Deliverable

D6.1

Requirements Specification for the Sharing Sensitive Scientific Data Test Bed

WP6 Sharing Sensitive Scientific Data Test Bed

8 Dec 2008

Version 2.1

Consequence

Context-aware data-centric information sharing

FP7-ICT-2007-1

ICT-2007.1.4. Secure, dependable and trusted Infrastructures

Grant Agreement 214859

LEGAL NOTICE

The following organizations are members of the Consequence Consortium:

Europäisches Microsoft Innovations Center GmbH,

BAE SYSTEMS (Operations) Limited,

Hewlett-Packard Italiana,

Imperial College of Science, Technology and Medicine,

The Science and Technology Facilities Council,

Consiglio Nazionale delle Ricerche,

Centre for Research and Telecommunication for Networked Communities.

The information in this document is provided 'as is', and no guarantee or warranty is given that the information is fit for any particular purpose. All warranties and conditions, express or implied, concerning the information, are excluded. The user uses the information at its sole risk and liability. Neither the Consequence Consortium, nor any member organization nor any person acting on behalf of those organizations is responsible for the use that might be made of the information in this document.

The views expressed in this document are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission or the member organizations of the Consequence Consortium.

This document is for general guidance only. All reasonable care and skill has been used in the compilation of this document. Although the authors have attempted to provide accurate information in this document, the Consequence Consortium assumes no responsibility for the accuracy of the information.

Information is subject to change without notice.

Mention of products or services from vendors is for information purposes only and constitutes neither an endorsement nor a recommendation.

Project acronym: Consequence

Project full title: *Context-aware data-centric information sharing*

Work Package:	**6**

Document title:	**Sharing Sensitive Scientific Data Test Bed – Requirements Specification**
Version:	**2.1**
Official delivery date:	**31 Dec 2008**
Actual publication date:	19 Dec 2008
Type of document:	Report (modify as necessary)
Nature:	Public (modify as necessary)

Authors:	Shirley Crompton, Benjamin Aziz, Michael Wilson, Alvaro Arenas, Brian Matthews
Approved by:	Representatives of EMIC and Create-Net

Version	**Date**	**Sections Affected**
0.1	April 2008	Valid for "versioned" deliverables like architecture document
1.0	October 2008	Initial version.
1.1	November 2008	Updated version with sections 4 and 5 and Appendix 2 completed. This version also incorporated partners' comments.
2.0	December 2008	Version updated in line with internal reviewers' comments.
2.1	8 December 2008	Version for external review.

Table of contents

1. Introduction

This document describes the Sharing Sensitive Scientific Data test bed. The test bed uses a multi-organisation research project in Structural Biology (SB) to highlight the data sharing requirements in scientific research as gathered from practising scientists and providers of large research facilities in the Science and Technology Facilities Council (STFC).

In a knowledge-based society, access to exclusive information confers competitive advantages. Public funding agencies and research organisations, including STFC, are keen to foster on-line community research resources in order to maximise access, to promote inter-disciplinary and cumulative research. In the academic domain, data arising from publicly funded research is considered public asset. According to guidelines of the Organisation for Economic Co-operation and Development (OECD) [1], this resource should be made openly available to the maximum extent possible. This principle underpins the agenda of the coalition of UK Research Councils (RCUK) to enrich society and contribute to the national economy through leveraging research to promote knowledge transfer and innovation [2]. However, to encourage a free flow of research data, there must be sufficient safeguards to protect a contributor's intellectual claims or property rights – which, after all, motivate research and innovation. Moreover, there may be legislation such as the Data Protection Act (DPA) 1998 which governs the publication of information. For instance, research involving the use of National Health Service (NHS) patient data must be anonymised before public release to protect the privacy of the human data subjects. Similarly, stakeholders to the research data may themselves have specific policies or requirements which limit how the data can be shared. For example, the STFC ISIS and Diamond Light Source (DLS) facilities require that experimental data obtained on their facilities by publicly-funded projects be released into the public domain after an initial period of exclusive use. It is sharing data during this exclusive time window that presents the biggest challenge to data owners and providers. The data owners may have commercial patents or scholarly articles pending and these would be invalidated by prior publication of the data. The challenge is how to protect data owners' intellectual property right (IPR) without imposing excessive restriction that may stifle data exchange and impede research activities. Once academic data is released into the public domain, there is generally little concern regarding access. The emphasis is shifted towards securing data integrity and to track derived data for IPR or quality management purposes.

