domain model without considerable further input from domain experts, nor without considerable experimentation with what is available within the existing schema.

Similarly, the data model will need to be further developed to ensure that usage requirements can be fulfilled, and these are only likely to be fully understood once real users are engaged with real prototype services.

Thus, Defra should expect that the eventual contract holders are likely to have to go through a phase of building and testing their eventual data models, even as the DTC project begins, and so interim plans for handling data and information will be necessary. A formal first draft of the data model will then be finalised by confronting the query model with the data model and user expectation, and the final versions will only be possible once realistic prototypes have been built and tested against the data.

What this means in practice for the data providers is that they need to concentrate early on documenting the key aspects of their data and metadata using whatever existing tools they have to hand, and share as early as possible, their expectations as to what information they require of each other.

The process should follow these principles when establishing structure:

1. Exploit what is already available,
2. Don't wait for standards to be finalised
3. Discuss repeatedly with the community of direct interest their requirements, but
4. Don't try and get total agreement from everyoneone, and then
5. Descope, and Despecialise.

Recognising the tension between the DTC project itself, and the larger communities which intersect within it, expect to use a "mapper/wrapper" approach to export and import to different communities, rather than a "compose it all within" approach.

The vocabularies used to populate the structure should be as far as possible decoupled from the structure itself so that the project can

1. Exploit existing vocabularies,
2. Build new ones and provide governance for them, and
3. Make them easy to use!

Expect this process to continue throughout the project, with mid-course corrections from internal pressures, and external drivers.

4.14 Summary of Information Viewpoint Recommendations for the DTC

Note that these are recommendations for the semantic structures, we address the encoding of this information in the computational viewpoint (see Section 5).

	Information type	**Recommendations**	**Discussed in**
A	Archive: should describe the encoding of the data itself using an appropriate format and convention. Is likely to vary between the various input streams.	Various	Computational viewpoint
B	Browse: should describe the information needed to choose between the various data objects available, provide provenance, and context.	Exploit a combination of WaterML, CSML and further DTC specializations of O&M. Ensure that export to GeoSciML can be supported. Consider relationship with MOLES.	Information viewpoint.
C	Character and Citation: should provide third party annotations as to the utility and reliability of the data.	Use annotations using Atom (RFC3287), with geotagging.	Computational viewpoint.
D	Discovery: should conform to legislative discovery requirements.	ISO19115 using the UK government Gemini2 profile.	Information viewpoint.
E	Extra: should provide extra information needed by domain experts.	Other documents which conform to well specified metadata models (e.g. SensorML) and documents which are in formats which can be persisted long term (.pdf – but not .doc)	Computational viewpoint
O	Ontologies and Vocabularies	Vocabularies should be created using standard thesauri tools and using SKOS and SKOS extensions to manage relationships (section 4.11)	Information viewpoint and computational viewpoint.
Q	The Query Model	This will need further elaboration early in the formal archiving project.	Information viewpoint.

Table 4: Recommendations for information structures for the DTC, classified according to the metadata taxonomy introduced in section 4.2

In interpreting this table it is important to recognise the important role of A file types and E type metadata. It is not practical nor possible for all aspects of data and metadata in a project this diverse to be encoded in the same standard information paradigms. The DTC project will need to use file type metadata along with extra metadata to archive the complete information resource. However, the aim of the DTC metadata model should be to ensure that all extra metadata is at least catalogued within the main data model – and that formats associated with extra metadata are both well documented and capable of long term persistence.

5 COMPUTATIONAL VIEWPOINT

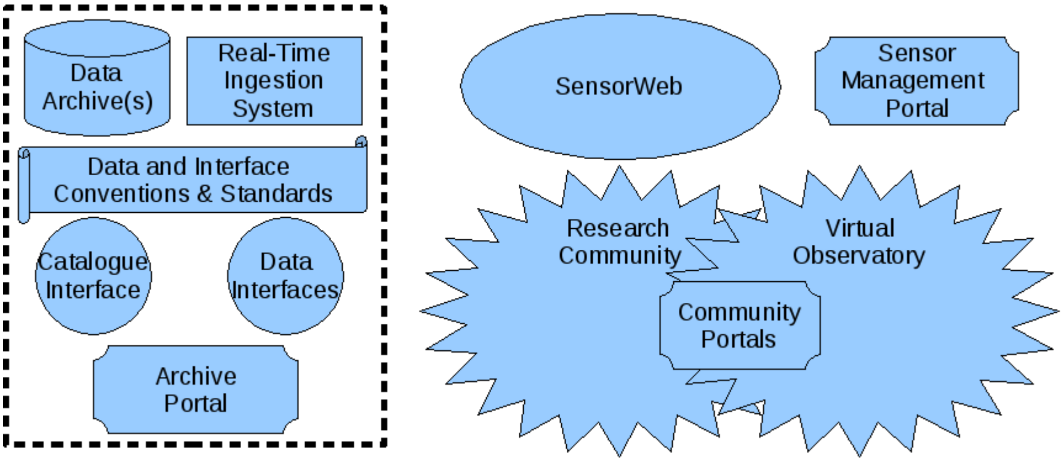

Figure 18: Schematic of the archive entities in the DTC context: the archive will consist of a portal, underlying interfaces (which conform to conventions and standards), and actual data which is ingested both in real time and as data becomes appropriately mature in the normal research cycle). We expect community portals as well which will harness their own internal information as well as interact with the archive contents through the service interfaces.

