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.