7. (b) **Objectives**

Please describe the general objectives of the project and the technical and scientific aims of the research which must be measurable and timebound, (please number the objectives). If your application is accepted, these objectives will be included in the agreement between you and the Department. Please, therefore, restrict your entry to the salient points and set these out clearly and concisely.

The general objective of this work is to provide a mechanism for storing, collating and accessing data from the Demonstration Test Catchments (DTCs) in a way that is usable for data providers, specialist data users and other stakeholders in the three catchments and elsewhere.

The proposed programme of work will be broken down into eleven Work Packages (WPs: described in 7.(c)), each of which is associated with a set of time-bound objectives that together will fulfil the general objective. This breakdown will make the work easier to manage and progress easier to monitor. The specific objectives are as follows, the numbers matching those of the WPs.

Note 1: The dates indicated assume a project start date of 1 Jan 2011.

Note 2: Because of the iterative nature of our approach, project deliverables will be made available in phased versions, so it is important to understand to what the dates below refer.

For documents (e.g. the data model), the date refers to the publication of the first complete version, which will be regarded as quasi-definitive for the project. This may however be subject to further updates during the project in the light of changing circumstances, in which case a final version will be released at the end of the project as part of the documentation set.

For software, the date refers to the release of the first, complete, tested beta version of the code in question, for final, formal evaluation by stakeholders. Earlier, partial, releases will be made available in accordance with the agile development methodology followed by the project.

1.1 To organise a Project Management Board for reporting to Defra (end March 2011).
1.2 To organise a Project Steering Committee, comprising scientists and technology experts, to act as “critical friend” of the project (end March 2011).
1.3 To produce a Quality Plan, describing the QA procedures to be applied during the project, and in particular to assure quality of project outputs (end June 2011).
1.4 To produce a Dissemination Plan and to manage dissemination and communication of outputs and achievements (end Sept. 2011).
1.5 To produce progress reports and a final report for Defra (end Dec 2011 and annually thereafter).

2.1 To engage with the data provider community through a series of targeted workshops (end Sept. 2011).
2.2 To deliver a Data Management Plan, as described in DTC-DMR-3 (end Dec 2011, with updates ongoing thereafter).
2.3 To deliver an ongoing online communication mechanism for data providers (end Sept. 2011).

3.1 To engage with the broader user community through a series of workshops categorised by DTC area (end Sept. 2011).
3.2 To deliver a prioritised list of functionality and features required by or requested by particular stakeholder communities (complete version end Jan. 2012).

4.1 To develop data models that define the semantics and syntax of the data and metadata to be managed by the archive, using a standard formalism such as UML (end Mar. 2012).
4.2 To define metadata schemas for the digital material in the archive, using relevant international and open standards (end Jun. 2012).
4.3 To implement content models for the different types of digital objects to be managed in the archive (end Dec. 2012).

5.1 To define formal vocabularies for describing the domain-specific objects, procedures and relationships with which the archive will have to deal, based as far as practicable on existing vocabularies and standards (first complete release end Mar. 2012).

6.1 To deliver an operational web-accessible archive capable of ingesting, storing and preserving the data objects identified in WP2 (end Jun 2014).
6.2 To deliver an archive that provides data curation and preservation services in accordance with the OAIS model (end Jun. 2014).
6.3 To deliver an archive that can ingest streams of data from the various sources identified in WP2, and which can be adapted to handle new data sources without major modification (end Jun. 2014).
6.4 To deliver an archive that can support a range of data object types (list to be determined in WP2) and can be adapted to handle new object types without major modification (end Jun. 2014).
6.5 To implement a flexible access control framework so that data is made available only in accordance with access rights, at a dataset and sub-dataset level (end Jun. 2014).

7.1 To specify and implement a generic and flexible framework for querying the archive (end Jun. 2014).
7.2 To implement a number of specific query services using the framework (end Jun. 2014).
7.2 To write guidelines on developing additional query services using the framework (end Jun. 2014).

8.1 To support the delivery and export of datasets in a variety of common formats, to be determined through engagement with potential user communities (end Jun. 2014).
8.2 To support delivery and export of data though a flexible and extensible mechanism, so that additional formats can be added as required without excessive modifications (end Jun. 2014).

9.1 To deliver access portals to accommodate targeted communities, as specified by Defra (end Sept. 2014)
9.2 To provide a framework facilitating the creation of additional portals subsequent to the project (end Sept. 2014).

10.1 To ensure that the archive, together with associated tooling and interfaces, meets the requirements of the various stakeholder communities (end Nov. 2014).

11.1 To host the data, archive and associated portals and tools after the lifetime of the DTC project with guaranteed longevity of access in the absence of continuing funding (mechanism in place at end Sept. 2014).
11.2 To provide access to the data, archive and associated portals and tools that is free at the point of access for all interested stakeholders (mechanism in place at end Sept. 2014).
11.3 To produce an Exit and Sustainability Plan that describes the options and potential business models for taking the system forward once the funding period is over (end Sept. 2014).

---

7. (c) **Approaches and research plan**

Outline the approaches to be used to achieve the objectives, describing the scientific context where appropriate. Set out the work plan for the life of the project stating clearly how you intend to proceed (please include a GANTT chart if appropriate). The Approaches should be given the same number, and in the same order, as the Objectives and must be clearly cross-referenced to the numbered Milestones set out in Section 8. Where there is more than one contractor, please show clearly the roles of each. If your application is accepted, the Approaches and Research Plan and Milestones will be included in any contract issued. Please, therefore, restrict your entry to the salient points and set these out clearly and concisely.

Description of Work Packages

The work is broken down into eleven WPs. Each WP will involve an approach that is designed to meet the objectives specific to that WP, as indicated by the numbering scheme: WP1 will use Approach 1 to meet Objectives 1.1, 1.2, …, and so forth. Each WP will contribute towards meeting a number of project Milestones, as described in Section 8. The temporal relationship among the WPs is shown in Figure 1.

Work Package 1: Project Management (Lead partner: FBA. Coordinator: M. Haft)

Objectives: 1.1-1.5
Milestones: 1, 2, 3

This WP will be ongoing throughout the project.

Approach 1a: Set up phase. A Project Management Board (PMB), comprising coordinators of the WPs, will set up the management structure for the DTC Archive consortium (FBA and KCL), will establish governance and communications procedures within the consortium, and will arrange all consortium contractual agreements regarding financial and other issues. It will also establish communication with the DTC consortia.