Figure 18 shows the major functional components which the DTC archive needs to deliver:

1. The physical archive itself, consisting of information entities (files and/or databases),
2. Service interfaces to the archive, and
3. A web portal to the service interfaces and underlying data.

It is expected that other activities may harness the Web Service interfaces as well.

Because there will be data for which IPR is an important issue, and because it is likely that some data may have periods of embargoed access, some form of access control interface will also be required.

The details of the software system used to construct theses entities will be up to those who deliver the archive, but to support the inter-working of the wider community with the archive, it will be important to decouple the portal from the archive services, allowing community portals to remotely interact with the archive, as depicted in Figure 19. The archive itself consists of the data with service layers plus a portal.

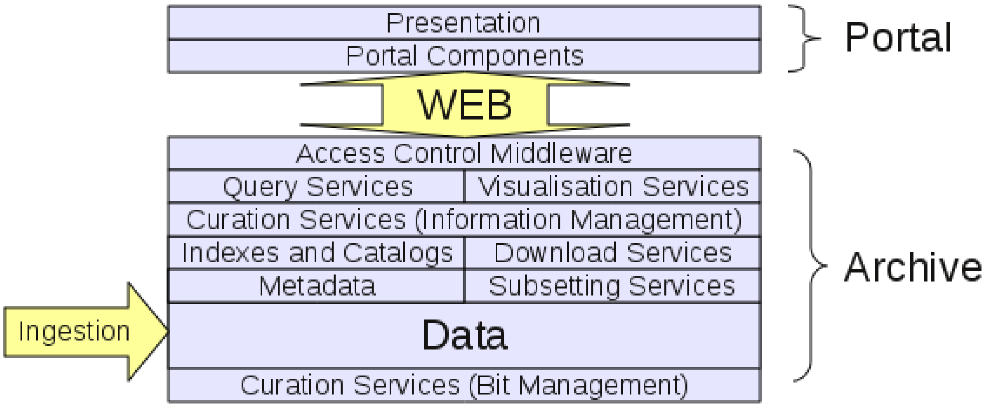

Figure 19: Layered approach to delivering archive and portal components

5.1 The Archive Itself

The archive itself should be constructed according to best practice standards, which at the moment consist of conforming to the Open Archive Information Standard (OAIS) for digital archives. OAIS mandates the existence of certain classes of information packages (which are described in the information viewpoint), but also certain management behaviours, the most important of which for our purposes is preservation planning.

Hence, the archive should also be constructed in such a way that:

- The archive contents could be easily moved to another provider should the service be deemed insufficient by Defra, and
- The archive contents are available to be consumed via interfaces in other community portals (as indicated in Figure 19, and
- The data can be persisted for decades, and
- The data as ingested is available without any transformations that lose information. (In practice, because no archive provider can guarantee lossless transformation, particularly through multiple transformations, original format data must be persisted alongside versions optimised for retrieval by the target communities).

To that end, we recommend that underlying data be stored as flat human readable files where data volumes are small, and as NetCDF binary files where data volumes are large, as both formats will be readable for the foreseeable future. Some geometries may also be stored in the ESRI Shape file format. Only one video and audio format should be accepted (to be decided). Human readable files should be constructed using a standard format to ensure adequate internal metadata: exactly which formats are supported is something to be established early in the project, but we would anticipate:

1. XML files which conform to one or more of the DTC supported application schema (WaterML, CSML etc),
2. ASCII files which conform to an appropriate metadata enhanced spreadsheet derived format (we recommend the BADC csv format[27], including appropriate metadata, but the eventual provider in consultation with the consortium should agree the appropriate format).

We strongly recommend that, regardless of the eventual solution provider, binary files using the NetCDF format should be constructed following the CF-conventions. If CF is found to be inadequate, then extensions should be sought through the CF process.

Storing data in these formats *does not preclude the use of databases* to provide indexing of the metadata (and for low volume data) the data itself. However, we strongly recommend against the only storage for the data (particularly the primary data as ingested) being a database (especially a commercial one) since experience suggests even where groups try and avoid vendor lock-in, some key information ends up being encoded in information structures which are difficult to preserve long term in a multi-vendor environment. To that end, open source tools provide the best long term prospects for preservation of indexing information as well. Note that these recommendations conform to the government policy on open source software[28], and in particular, the consideration of exit costs, which are especially important for preservation of data.