The PMB will establish a Project Steering Committee (see **The Steering Committee**, below), which will comprise information scientists and technology experts as well as freshwater scientists, to act as "critical friend" of the project; the Steering Committee will liaise with the PMB to produce a Quality Plan. The PMB will produce a Dissemination Plan for publicising the outputs of the project, based on the proposed outputs in Section 14.

Approach 1b:Ongoing. The project will be managed in accordance with Defra's Joint Code of Practice for Research, and the Project Management Board will establish quality procedures that will be applied throughout the project. The project will require close liaison between FBA and KCL, so there will be monthly team meetings or teleconferences, face-to-face project meetings every 3 months, and more informal communication as required. Communication with the DTC consortia will be managed through **WP2**.

The Project Management Board will oversee ongoing communication of outputs via publications, conferences, etc.; these will be in addition to the communication to stakeholders described under **WP2 and 3**.

The Steering Committee will meet every six months and will oversee quality assurance through the lifetime of the project.

Realisation of objectives will be reported within annual interim reports to Defra, a final report to Defra and via regular updates posted on the project website.

Work Package 2: Data Provider Engagement (Lead partner: FBA. Coordinator: M. Dobson)

Objectives: 2.1-2.3
Milestone: 5

This WP will be ongoing throughout the project.

Approach 2: As the DTC consortia will be providing the majority of the data to be archived in the proposed system, engagement between the DTC data providers (and Defra) and the DTC Archive consortium will be essential to the success of the proposed project. This liaison will be ongoing throughout the project, but will be particularly focused at the start of the project, to finalise detailed requirements of results. The primary means of engagement will be through a series of workshops to be held with data producers.

This will commence with a meeting of DTC coordinators, to ensure consistency of approach and data format across the three projects. This will be hosted by the FBA, to enable the coordinators to meet associated staff and become familiar with the set-up.

Workshops will then be held with each of the catchment teams, involving coordinators of the various investigative areas that will generate data (surface water quality, groundwater quality, ecology, etc.), in order to understand data types and to negotiate efficient flow of data which, in turn, will inform the activities in **WP4 and 5**. A draft Data Management Plan (DMP) will be drawn up following these workshops. Actual timescales for data delivery will be determined by the work plans of the three DTCs.

Once data capture mechanisms are in place, a pilot period of data supply will be followed by further workshops with each catchment team, to identify and troubleshoot problems and to make final modifications to the DMP.

Subsequent workshops will be held every six months following completion of the DMP to allow interchange of ideas and experiences, along with modifications to data collections that may be introduced to support further short-term studies. These will be held at various locations to minimise individual travel time, and could include Defra (Nobel House) and Environment Agency (EA) (London or regional offices). They will be held in association with stakeholder forums (**WP3**) to ensure effective understanding of end-user requirements. These workshops will ensure both smooth flow of data and a coordinated approach to wider communication of outputs (see **WP1**).

Appropriate online communication mechanisms among DTCs and the Data Archive team will be installed. The specifics will be decided at the workshops, but are anticipated to follow the model currently followed in the FISHNet and FISH.Link projects (see **Previous and related work**, below), which hold regular Skype meetings, and communication via the JIRA system.

Work Package 3: Broader Community Engagement (Lead partner: FBA. Coordinator: A. Powell)

Objective: 3.1-3.2
Milestone: 4

This WP will be ongoing throughout the project and will build upon the FBA's considerable experience in community and stakeholder involvement in freshwater biology and conservation.

Approach 3: The details will depend upon the stakeholder engagement mechanisms already put in place by the three DTC projects, but we anticipate that the key elements will be as follows:

- Initial stakeholder forums (one per catchment), with the aim of exploring the kind of data being generated and how these align with stakeholder expectations; this information will inform the activities of **WP4 and 5**.
- Ongoing communication via stakeholder forums to be run in sequence with the six-monthly data provider workshops (**WP2**).
- Regular communication (e.g. e-newsletter) on current activities and events.
- Working with DTC project teams to provide demonstrations of data and outputs.

Web outputs, both of specific datasets to stakeholders (which can be restricted in access if required) and wider public outputs, will be the primary delivery output of **WP9**. These will be designed in consultation with key stakeholders to ensure that they are intuitive and straightforward to use.

During the course of the project we will work with DTC project coordinators to publicise the work through the FBA's own network of contacts, including FBA members, biological recorders, rivers trusts and wildlife trusts.

WP4 and 5, which will run in parallel, are essential for setting the standards for data handling (WP 6-8). Therefore they will run towards the beginning of the project, with some ongoing later activity.

Work Package 4 – Data and Content Modelling (Lead partner: KCL. Coordinator: M. Hedges)

Objectives: 4.1-4.3
Milestones: 7, 9, 10

Approach 4: This WP will proceed in three phases: (i) the development of an overall data model (or models) describing the semantics and syntax of the data and metadata for the archive; (ii) the creation of metadata schemas for the various types of digital object that will be managed by the archive; (iii) the creation of content models specifying the detailed contents of the digital object types at an implementation level.

It is inevitable that some changes in scope or requirements will be identified as the project progresses, resulting in changes to these various models. To accommodate this, we will establish procedures for managing these changes, which will ensure that changes are propagated through from the requirements to the system implementation.

The vocabularies and ontologies to be used in archive are being addressed for convenience under a different work package, **WP5**. Note however that there will significant interaction between **WP 4** and **5**, as these vocabularies will feed into the data and content models, and relationships and entities identified in the data model will need to be incorporated into the formal vocabularies. The two work packages will proceed in parallel, as shown in the GANTT chart below.

(i) The overall data model will be developed in consideration of the requirements already scoped and documented in the report "Demonstration Test Catchments – Data Management Requirements". A key aim of the model with be flexibility and extensibility – flexibility to support the wide range of types of digital material that the archive will be required to manage, including both raw data and processed data (such as simulations and analyses), and extensibility to allow additional categories of data to be incorporated without excessive modification. The data model in will be described in UML, with additional clarifications in textual form.

(ii) Appropriate metadata schemas will be defined (as XML schemas) for the various types of data that will be ingested into the archive (as outlined in DTC-DMR-3.3) and, where appropriate, metadata will be used to describe digital data at lower levels of granularity than that of the entire dataset (i.e. within datasets). We will draw on the principles of the 'Observations and Measurements' standard (the forthcoming ISO 19156), generic metadata standards such as MODS and standards for scientific datasets such as the Core Scientific Meta-Data Model (CSMD), as well as discipline-specific standards such as WaterML, SensorML, GeoSciML, MOLES, Ecological Markup Language (EML), the AgMES extension to the Dublin Core Standard for Agricultural information and CSML, for example by using a generic schema such as MODS at the top level with embedded extensions to accommodate domain-specific requirements.

Using the CSMD, which was developed at the Science and Technology Facilities Council (*citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.106.2880*), would in particular facilitate future extensions of the system by providing a generic metadata environment for supporting data-intensive scientific applications, as well as interoperability with other scientific data archives.

(iii) We will also define a set of content models corresponding to the various types of digital object that will be managed by the archive. The archive will be implemented using the Fedora repository software (see **WP6**), and in Fedora representations of digital objects are formalised as "content models" (see http://sourceforge.net/apps/mediawiki/ecm/index.php?title=Main_Page), which may be regarded as complex "data types" or templates for digital objects. These content models facilitate the management and reuse of digital objects by representing their internal structure and behaviour formally, providing consistent, standardised and interoperable patterns for representing objects, resulting in collections of similar objects that can share common processing. Note that, although these content models are linked to Fedora in that the repository software contains built-in support for them (e.g. validation services), they are defined in an entirely open and implementation-independent fashion using XML schemas (including some RDF, RDF(S) and OWL elements).

Work Package 5 – Vocabulary and Ontology Development (Lead partner: FBA. Coordinator: H. Schwamm)

Objective: 5.1
Milestones: 8, 15

***Approach 5*:** Semantic approaches allow sophisticated queries to be made across datasets. Appropriate query services will be developed in **WP7**, but a key pre-requisite for this is the development of appropriate vocabularies.

We will as far as practicable use standard, formal vocabularies to describe domain-specific objects, procedures and relationships, so that the archive is understandable by stakeholders, and enables it to interoperate and exchange data with other systems. As the data generated will relate primarily to agricultural impacts, these will include AGROVOC, the FAO agricultural terminology tool used for the international agricultural information System AGRIS and the CABI (Commonwealth Agricultural Bureau) thesaurus (which contains agrochemical vocabularies) in addition to the Defra Farm Types, Farmers Attitudes vocabularies. The physico-chemical vocabularies will be based on existing IUPAC vocabularies and other appropriate vocabularies/Ontologies such as those in the Semantic Web for Earth and Environmental Terminology (SWEET) modules. In the aquatic domain, the Aquatic Sciences and Fisheries Information System Subject Thesaurus will be assessed for its relevance. The FBA's Freshwater*Life* programme and the FISH.Link project have been developing a number of domain ontologies that are relevant, and others are identified in DTC-DMR-4.11; these will provide the basis for vocabulary development.

However, it will be necessary to extend and harmonise these existing vocabularies in order to support the proposed archive, a process that will be facilitated by the series of targeted workshops to be held during the first year of the project as part of **WP2 and 3**, primarily with the data provider community, but also with the broader community of potential users of the data. Should it prove necessary to produce any new specialist domain Ontologies (e.g. for farm surveys) these will be constructed to OWL2 logic standards using robust knowledge elicitation techniques.

This WP will initially involve assessment of current vocabularies to identify the most appropriate for the data types and end users, with some additional work over year 2. During year 3, it is anticipated that domain ontologies specific to the work will be identified and will require refinement.

In addition to the choices of vocabularies and their refinement, WP5 will work closely with WP4 and WP7 to define SKOS models for implementation and semantic web utilization of the developed vocabularies. Again, existing SKOS models (e.g. the AGROVOC SKOS Model) will be evaluated for use before any specific development is undertaken.

The development of vocabularies is a dynamic process, and as with **WP4** it will be necessary to establish procedures for versioning vocabularies and for managing changes, to ensure that changes are propagated through into the query services in a consistent manner.

Namespaces for open dissemination and re-use of vocabularies will be established and maintained.

WP 6-8 are inter-related and will run in parallel, although with **WP8** starting slightly later than the others. They will be initiated once the groundwork has been carried out by **WP 4 and 5**. They will therefore not commence until the main activity in **WP4 and 5** is complete, and will run until towards the end of the project.

Work Package 6 – Archive Implementation (Lead partner: KCL. Coordinator: M. Hedges)

Objectives: 6.1-6.5
Milestones: 11, 16

Approach 6: For implementing the archive we will use at the core the Fedora digital repository software. Fedora offers a flexible content model architecture that supports the representation of compound digital objects and aggregations, and allows multiple heterogeneous metadata schemas to be associated with an object. It contains built-in support for representing (as an RDF graph) the structure of compound digital objects and relationships between objects. Fedora is open source and uses open standards for its representation of its digital content; the architecture of Fedora is essentially service-orientated, with all functionality, data and metadata being exposed as web services (both SOAP and REST).

Fedora will provide the core repository, together with a API of web services providing low-level functionality. We will supplement this with a range of additional services, following the same web service paradigm, that implement the higher-level functionality required for the archive. Our general approach will be to develop services that implement atomic units of functionality that can then be combined to create workflows implementing a variety of more complex requirements.

These services will address the following areas:

1. Curation and preservation services, including bit-level and representation-level services (e.g. checksums, PREMIS preservation metadata services, extraction of Transformational Information Properties/Representation Information, file format conversion)
2. Ingest services, e.g. data validation services, services to create digital objects that conform to appropriate content models.
3. Metadata and indexing services
4. Query services
5. Data export, download, derivation and visualisation services

(4) and (5) are covered separately in **WP 7 and 8**, respectively. In **WP6** we address the core archival services, relating to (1), (2) and (3).

We will use these services to build up configurable ingest and preservation workflows. This framework will allow the automated ingestion of data from instruments and other sources, and its modular/configurable nature means that it can be adapted to new data sources and types with relatively little effort, by changing or replacing components. This flexible approach to moving data into (via a configurable set of workflow components) and out of (via web services) the archive will facilitate greatly the archive's ability to interface to and interoperate with other systems, including those listed in Archive_Spec_V4 (Requirements and Methodology, bullet 6), which will be addressed specifically.