These recommendations, based on long term-preservation requirements alone, match well to the requirements identified from the community:

The workshop identified that the community was currently using a range of comma and tab delimited “spreadsheet class” files (as well as native vendor formats such as Excel). Clearly standardizing onto a common standardized vendor-independent format for these is an important requirement – hence the

[27] BADC CSV Format for Data Exchange: http://badc.nerc.ac.uk/help/formats/badc-csv

[28] Government policy on open source: http://www.cabinetoffice.gov.uk/media/318020/open_source.pdf

recommendation to use BADC CSV (or any suitable equivalent)

Since ESRI shape files are in common use, and the specification for such files is in the public domain, and as far as we know is not encumbered by IPR, we would recommend accepting Shape files into the archive as well. Similarly, appropriate video and audio formats should also be accepted. Here, suitable should be formats which are unencumbered by IPR issues (however, given the gamut of proprietary solutions, it may be necessary to support input/output filters into commonly used formats).

The community is using a wide variety of grid formats, most of which appear to be home-grown. However, other communities are beginning to converge on the use of NetCDF as a well documented, efficient binary format, for which tools are available on most platforms, as well as in most programming languages. We strongly recommend the usage of NetCDF within the DTC project for high volume binary data.

While GML[29] based formats are not yet common in this community, the advent of GeoSciML and WaterML is changing behaviour, and we would expect them to become more common. Because they come with appropriate structures for describing the entities important to the DTC information requirements, data in those formats should be supported. However it is likely that they will need augmentation with properties described in a DTC specialization of O&M as described in the information viewpoint.

KML[30] is starting to be in use in the community, and while this is to be encouraged, we see KML as an interface language, not an archive language, and we recommend that data is not imported into the archive in KML. We do, however, support the export of data in KML format.

Further ancillary documents will be required in the archive, but these need to be in formats which have longevity in mind: to that end, the archive should not permit the storage of custom word processor formats (such as .doc and .docx), but admit document with open standards and open implementations (.pdf).

5.2 Required Data Interfaces

There are a number of key interfaces that the archive needs to support:

1. Data Ingestion Interfaces – to support getting data into the archive.
2. Data Querying Interfaces – so that data and metadata can be found
3. Data Download Interfaces (including subsetting) – to support the retrieval of appropriate data and metadata.
4. Data Visualization Interfaces – to deliver commonly required ways of graphing the data..
5. Data Processing Interfaces – to process data from its stored form into alternative forms, where either common processing options can be provided or data volumes are large enough that server-side processing is preferred to data downloading. It is not obvious that the DTC project is going to involve much, if any, of the sort of data which falls into the second of these two categories.
6. Data Catalogue Services – to support the query services, and to ensure the data is visible in other catalogues, including those of the UK government, the EU and the Global Earth Observing System of Systems (GEOSS).

5.3 Appropriate Interface Standards

There are two classes of interface standards of interest: those which constrain the semantics of the interface, and those which constrain the methodology of passing information through the interface. There are three of the latter which are of interest here:

[29] GMLis Geography Markup Language: http://www.opengis.net/gml

[30] Keyhole Markup Language (KML): http://www.opengeospatial.org/standards/kml

1. SOAP based interfaces: which exploit the Simple Object Access Protocol using a remote procedure call like methodology to encapsulate objects as SOAP XML payloads,
2. So-called POX interfaces, where "plain old XML" documents are passed through interfaces, and
3. RESTful interfaces, which conform to the Representation State Transfer principles for distributed hypermedia to allow the use of standard HTTP methods to manipulate objects which carry well defined identifiers accessible to the Web.

It will be seen that interfaces which conform to semantic constraints may exploit one or more of the above as alternatives interfaces. Where alternatives are available we strongly recommend the use of RESTful interfaces as it is our belief that these are easier to engineer, easier to consume, and are likely to have the most longevity. However, from an information interoperability point of view, the most important constraints are those around the nature of the semantics of the interfaces. To that end, the following OGC standards provide the best available semantic standards for interfaces, and we therefore recommend that the DTC utilise these interfaces where possible:

WCS – Web Coverage Service: gridded data download

WMS – Web Map Service: data viewing/maps

WFS – Web Feature Service: data querying and filtered access

WPS – Web Processing Service: remote processing

SOS – Sensor Observation Service: querying one or more sensor systems

KML – Keyhole Markup Language (Google Earth and Google Maps)

Details of these standards can be found on the OGC standards website[31], some of which are also described briefly below.