Automation of access management and rights issues will be an important part of our approach, and we will investigate a number of ways of representing access rights in a machine-actionable fashion within the archive, for example XACML, XrML, ONIX and PREMIS. This information will be used by the archive access layer to manage access to the data, and will also enable machine-actionable rights information to be attached to any dataset that is transferred to 3rd party systems. As well as addressing access rights for individual data sets, we will address rights *within* datasets, for example access at particular levels of granularity or spatio-temporal resolution (e.g. http://homepages.inf.ed.ac.uk/fgeerts/pdf/demovldb.pdf). Appropriate rights metadata will be associated with datasets as art of the ingest workflows.

The development in this WP will not be undertaken from scratch; it will build upon, adapt and enhance a body of work already carried out by KCL on a number of other digital archive projects (see ***Previous and Related Work*** below)

Work Package 7 – Data Querying (Lead partner: FBA. Coordinator: M. Haft)

Objectives: 7.1- 7.3
Milestones: 12, 17

Approach 7: The ability to query the data in the DTC Archive will be fundamental to its successful use. The DTC Archive will make use of the semantic querying tool SPARQL to allow it to query data in the Archive by using vocabularies developed in **WP 5**. The FBA and KCL are already developing these tools in conjunction with the University of Manchester as a part of the FISH.Link project.

The FISH.Link project has already determined that the most common queries of freshwater scientists will involve data queries along the time axis and spatial queries, along with other parameters and thresholds of those parameters, e.g. "show me all streams in area x with a mean temperature under 5°C between date a and date b". The FISH.Link project has also determined that the current technological challenge to doing this is not making semantic queries (for which there are a variety of tools currently available, such as SKOS and SPARQL); nor is it retrieving whole files such as an excel spreadsheet relating to the query; but rather it is in finding sub-sections of a given dataset from many different files and extracting them into a derived dataset matching the semantic query that is made.

The DTC Archive project will therefore work closely alongside the FISH.Llink project, which is already investigating and planning to implement the technology to do this. A variety of potential tools already exist to convert Microsoft files to RDF as well as potential direct querying of data in a specified format such as the netCDF. The precise method for querying and retrieving data will involve close liaison with the data providers and users as detailed in **WP 2 and 3** as the file formats used to store data will greatly affect this.

A data querying portlet/tool will be developed (see **WP9**) to allow for the implementation of the above.

Work Package 8 – Data Visualisation, Derivation and Export (Lead partner: KCL. Coordinator: M. Hedges)

Objectives: 8.1-8.2
Milestone: 18

Approach 8: As the majority of the data to be dealt with by the DTC Archive will be over the catchment scale, GIS visualisation will be a minimum requirement for allowing users to view and interpret the data easily. The DTC Archive will make use of pre-existing data such as openly available Ordnance Survey GIS data and other freely available GIS data and systems in order to implement geographic visualisation. An appropriate software tool, such as the industry standard ArcGIS, will be used to create mapping services. Tools will also be developed to allow various data export formats for use by researchers, the specific formats being determined by **WP 2 and 3**.

Export of data will be developed in conjunction with **WP 6 and 7** so that researchers can derive and export subsets of the data they are interested in an appropriate format for their use. The FBA is involved in both the National Biodiversity Network and Global Biodiversity Information Forum, two biological distribution mapping initiatives; it also has experience of integrated information systems at a catchment scale; it will therefore draw on this experience in designing layers for this output.

The project will also enable the exposure of data (incorporating ownership and privacy requirements) using linked data standards thus supporting compatibility with the data.gov.uk project. Tools for doing this are currently being developed by the FISH.Link project.

Work Package 9 – Portal(s) Implementation (Lead partner: FBA. Coordinator: M. Haft)

Objectives: 9.1-9.2
Milestones: 14, 19

This WP will commence after the main activities in **WP6-8** are complete, as it depends upon them for its effective delivery.

Approach 9: Our development approach will be to make the data and derived functionality (visualisation etc.) available via atomic web services that can be combined into portlets to meet the needs of particular communities. We will develop as part of the project a number of such portlets, a suggested list of which is presented below, all of which will be compliant with JSR168 and JSR 286 portlet standards and will be deployed via the existing Freshwater*Life* portal, thereby allowing for the sharing of these portlets with other JSR 168 and JSR 286 standards compliant portals. Additionally, with the use of pre-existing tools, personalised JSR compliant portlets can easily be exported to non-JSR compliant portals such as iGoogle and Netvibes. The source code for portlets can also be ported to non-portlet environments with relatively little difficulty, thus allowing the re-use of this code to construct a bespoke portal by a third party without the need to adhere to the JSR portlet standards.

This flexible, layered approach will make it as easy as possible to integrate feeds and other outputs from the DTC archive either into newly developed portals developed by third parties or into existing portal environments and systems developed by other projects.

Suggested portlets:

- Data upload/ingest portlet

    - Adding data into the Archive; this portlet will be related to the Sensor portlet (see below)
- Search and browse portlet
    - Browse and discovery of data in the Archive
- Data query portlet
    - Tool for interrogating archived data (details in **WP7**)
- Maps portlet
    - Visualisation of data using standardised geographic tools
- Data export portlet
    - To allow raw, quality controlled or derived data to be exported in a variety of formats as determined during the project.
- Sensor portlet
    - The precise requirements of the sensor portlet will be determined by **WP 2 and 3** but are likely to make use of existing standards and tools such as SensorML

The list of suggested portlets presented is not complete, as the full list of portlets and other tools to be developed by the DTC Archive project will be determined over the course of the project via **WP 2 and 3**. Integration of the FBA library catalogue into the current Freshwater*Life* portal (currently being implemented as part of the FISHNet project) will also allow data in the DTC Archive to be linked to supporting bibliographic information as an aide to researchers and data users.

Note in particular that variants on individual portlets (e.g. for searching or querying) may be appropriate for different audiences or stakeholder communities – including professionals, practitioners and the general public – and that portlets will be combined to produce custom interfaces to support the needs of these different audiences. Maximising flexibility is the key to our approach here – atomic services, which can be configured to create a variety of portlets, which can in turn be aggregated to produce customised portals.

Work Package 10 – Evaluation (Lead partner: FBA. Coordinator: M. Haft)

Objective: 10.1
Milestones: 6, 21

This WP will be ongoing throughout the project.

Approach 10: The close involvement of the various stakeholder groups is key to ensuring that the system resulting from the proposed project meets those stakeholders' varied requirements. For convenience, the requirements may be considered under two headings, which also correspond to two broad categories of stakeholder: (i) archival requirements, corresponding to the needs of the data provider community, and (ii) user-facing requirements, corresponding to the broader group of potential users of the system.

We will identify and engage with potential users as part of **WP3**. The requirements of these users will refined during the process of engagement, as they see what is possible from the proposed system. Thus, development will follow a user-driven, rapid prototyping methodology, involving incremental cycles of implementation and evaluation in close collaboration with the stakeholder communities. Primarily, these will be the catchment-specific communities identified by the three DTC projects, but we will also solicit feedback more widely (although in less depth), for example in the extensive Freshwater*Life* community. Later (near "final") versions of the system will be trialled with stakeholders, and will evaluate the degree to which the requirements, use cases and "user stories" have been met.

The primary area of functionality that the system will provide will be as an archive for the DTC data. It will be easier to define these requirements at an earlier stage in the project, and indeed they will be described in detail in the Data Management Plan to be developed and agreed with stakeholders during **WP2**. An Evaluation Plan will be written on the basis of the Data Management Plan, detailing the specific criteria and tests against which the archive will be evaluated, and it is this document that will serve as a baseline for acceptance of the system. Notwithstanding this, it is likely that the requirements may evolve in response both to external changes and to internal project research, so we will still follow an iterative approach with continual stakeholder engagement.

Evaluation will be ongoing, but in the Project Plan we will define particular “evaluation milestones” at certain points in the project lifecycle, to ensure that the project keeps on track.

Work Package 11 – Sustainability and Appropriate Business Models (Lead partner: FBA. Coordinator: M. Dobson)

Objectives: 11.1-11.3
Milestone: 20

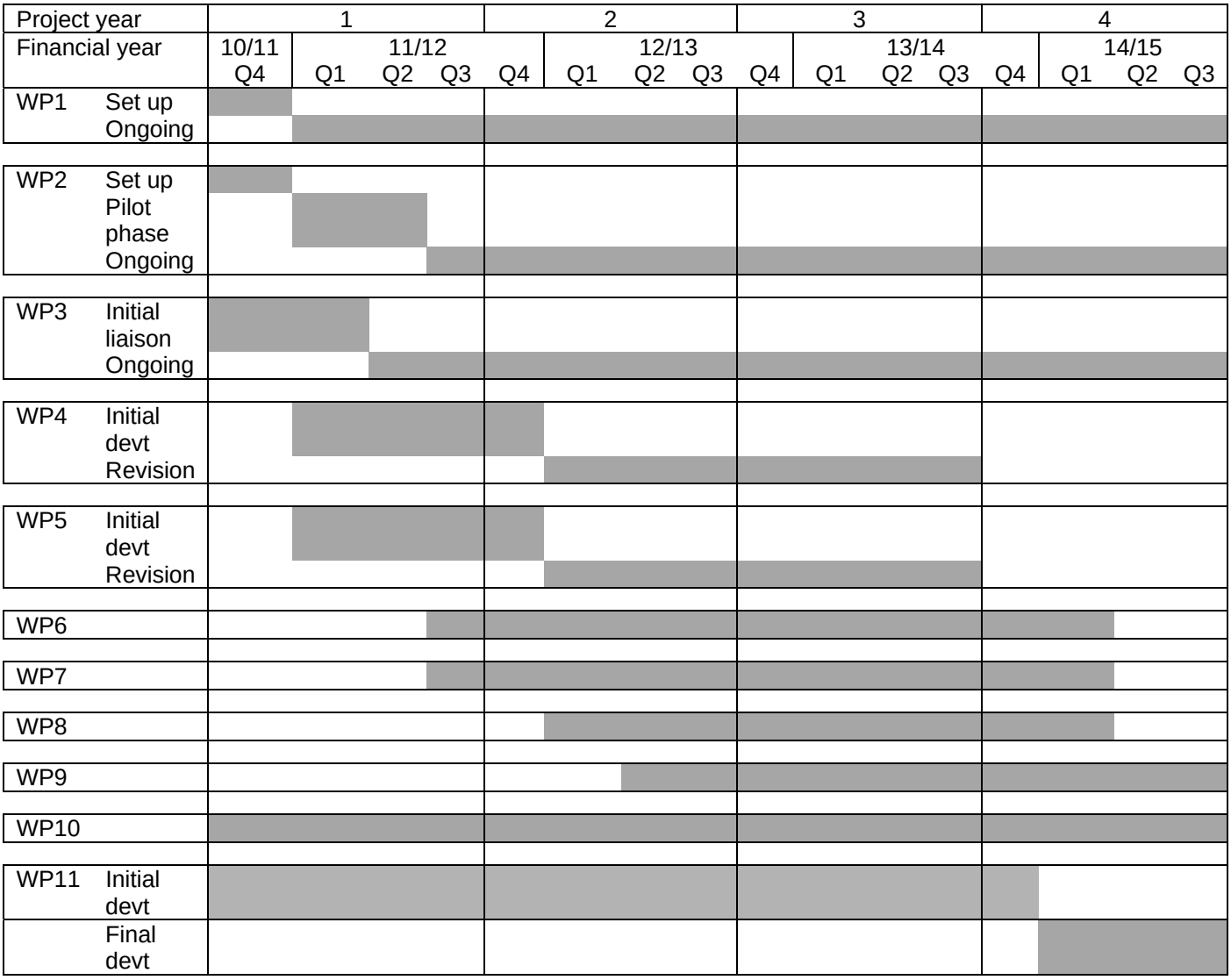
Approach 11: This WP will be active towards the end of the project, as its actions are dependent upon financial and other conditions as the project comes to its close.

Sustainability is a key issue for all web-based services and digital resources, an issue that is being examined and addressed in detail by the FISHNet and FISH.Link projects. Moreover, the FBA recognizes that data provision and archiving is extremely important in the digital age. It forms a natural extension of the library services originally established by the FBA in the 1930s and remains an important component of its charitable objects. Integrating digital archives into the management of the FBA library and associated services is an important long-term goal for the FBA. Indeed the FBA Business Plan states that: “holding information resources is considered to be important for the FBA’s existence” and has as one of its objectives: “to maintain the FBA’s position as a major holder of information on freshwater sciences”.