5.3.1 Web Coverage Service

The Web Coverage Service defines a web service interface for subsetting multi-dimensional spatio-temporal gridded datasets in binary format. The standard has had a chequered history, with major changes between versions 1.0 (OGC document 05-076), 1.1.0 (OGC document 06-083r8) and 1.1.2 (OGC document 07-067r5). The OGC WCS Standards Working Group recently abandoned an evolutionary effort to develop version 1.2, instead opting for a major overhaul to version 2.0, currently (April/May 2010) undergoing vote. The main feature of the new version is a GML-based representation of a coverage, and a factoring of the specification into 'core' plus 'extensions', providing different conformance points for implementations depending on their required capabilities. The operations supported by WCS include:

- GetCapabilities (mandatory): general service metadata, plus a description (name, label, keywords, spatiotemporal bounding box) of offered coverages
- DescribeCoverage (mandatory): more detailed metadata (detailed spatiotemporal domain locations, description of 'range' values, supported coordinate reference systems, supported formats, supported interpolations) for a specific coverage
- GetCoverage (mandatory): returns a coverage in a specified reference system and format based on a spatiotemporal subset bounding box, and multidimensional (pixel) resolution (if interpolation is supported)

5.3.2 Web Map Service

The Web Map Service version 1.3.0 (OGC document 06-042) has been standardised as ISO 19128, and provides a web service interface for generating rendered geospatial 'map' layers. While there is no mandated relationship between WMS layers and an underpinning data model, the approach adopted by

[31] OGC Standards Website: http://www.opengeospatial.org/standards

INSPIRE (and often used more generally) is by default to define a layer for each 'feature type' – e.g. a hydrological model may have separate map layers for water bodies, catchments, man-made objects, reporting units, etc. A map may be queried for additional information on a specific rendered feature. The operations supported by WMS include:

- GetCapabilities (mandatory): describes the list of layers supported by the WMS, providing metadata (title, name, abstract, keywords, available rendering 'styles', available CRS/projections, bounding box, whether the layer is queryable, opaqueness for overlaying, etc.) for each layer
- GetMap (mandatory): a map is returned to the user in a specified image format and pixel size for a particular layer within a requested region (and time/elevation if required), and using a specified (or default) styling
- GetFeatureInfo (optional): a request effectively identical to GetMap is provided, with additional 'pixel coordinates' and information is returned on the feature located at that position; the content and format of the returned information is left to the discretion of the WMS provider (by default the INSPIRE WMS View Service will return a GML instance of the relevant feature)

5.3.3 Web Feature Service

The Web Feature Service 1.1.0 (OGC document 04-094) is currently undergoing standardisation as ISO 19142. It must be understood alongside the sister specification on Filter Encoding (OGC 04-095, ISO 19143). WFS defines a web service interface for querying an 'opaque feature store', regarded as a database of feature instances (a feature is an 'abstraction of a real world phenomenon' and may represent any specific class of objects defined in a data model). The data is typically returned encoded in GML. While a number of early implementations of WFS did a more-or-less direct 'translation' of existing relational tables to a flat GML structure, more recently the importance has been recognised of the WFS also mapping from an underlying database structure to a 'GML Application Schema' defined *a priori* (i.e. output feature types independent of relational table structure). It has been demonstrated[32] that the WFS interface may also be used for providing access to gridded binary data traditionally served via WCS. The operations defined in WFS are:

- GetCapabilities (mandatory): provides service metadata (e.g. which optional operations are supported), and lists the feature types and other GML objects (e.g. dictionaries, CRS definitions, etc.) available from the service
- DescribeFeatureType (mandatory): returns the GML application schema corresponding to a particular named feature type
- GetFeature (mandatory): retrieves a set of feature instances matching a defined selection clause (encoded according to the Filter Encoding specification) and a projection clause (i.e. requested feature types); xlinks to remote resources may be resolved
- GetGmlObject (optional): allows retrieval of specific features and other GML objects by identifier
- LockFeature (optional): allows locking feature instances to ensure consistency over the duration of a transaction (e.g. for update using the Transaction operation)
- Transaction (optional): an operation to insert, update, or delete feature instances from a server

5.3.4 Web Processing Service

The Web Processing Service version 1.0.0 (OGC document 05-007r7) enables computational processes to be encapsulated and exposed through an OGC web service interface. The specification is currently undergoing revision to version 2.0. The types of processes exposed generally require some kind of input data and generate output. Defined operations include:

[32] Lowe, D. and A. Woolf (2008): What's this "coverage vs. feature" nonsense? OGC Technical Committee Meeting, Valencia, Spain, 01-05 Dec 2008 [http://epubs.cclrc.ac.uk/work-details?w=49662]

- GetCapabilities (mandatory): returns service metadata, including names and descriptions of each process offered by the server
- DescribeProcess (mandatory): describes a specific process in more detail, including required inputs and formats, available outputs, and whether an asynchronous mode is available (e.g. storing results and reporting process status via polling)
- Execute (mandatory): runs a specific process on the server, with defined input data sources and parameters, and returning the results to the user

5.3.5 Sensor Observation Service

The Sensor Observation Service (OGC document 06-009r6) is the workhorse of sensor webs; it defines a general API query mechanism that may be run against networks of sensors to retrieve sensor observations filtered in specific ways. Sensor may be static, in-situ, mobile, remote, and may be aggregated into collections for efficient interaction. Sensor collections are grouped into logical collections for access via SOS, with each 'observation offering' characterised by:

- a specific 'procedure' (one or more instruments)
- time periods for which observations are available
- observed properties (e.g. temperature, rainfall, etc.)
- geographical region that contains the sensors
- sensor target locations (also called 'features of interest')

These logical groupings should be configured in order to define a coherent collection – e.g. two weather stations far apart in space or operated during different historical periods should probably be factored as distinct observation offerings. On the other hand, two sensors (wind and temperature, say) on a single weather station may sensibly be grouped within a single offering – data for both sensors will usually be available at the same time. 'Sparseness' of data within an observation offering should be avoided as far as possible.

The '52° North' consortium (http://52north.org/maven/project-sites/swe/) has developed a well-known and widely used open source implementation of SOS.

The operations defined by SOS include:

- GetCapabilities (mandatory): returns service metadata, including observation offerings (time periods, procedures, observed properties, features of interest, etc.), and observation filtering capabilities (e.g. in time and space)
- DescribeSensor (mandatory): a particular procedure (instrument(s) or sensor system(s)) is described in more detail using SensorML or TransducerML
- GetObservation (mandatory): in response to a request, returns matching observations; the request may filter on any dimension of the requested observation offering (observed property, space/time, procedure, feature of interest, etc.) though only the target offering and observed property are mandatory request parameters
- RegisterSensor (optional): a new sensor system may be registered under an 'SOS transactional profile'
- InsertObservation (optional): new observations may be inserted from a sensor system
- GetObservationById (optional): returns a specific observation by identifier
- GetResult (optional): a previous GetObservation request may have been made defining a 'result template', that may be re-queried with the GetResult operation without submitting all filtering parameters

- GetFeatureOfInterest (optional): returns full detail for an observed feature of interest advertised in the service metadata
- GetFeatureOfInterestTime (optional): returns time periods of available observations for a specific feature of interest
- DescribeFeatureType (optional): returns the GML application schema of a defined feature of interest
- DescribeObservationType (optional): returns the specialised observation type for a specific observed property
- DescribeResultModel (optional): returns the schema for a specified result type

5.3.6 Keyhole Markup Language

KML was developed by Google as an XML language for describing rendered scenes in Google Earth. It was subsequently handed over to OGC for ongoing development (Google recognised it should not compete in the business of developing geospatial technology standards). The objectives of the OGC standardisation effort are:

- to ensure there is a single international standard for geographic annotation/visualisation on web-based online and mobile map browsers and virtual earth browsers
- that KML be aligned with international standards to ensure interoperability
- that OGC and Google work collaboratively to engage and inform the KML developer community
- that the OGC process be used to ensure proper life-cycle management for the KML standard

5.4 Data Discovery Interface Conventions and Standards

Data discovery is a rapidly moving landscape: government in the UK is recommending the use of linkeddata principles, while INSPIRE is recommending the use of the OGC/Catalog Web Service (known as the Catalog Servcie for the Web, CSW). Rather than the DTC project trying to follow discovery service evolution, the DTC should simply allow it's data to be harvested in GEMINI2 format into as many other services as possible, and allow them to worry about the changing landscape.

One option would be to exploit the NERC data discovery service (DDS). Currently that would require depositing the GEMINI2 records into an OAI/PMH (Open Archive Initiative Protocol for Metadata Harvesting) compliant digital repository, but is likely that the simpler option of pointing to the records from an Atom feed would provide more flexibility in the near future. The NERC DDS would harvest this metadata, and make it available in the NERC portal, and forward it into the national and international repositories.

5.5 Vocabulary Service Interfaces

Currently there is no standard vocabulary service interfaces. The best available interface with which the authors are familiar is the NERC vocabulary server interface[33], which provides both a SOAP[34] and raw XML over HTTP interface, with new REST[35] interfaces under development. RESTful interfaces to SKOS are also under development elsewhere, but there are subtle mismatches between SKOS and REST[36]. The NERC vocabulary group are likely to track standards developments, so we recommend the DTC follows their roadmap and interfaces as they evolve.

[33] See http://www.bodc.ac.uk/products/web_services/vocab/

[34] SOAP: Simple Object Access Protocol, see http://www.w3.org/TR/soap/

[35] See http://en.wikipedia.org/wiki/Representational_State_Transfer for an overview of RESTful web services and principles.

[36] Simon Cox, CSIRO Australia, personal communication.

5.6 User Authentication

There are a plethora of Authentication (and Authorisation) standards, and rather than review them here, we simply recommend that all resources exposed by interfaces to the archive are associated with an access control policy implemented via middleware services with respect the following:

1. OpenID is used for authentication as it is relatively easy to implement, and anyone looking to deliver the archive is likely to be able to wrap services using OpenID without too much effort (It will likely be important to minimise the effort expended on access control as a proportion of the funded activity.) If other authentication protocols become prevalent, gatewaying to and from them via OpenID should be possible.
2. Authorisation is delivered via similar protocols to those now being deployed for the Earth System Grid.