The DTC archived data will be one of many datasets that the FBA maintains as part of a continuing series of projects, from which the entire stored archive benefits and this process will continue. As the nature of funding opportunities changes over time, it is not possible to provide specific details at this time, but this approach has so far allowed the FBA to maintain digital datasets for many years. In order to produce an optimum outcome for the DTC archive we will produce an Exit and Sustainability Plan that describes the options and potential business models for taking the system forward once Defra funding runs out; this will build on an equivalent plan being developed for the FISHNet project.

Following appropriate stakeholder consultation, this data can be made accessible beyond the end of this contract: either accessible to all, accessible to restricted users via password, or a combination of the two depending upon each individual data component.

Figure 1. Gantt Chart showing timelines for Work Packages.



Development Approach

WP 2 and 3 will produce analyses of the requirements for the proposed system, from the point of view of two distinct but overlapping communities (data providers and stakeholders). These analyses will provide the baseline functionality for the development. Note that, while we expect the requirements from **WP2** to be very well defined, those from **WP3** will be more speculative, containing scenarios and "stories" of how the targeted stakeholder communities might interact with the system, and how they might use it in their activities; as such it will be incomplete and one cannot assume that all the functionality requested will be practicable. Consequently, to ensure that the resulting system meets the needs of the targeted communities, we will follow a phased, user-driven, evolutionary approach, involving incremental cycles of implementation and evaluation in collaboration with the stakeholders.

Specifically, we will follow an agile approach to software development, which involves working in short developments cycles or "sprints" comprising about 4-5 weeks of work, with release of a "working" software to interested stakeholders at the end of each sprint; in our approach, we anticipate releasing to targeted stakeholders to begin with, then opening up as the project progresses. At this point project stakeholders can provide feedback on the new release, report bugs, test new features, request improvements to existing features, and so forth. Associated with each sprint cycle, the project team will hold a meeting to review the previous sprint and to plan the next one: functionality and features are added to the schedule for the sprint, and notes from the planning meetings will be made available to stakeholders so that they can see what is planned for the sprint and the justification for the choices made.

Thus the overall direction of the project is defined by the set of required, requested or desired functionality and features that have been identified (and prioritised) with the stakeholders – in the agile methodology we are following, this is termed the "product backlog". The monthly sprints represent the process by which the progress of the project is monitored and evaluated, and its direction readjusted. When each sprint is planned, functionality and features are taken from the "product backlog" and put on a separate list called the "sprint backlog", and developers then work on adding these features to the software.

A final level of agility is provided though daily meetings of the development team of no more than 15 minutes in length, at which issues relating to the day's work are discussed. Because the progress is reviewed on a daily basis, developers are able to respond quickly to any problems that may arise, and to change their detailed, short-term plan accordingly. This means that these problems do not impede their ability to release a product at the end of the sprint, and it because of this agility in the face of events that the term agile is used to refer to this development methodology.

Project Management Structure and Reporting

The project will be coordinated by Dr Michael Haft (FBA), with the close assistance of Dr Michael Dobson (FBA) who, as FBA Director, has overall responsibility for ensuring that the supporting infrastructure and services are in place. The Project Management Board will comprise the leaders of the Work Packages, and a Steering Group of external experts will provide independent advice and guidance. Further details are provided in **WP1**.

Reporting to Defra will be as described in **WP1**. Reporting to data providers and stakeholders will be as described in **WP2 and 3**.

The Steering Committee

An external Steering Committee is seen by the partners as a valuable mechanism for ensuring high quality output. The FBA regularly uses such bodies, and its current Data and Information Services Advisory Group will form the basis of the Data Archive Steering Committee, with at least two members in common. Agreements to sit on this type of committee have already been secured in principle.

Project Partners

The project partners are the Freshwater Biological Association (FBA) and the Centre for e-Research at King's College London (KCL). The FBA and KCL have a good working relationship arising from their existing FISHNet and FISH.Link collaborations (see below).

The **FBA** is a scientific society, established in 1929 and a registered charity since 1966. Its primary objective is to meet the information needs of those involved in freshwater research and management by maintaining specialist information resources, along with effective knowledge transfer and knowledge exchange mechanisms. The FBA supports freshwater science through specialist events and publications, research grants, research and information services, provision of research facilities, and training. It has long-standing links with scientists and managers working in fresh waters throughout the world, and supports a small scientific staff and research portfolio of its own. In recent years it has been heavily involved in wider community engagement in freshwater science, and now has considerable experience in community and stakeholder involvement in freshwater biology and conservation, including governance positions in some of the major networks: Riverfly Partnership, National Biodiversity Network; National Trust; Royal Society of Wildlife Trusts; plus local initiatives such as South Cumbria Rivers Trust, Cumbria Wildlife Trust and the Lake District Still Waters Partnership. The FBA is a membership organisation, and its 1500 members include freshwater researchers, educators and enthusiasts from academia, industry and the voluntary sector around the world. The FBA has held information resources since is foundation and now has one of the world's largest specialist freshwater libraries, along with an extensive data archive. Since 2001 it has actively moved into digital data archiving and sharing, initially through the Freshwater*Life* initiative (see below) and more recently through data management research projects, particularly those funded by the Joint Information Systems Committee (JISC: see below). It provides online data and information resources for clients including the EA, and hosts websites on behalf of external organisations and consortia (e.g. www. riverflies.org; www.windermere-science.org.uk).

The **Centre for e-Research** (CeRch) at KCL is a research centre located in Information Services and Systems (ISS) that aims at facilitating interdisciplinary, institutional, national and international collaboration. The Centre is unusual in that it is both an academic centre: researching, publishing and teaching in its areas of expertise, including contributing to the UK Research Excellence Framework (REF) and developing and teaching on the MA in Digital Asset Management (MADAM) Programme; and a contributor to institutional activities in support of research across disciplines, including research data management, research infrastructures, digital archives, and digital curation and preservation. This unusual combination of research, development, and practice provides the centre with a unique insight into how theoretical work can be applied and translated into practical applications and services. The Centre comprises a mix of academic researchers, librarians, systems architects and analysts, developers and programmers, and departmental and project managers and administrators. The Centre works collaboratively with researchers, research teams and groups, and has partners in research projects across KCL. It also works in partnership with other UK HE, with European institutions, and internationally with HE library and research institutes. It has an extensive project portfolio funded by JISC, the Arts and Humanities Research Council (AHRC), the Engineering and Physical Sciences Research Council (EPSRC), and the European Commission.