However, because access control tooling is currently immature enough that there are no clear widely deployed standards, should other protocols be recommended by the eventual archive providers, along with evidence of at least some widely deployed communities using them, the only other criteria by which they should be judged is what proportion of the project cost would need to be spent on delivering access control.

5.7 DTC Interface Standards

The interface requirements were outlined at the beginning of section 5, and the available interface standards described in section 5.3. In this section we make our recommendations for how these interface standards should be used in the DTC.

However, before doing so, it is useful to understand the relationship between the Sensor Observation Service and other OGC services (see figure 20). The key point to note is that while SOS is the natural fit for manipulating and interrogating sensors and the WFS and WCS protocols are probably natural fits for wider interoperability, they do play together quite naturally, and can be combined sensibly in a DTC portal.

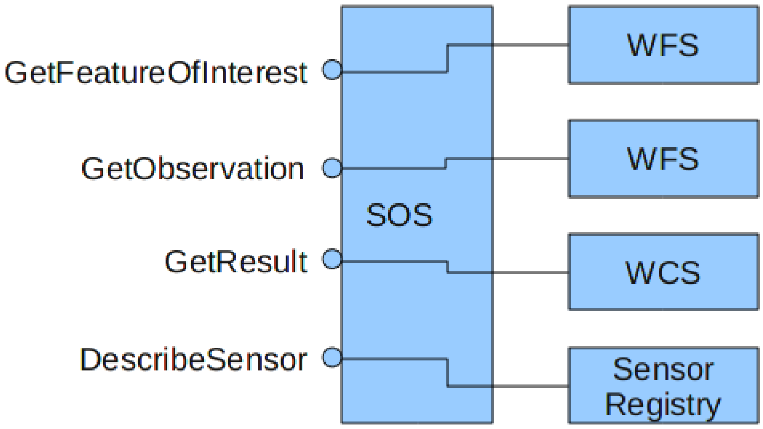

Figure 20: Schematic of the relationship between the SOS and three key underlying services: WFS, WCS, WFS and a registry of sensor descriptions. The "Get Feature of Interest" interface describes what has been observed via a WFS to underlying data, the "Get Observation" interface describes the observation method itself (including possibly a description of the sensor itself), and the "Get Result" method returns an actual data observation (via a WCS). (Modified from a presentation by the O&M author, Simon Cox.

Table 4 outlines the recommendations for the DTC interfaces.

Interface	Notes	Protocols and Standards
Data Ingestion	Real Time	SOS
	Scientific Products	Manual upload
Download	Service	WCS + SOS +WFS
	Portal	WGET Scripts + HTML
Visualisation	Simple Maps	WMS + KML
	Plots etc	Customised WPS with customised service interface
Querying	Within and between Features	WFS
	Customised Portal Functions	Expose customised interface to xquery, sparql and sql as appropriate.
Catalogue	To support INSPIRE etc	CSW + NERC bespoke
Vocabulary Services	To Edit, Query and Download	NDG Vocabulary Service Interfaces.
Authentication	On archive and in portals	OpenID

Table 5: Recommendations for the DTC interfaces (from requirements to standard). To avoid confusion recall that this table applies to the tools that a portal should interact with that provide higher level functionality, it does not define portal functionality per se.

5.8 Data Portals

The services listed above should be delivered by a dedicated DTC project portal. The portal should provide access to the metadata, data, other products as well as relevant information about the project itself. It should also deliver customised report pages that are defined by the project requirements.

During the workshop it became apparent that there were a wide range of requirements of portals, and in fact that many of those requirements could only be delivered by project specific portals. To that end, the architecture outlined above would support customised portal development within the DTC research consortia.

The requirements of a "vanilla DTC archive portal" could be established early in an archive project, but should not be onerous, provided the main user interaction functionality was developed within the research consortia.

Defra will need to decide the scope of portal development required as part of the data management activities.

We recommend that the leader of the data management activities should hold an early consultation process with the main stakeholders to define the scope of portal services required.

5.9 Summary of Computational Viewpoint Recommendations

The following recommendations are made in the computational viewpoint section:

1. The archive itself should be constructed according to best practice standards, which at the moment consist of conforming to the Open Archive Information Standard (OAIS) for digital archives.
2. Underlying data should be stored as flat human readable files where data volumes are small, and

as NetCDF binary files where data volumes are large, as both formats will be readable for the foreseeable future.

3. Binary files using the NetCDF format should be constructed following the CF-conventions
4. Human readable files should be constructed using a standard format to ensure adequate internal metadata.
5. BADC CSV, or a suitable equivalent, should be provided as the human-readable file format.
6. If a database solution is employed there should also be a version of the data stored in a file format.
7. The project should use the various standards and interfaces outlines in table 4 of section 4 of the document.
8. OpenID is used for authentication. If other authentication protocols become prevalent, gatewaying to and from them via OpenID should be possible.
9. Authorisation is delivered via similar protocols to those now being deployed for the Earth System Grid.
10. The leader of the data management activities should hold an early consultation process with the main stakeholders to define the scope of portal services required.

6 SUMMARY OF RECOMMENDATIONS

This section draws together all the recommendations made in this document.

6.1 Summary of Enterprise Viewpoint Recommendations

This section provides a list of enterprise level recommendations for the DTC archiving activity.

1. When issuing an eventual specification for data archival, Defra should make the scope of intended users clear, so as to both manage expectation and maximize benefit for the desired stakeholders.
2. Any provider should include in their bid details of their exit plan for the data as to what would happen when (not if) Defra withdraw funding for the DTC archive.
3. Defra should include all points listed in Table 1 (section 3.2) in the eventual specification of the requirements of the data archive.
4. The DTC project should identify the most important third party datasets and then start negotiation for access as soon as possible. It may well be that the target datasets already conform to international standards, have managed vocabularies and fit within agreed data models. However this is unlikely and it is recommended that the programme sets aside some contingency funds to help with establishing access to specific datasets.
5. Both raw data and quality assured data should be preserved in perpetuity in the Data Archive.
6. The DTC data models are developed conform to both INSPIRE specifications and data.gov.uk requirements.
7. Defra requires project participants to engage with data model specification efforts throughout the project.
8. Open source software is used wherever dependency on data model extensibility is expected.
9. That a data policy document be drawn up based on the points listed above, and agreed by the DTC Project Board (and thus all DTC project participants) which should include clear guidelines as to when embargo periods are suitable, and for how long.
10. That the archive uses a registration system to control access to data, and potentially to some (but perhaps not all) data visualisation functionality.
11. That the DTC project construct a suitable data license, and ensure that all data users sign up to the terms and conditions of the license before access to the data is provided. (This last implies that all data portals implement some method of ensuring that a license agreement has been entered into).
12. The National Grid Reference datum should be used where possible for geospatial coordinates (possibly in addition to other coordinates where appropriate).
13. The DTC project should appoint a Programme Data Manager who will be responsible for coordinating data management activities. This person should be responsible for liaising with those responsible for local data archives (should they exist), liaising with the central Data Archive, establishing, implementing and monitoring programme data policy and establishing mechanisms to support scientists in getting the most out of their data.
14. Each Consortium involved in the DTC should appoint an individual Data Officer who has responsibility over how that Consortium interacts with the wider data management activities within DTC.
15. Defra should negotiate with the Centre for Ecology and Hydrology to ensure that any specific DTC vocabularies are incorporated in the vocabulary management activities at CEH.

6.2 Summary of Information Viewpoint Recommendations

Note that these are recommendations for the semantic structures, we address the encoding of this information in the computational viewpoint (see Section 5).

	Information type	**Recommendations**	**Discussed in**
A	Archive: should describe the encoding of the data itself using an appropriate format and convention. Is likely to vary between the various input streams.	Various	Computational viewpoint
B	Browse: should describe the information needed to choose between the various data objects available, provide provenance, and context.	Exploit a combination of WaterML, CSML and further DTC specializations of O&M. Ensure that export to GeoSciML can be supported. Consider relationship with MOLES.	Information viewpoint.
C	Character and Citation: should provide third party annotations as to the utility and reliability of the data.	Use annotations using Atom (RFC3287), with geotagging.	Computational viewpoint.
D	Discovery: should conform to legislative discovery requirements.	ISO19115 using the UK government Gemini profile.	Information viewpoint.
E	Extra: should provide extra information needed by domain experts.	Other documents which conform to well specified metadata models (e.g. SensorML) and documents which are in formats which can be persisted long term (.pdf – but not .doc)	Computational viewpoint
O	Ontologies and Vocabularies	Vocabularies should be created using standard thesauri tools and using SKOS and SKOS extensions to manage relationships (section 4.11)	Information viewpoint and computational viewpoint.
Q	The Query Model	This will need further elaboration early in the formal archiving project.	Information viewpoint.

Table 6: Recommendations for information structures for the DTC, classified according to the metadata taxonomy introduced in section 4.2

In interpreting this table it is important to recognise the important role of A file types and E type metadata. It is not practical nor possible for all aspects of data and metadata in a project this diverse to be encoded in the same standard information paradigms. The DTC project will need to use file type metadata along with extra metadata to archive the complete information resource. However, the aim of the DTC metadata model should be to ensure that all extra metadata is at least catalogued within the main data model – and that formats associated with extra metadata are both well documented and capable of long term persistence.