Previous and Related Work

Freshwater*Life*

The FBA Freshwater*Life* programme was established in 2001 by partners from academia, regulators such as the Environment Agency and the private sector to promote information sharing and good practice to the freshwater community, through awareness raising and by provision of electronic information facilities. The Freshwater*Life* network provides a convenient and efficient means of engaging with individual researchers and institutions with an identified interest in improving data management and information flow. Programme objectives are set by the partners, collaborators and users and reflect the needs of researchers, educators and other information users. The varied work undertaken by the programme team includes many aspects of data management, application of newer technologies - such as RDF and ontologies - to freshwater science, facilitating information exchange and encouraging communication, debate and sharing amongst participants. The cross-disciplinary team specialises in providing an interface between researchers, conservationists, decision makers, educators, enthusiasts and IT specialists.

FISHNet

FISHNet (Freshwater Information Sharing Network) is a JISC-funded project that aims to help freshwater scientists more easily share scientific data. It is collaboration between the FBA and CeRch.

The aim of the project is to understand the needs of freshwater scientists with regard to sharing data and exploring ways to facilitate this. This project has a relatively short timescale of 18 months (October 2009 to March 2011). Time and effort are often a problem with the long-term curation of datasets. We also know that for many researchers, data management and infrastructure can be an issue for which support is not always available from their institutions. The freshwater science community is also spread thinly across the country in many institutions, and so there are many potential benefits of having a network of properly and effectively managed data available to those working in this field.

The project initially involved gathering information from freshwater researchers as to how they use and manage data, along with their concerns over data protection, copyright and access control. This information is now being used to implement a pilot web tool to help share and manage these data. The final FISHNet system will be available online at the Freshwater*Life* website by the completion of the project in March 2011.

FISH.Link

Motivated by the large quantity of diverse data in the freshwater biology community, the JISC-funded FISH.Link project, shared by FBA, CeRch and the University of Manchester, will provide a demonstrator of the benefits of publishing data by illustrating how data can be combined, re-purposed and reused with attribution and provenance information to promote data sharing. The project intends to support the sharing and integration of research data through the application of lightweight vocabularies and vocabulary mapping, facilitating integration of data sets, and moving towards the Web of Data that forms the current Linked Open Data vision. A case study that addresses a real scientific question is being used to provide motivation, requirements and support evaluation. The project runs from August 2010 to July 2011.

Other CeRch projects

As well as these projects relating specifically to freshwater science, CeRch has been involved in various projects addressing research data management, digital curation and related issues, and the experience gained from these will be leveraged in the proposed DTC archive project. The relevant projects include the following:

- BRIL, which is developing an archive for the experimental data and processes of biophysics researchers within the School of Biomedical and Health Sciences at KCL.
- ASPiS, a joint project with the Science and Technology Facilities Council (STFC) that investigated the use of the iRODS data grid middleware for managing large distributed research datasets, in particular addressing issues of access management and data provenance.
- PEKin, which is developing a preservation archive for a variety of research data and records systems across KCL.
- gMan, implementing a Virtual Research Environment for humanities researchers using EU research infrastructures and tools.
- CMES, an AHRC-funded project developing reusable content patterns for complex digital resources in the humanities
- Kindura, a collaboration with the Science and Technology Facilities Council looking at the use of cloud technologies to support research data repositories and data-intensive research computing.

The KCL staff include information scientists with expertise in digital archives and curation, and with practical experience of implementing archival and curation systems for research data. They will ensure that the project team has an appreciation of the state of the art in these fields, and of the issues involved in implementing such systems in practice. They will also facilitate communication with the wider information science, digital curation and e-science communities, as part of the project's dissemination programme.

Other FBA projects and strengths

ASFA (FAO Aquatic Science and Fisheries Abstracts) has commissioned the FBA to carry out several projects over the past year, including updating a geographical vocabulary for abstracting and indexing purposes; and digitisation of grey literature and making documents available via open access repositories.

In addition to its own titles, the FBA publishes an online first journal – *Inland Waters* – on behalf of the International Society of Limnology.

FBA staff have been involved in various web portlet design projects in the past, notably for UN FAO (oneFish, rural finance, etc.); these are detailed in the CV of Ian Pettman.

FreshwaterLife staff were involved in defining and improving knowledge elicitation techniques for the production of specialist domain Ontologies. They have also worked with the NERC Ontology staff and the Ordnance Survey staff in relation to environmental and geospatial Ontologies. Most of these initiatives arose from their involvement with the EPSRC funded Ontogenesis network of which they were founder members and co-instigators.

The FBA has representation on BSi, CEN and ISO Biological Methods Committees; providing a strong grounding in standard setting and quality assurance. These are detailed in the CV of Roger Sweeting.

The presence on FBA staff of trained and experienced biological scientists, including Michael Dobson, Michael Haft, Anne Powell and Roger Sweeting, will ensure that the data archive team has a clear understanding of the data being produced and the uses to which it can be put, which will greatly facilitate communication with the DTC teams.

We believe that this combination of experience, that incorporates both domain expertise in the relevant scientific disciplines, and expertise in the theoretical and practical issues around data archives and digital curation, will be the key to a successful project.

Appendix – Points of Clarification

Storage Capacity

The FBA is investigating various storage options based on estimates of the amount of data that catchments are likely to produce. These estimates are based on information provided by the Eden Catchment team. Dataloggers are not anticipated to produce large quantities of data; video and photos will be the main contributors' storage requirements. The FBA's data storage solution is based on the best estimate to be made on likely data volumes but is also expandable to accommodate future needs. This is derived partially from the storage requirements of previous projects handled by the FBA and by a generous estimate of numbers of photographs and lengths of video footage that may be generated within each catchment.

The FBA's proposed solution is already being implemented and involves adding a Dell EquaLogic PS4000e which will provide an estimated further 9TB of usable storage capacity to the FBA's hardware infrastructure. This solution is scalable with the purchase of more disks and/or a further SAN Unit depending on requirements. This additional storage will not affect performance in any way. The pinch point in performance that would be felt first is the FBA's bandwidth on its connection to JANET via University of Lancaster, which is currently in the process of being upgraded the University.