6.3 Summary of Computational Viewpoint Recommendations

The following recommendations are made in the computational viewpoint section:

1. The archive itself should be constructed according to best practice standards, which at the moment consist of conforming to the Open Archive Information Standard (OAIS) for digital archives.
2. Underlying data should be stored as flat human readable files where data volumes are small, and as NetCDF binary files where data volumes are large, as both formats will be readable for the foreseeable future.
3. Binary files using the NetCDF format should be constructed following the CF-conventions
4. Human readable files should be constructed using a standard format to ensure adequate internal metadata.

5. BADC CSV, or a suitable equivalent, should be provided as the human-readable file format.
6. If a database solution is employed there should also be a version of the data stored in a file format.
7. The project should use the various standards and interfaces outlines in table 4 of section 4 of the document.
8. OpenID is used for authentication. If other authentication protocols become prevalent, gatewaying to and from them via OpenID should be possible.
9. Authorisation is delivered via similar protocols to those now being deployed for the Earth System Grid.
10. The leader of the data management activities should hold an early consultation process with the main stakeholders to define the scope of portal services required.

Appendix A: Workshop Notes

In this appendix we tabulate as bullet points some of the key requirements established in the workshop which are not detailed elsewhere.

There has been little attempt to rationalise this list. See the main document for rationalised lists.

1. Input data streams
    - From all sensors on the trailer as it will be purchased (tbd).
        - Common spec, but flexibility around it.
    - Field experiments also deployed separately
    - Some edge-of-field monitoring
    - Third party data
    - Calibration data key
2. Conventions
    - Need to be the same across all parties
3. Typical Measurements
    - Tipping buckets
    - Rain gauges
    - Turbidity sensors
    - Water quality samples
    - Met Instruments (local and third party data)
    - Radar information
    - Laboratory instruments to be handled the same as field instruments
    - Manually sampled faecal indicators
    - Genetic sampling of biological samples
    - Bore hole data (in some cases dedicated, in others from BGS etc).
    - Aiming to map field by field, repeated each year.

4. Note that long term storage of physical samples is out of scope.
5. Need to discuss with EA what additional data should be acquired or catalogued and accessible.
    - Local Gauging stations
    - Other EA data within catchment.
    - EA think up to 140 GIS layers could be needed including: OS data, geology, EO data, many others. Need to establish the IPR position. Land cover data (EO).
6. Data Structures
    - Point Series
    - GIS features
    - Acoustic Doppler Current Profiler (ADCP)
        - Measuring spatial profiles of river velocity to produce profile series
    - Borehole vertical profiles

   - Velocity Profiles
   - High Resolution terrain models from LIDAR/Laser scanning.
   - Genetic Sequences
7. Models (usage includes ensemble sensitivity analyse to evaluate policy etc).
   - Groundwater models
   - Run-Off models
   - Pesticide Leaching Models
   - Interpolations and Aggregations in space and time
   - Statistical Models
   - Assume that model output and input types map onto the observation output and input types.
8. In principle there is a requirement to capture model descriptions (versions, codes etc), but we expect the DTC researchers to be tweaking models as they go, and are unlikely to document versions in detail.
   - In this project there is not a requirement to capture model outputs and description, however, if the model is itself a key part of the evidence for a policy decision, we should capture some model provenance.
9. Soft Data
   - Farm practice data
      - from discussion with farmers
      - Might have a common format for those interviews
      - Need transcript to be kept as raw data
      - Participants need to be able to map transcript to a common vocabulary
   - Web cam
   - Still images
   - Real-time footage (tied to instruments such as flow gauges, and to event logs).
   - Crop yields
   - External agronomic data
   - Pesticide usage
   - Questionnaires
   - Videos of workshops

Appendix B: Portal Exemplars

This appendix provides links to some example portals that provide a front-end to similar or related activities.

1. Healthy Waterways Partnership in South East Queensland, Australia: has a very effective way of connecting users to catchment management.

   http://www.healthywaterways.org/Home.aspx
2. Swiss Experiment – Interdisciplinary Environmental Research:

   http://www.swiss-experiment.ch/index.php/Main_Page
3. Catchment Hydrology And Sustainable Management (CHASM)

   http://research.ncl.ac.uk/chasm/

4. The LOIS River Monitoring Network:

   http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6V78-3SVYTV1-10&_user=910841&_coverDate=02%2F24%2F1997&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_searchStrId=1237376812&_rerunOrigin=google&_acct=C000047841&_version=1&_urlVersion=0&_userid=910841&md5=953df3101fb77a58716da60fcc16bafc

5. SWIMA at Nottingham University – a small scale sensor web demonstrator project:

   http://cgs.nottingham.ac.uk/cgs/projects_swima.html

6. Sensors Anywhere - . A large deployed end-to-end EU project using sensor webs in limited environments:

   http://sany-ip.eu/

7. The Tasmanian South Esk Hydrological Sensor Web:

   http://wron.net.au/au.csiro.OgcThinClient/OgcThinClient.html