Long-term Maintenance & Exit Strategy

The FBA has maintained a library, archives and long-term datasets for 80 years as a key part of its charitable objects and will continue to do so in the future. The FBA is committed to finding solutions for reducing the running costs of the archive and has explored a number of different business models. In the unlikely event that the FBA can no longer house the archive, the virtual server infrastructure used by the FBA allows for easy transfer of the archive to an alternative host with a similar system.

Potential revenue sources for reducing the running costs of the DTC Archive in the future, in the absence of further funding from Defra, include models already used in many successful Open Source Software projects, where the software (or in the case of the DTC Archive, any data not subject to restrictions in use) is made freely available but related report/services are sold. The FBA has, in the past, provided for sale various data-mining activities and information services relating to the FBA's hard copy library. Similar

services could be provided for digital data in the future; the precise nature of these is difficult to determine at the present time, as they relate to the nature of the data to be collected and the type of tools that individual catchments would require. Therefore service will be determined as the project progresses and engagement is established with the various catchments. However, access to the data via the portals developed in the project will remain free at the point of use.

Specific business planning in relation to long-term data maintenance appears at the end of the Gantt chart, but it will be under consideration from the beginning of the project. The Gantt chart reflects the extended period during which we anticipate that the long-term maintenance strategy will initiated and developed, along with a shorter period towards the end when it will be finalised.

Other Data Storage, including the Defra Greenhouse Gas Platform

There are a variety of metadata standards that are applicable to farm practice data, video and photographs. It is important that the choice of standards is made in consultation with the data providers for the DTC projects, as this is not a one-way process but requires data providers as well as the DTC Archive having input into the selection of the appropriate standard, which must also take account of the final likely use of the data.

Where the data type overlaps with ongoing projects elsewhere we will engage with the coordinating organisations to achieve consistency (e.g. the Nerc EVOp, North Wyke's farm platform (for farm practice data), and the GHG data model experts in the Defra AC0114 project (Dr Steve Anthony leads, Dr Bryan Lawrence of CEDA to consult on data model). Liaison with the Defra GHG platform will be an important part of ensuring GHG data compatibility. The FBA has already spoken to Steve Anthony at the GHG Platform to discuss how to proceed. The DTC Archive Project manager (Dr Michael Haft) has direct experience of GHG and land-use modelling which will aid in facilitating a close liaison with the Defra AC0114 project.

INSPIRE & Data.gov

The metadata standards used by the FISHNet project already involve EN ISO 19115 and EN ISO 19119 upon which INSPIRE is based. Extending these standards is a relatively straightforward task which will ensure that data complies with INSPIRE.

The FISH.Link project involves linked data approaches which are compatible with data.gov.uk; more information on how Defra publishes to data.gov.uk will be sought to ensure compatibility.

Both the FISHNet and FISH.Link projects are funded by JISC and are managed separately from the DTC Archive project; however they are producing tools and expertise which is of direct use by the DTC Archive Project.

Existing National Datasets and the Virtual Observatory

The FBA has had long-term close involvement with many dataset holders including EA, NE, SEPA, SNIFFER, Riverfly Partnership, Pond Conservation, NERC/CEH. The FBA shares ownership (i.e. joint IP) with CEH of a range of internationally important physco-chemical and biological datasets relating to the Windermere catchment and other Lake District water bodies. We have been able to gain access and use of data from Biosys, RHS, Rivpacs, National Pond Survey, Species Dictionary, uBio and GBIF for purposes of demonstration connected with Freshwater*Life* (FwL) and now for Fish.Link (macrophytes in tarns, water chemistry, altitude data, etc.).

We worked closely with FishBase in the early days of FwL and the FBA is a long time member of the NBN Trust (Dr Powell serves as an NBN Trustee); we have been able to link our species portals to NBN maps on the Gateway to present detailed of freshwater species and their distribution on FwL.

The FBA has had involvement in EU data management, for example in the STAR, Universe, FAME and other European projects related to the WFD.

The role of amateurs in monitoring and data gathering and the development of “citizen science” is of great interest to us as we have been working with and training freshwater recorders for many years. The FBA has worked with Cumbria Wildlife trust volunteers for the last 5 years (Cumbria Tarns Project) and are exploring the freshwater data held by the 47 wildlife trust through the Water for Wildlife manager at RSWT. The value of protocols and standards as well as careful training and quality control are key elements of our involvement in citizen science initiatives.

We are very interested in and have kept contact with the self monitoring initiative under the Catchment Sensitive Farming initiative – with all its challenges!

FBA personnel are looking forward to exchanging experience with, and learning more about the NERC VO. As the VO is an initiative to allow multiple physical locations to be used for research and experimentation, through effective data sharing, the DTC Archive is an obvious partner; indeed the DTC is seen as an exemplar of catchment data provision (Andrew Impey, NERC, pers. comm.), so the DTC Archive will become a model for future catchment monitoring initiatives. We will negotiate with NERC their access requirements in order to link the DTC Archive to the VO.

Access and Data Protection

Metadata defining access to, and rights concerning, digital objects in the repository will be included as part of the object definitions, in a machine-readable fashion. There are various standards for this; we will most likely use XACML for access control and PREMIS for rights information. XACML is an XML-based standard for expressing access control policies that is already supported by Fedora (the repository software we will be using). PREMIS is a standard for preservation metadata that is supported by the Library of Congress.

Datasets will be ingested into the archive via a configurable workflow system, whereby different automated actions may be taken (e.g. metadata extraction, creation of preservation versions) depending on the category of dataset. One of these actions will be to associate the appropriate rights/access metadata with a dataset. The categories and a set of access/rights profiles will be defined during the project, and will be extensible on both sides – we can add more categories of dataset and more access profiles as we see fit. Of course a default, restrictive, profile can be associated with non-configured categories of dataset, and access rights can then be assigned via the user interface.

As well as addressing access rights for individual digital objects (i.e. datasets), we will address rights *within* datasets, for example with regard to anonymisation or level of spatio-temporal resolution. We will take a decision on when this should be done – on ingest (e.g. creating more than one version with different access restrictions) or on the fly (i.e. generating the appropriate version when a user requests access) – during the project once the datasets are available. These two options need not be exclusive.