As a pragmatic solution, stakeholders commonly use legally binding data sharing agreements to control how their data is shared and disseminated. These agreements contain policy statements on the access, usage conditions and obligations for specific sets of data as well as references to external data sharing policies or protocols, like those of the funding agency and university hosts. Such agreements are usually drafted by senior managers and lawyers to express what can be decided in court should a breach occur. Enforcement is generally left to the discretion of the data owners, publishers and providers. In the academic domain, enforcement may range from simple mutual trust between individual researchers on one end of the spectrum, with data consumers expected to voluntarily observe the ethical and legal obligations pertaining to the data; to a complete lack of trust at the other end, with sensitive data secreted away on private repositories accessible to the selected few. A system based on mutual trust is simple to operate but not adequate to prove compliance as obligated by many data sharing policies [see 3,4,5] or regulatory legislation. For instance, DPA requires an organisation to deploy appropriate technical and organisational security measures when processing personal data. The data processor must be able to demonstrate that such measures exist or risks prosecution. Similarly, there are many aspects in data sharing agreements relating to usage that are not addressed by a simple grant or no grant type access control system. For instance, a well-known drug company only uses approved, self-contained High Performance Computing (HPC) infrastructures to run simulations involving proprietary data; computing grids with nodes distributed across administrative domains are considered too risky.

Inter-disciplinary study and collaborative research with industry is changing the way researchers interact, share data and manage intellectual properties (IP). With increasing commercial exploitation and ambitious international experiments tackling grand research challenges, research data is becoming too expensive or even impossible to replace. To promote a free flow of research data in this complex environment, there is a need for a secure data sharing and dissemination framework that addresses issues such as context-aware usage and obligations, data integrity, derived data, privacy and confidentiality.

In the next section, we outline the test bed scenario for sharing sensitive scientific data and the selected use cases. The use cases are described in an informal format focusing purely on user and business requirements, and the constraints of the existing infrastructure. Section 3 presents the main requirements for the test bed; Section 4 assesses the risk and threat associated with different levels of trust in the framework and Section 5 outlines the evaluation plan. This document highlights the requirements gathered to date, it is envisaged that further requirements will emerge as development of the test bed progresses. In addition, it should be noted that the terms ‘data sharing agreement’ or ‘policy’ or ‘protocol’, and ‘high level policy’ or ‘agreement’ all refer to a textural document unless otherwise stated. A glossary of abbreviated and domain specific terms is provided in Appendix 1.

2. Scenario

2.1. Background

STFC hosts large science facilities that are used by over 15,000 scientists annually from around the world to produce data about the structure of materials. The data can be used in many ways, from advancing fundamental knowledge of the universe to practical applications to create new products, improve manufacturing technology and design new drugs. There are many individuals and organisations involved in the research processes and the different parties will make agreements with each other on how to share data from their collaboration. In this data sharing scenario, the STFC facilities act as a single source or hub of data connecting multiple user groups or spokes, each obtaining their own bespoke data which is protected by system level policies refined from the agreements mentioned above. The users participating in the data sharing processes are affected by these agreements but may not themselves be signatories. For example, a funding agency may have made a generic, high-level data sharing agreement directly with the facilities regarding the management of data produced on the facilities by all its funded researchers.

STFC facilities use an integrated e-infrastructure (see 2.4.1) to implement a data management framework. The data management workflows harvest metadata and raw data directly from experiments conducted at the facilities’ beamlines. The data and metadata are maintained in community-based repositories as a public research resource. Scientists can use the e-infrastructure’s Data Portal to quickly access their current and past data, search related experiments and publications, etc. As a data hub, each facility is bound by multiple bilateral data sharing agreements with separate end parties. These high-level agreements are first interpreted by administrators at design time and then implemented by developers using a combination of metadata (see 2.4.2), application logic and system processes to provide role-based security to the e-infrastructure. Changes to authorisation policies will require manual interventions – which are costly, time-consuming and error-prone. There are also no mechanisms to monitor consistency and address conflicts in low level policies derived from the different data sharing agreements, or to enforce these policies once data is disseminated outside the e-infrastructure.

At the moment, the assignment of project role for access control purposes is automatically cascaded from the research proposal system. All named investigators on the proposal will be given the role of Principal or Co-Investigators. The system, as it stands, has limited capability for users to assign roles and tailor access to their data. In line with RCUK funding practice, the users, as grant-holders, have to comply with our data sharing policies as a grant condition. Therefore, the onus is on the users to ensure that their specific data sharing arrangements are compatible with our data management policies. If users need to keep their data confidential for different periods and between different groups, they will need to apply for exemptions and submit a justification for each special case. Optimising this process by allowing users to specify data policies for their data would lighten the data management burdens on both the users and the facilities. It will also help improve security through minimising human interventions and the early detection of potential breaches arising from policy conflicts. Data producers will be more inclined to share research outputs on-line if they are confident that their work will only be used according to their data sharing policies. By improving security, we could encourage users to deposit their refined or analysed data with us. The early retention of research outputs in a managed infrastructure will minimise the loss of knowledge through neglect or maliciously acts and promotes a timely release of valuable research data as recommended by the OECD guidelines [1].

Our e-infrastructure represents one means of sharing data in the lifecycle of a research project. There are situations when research data need to be disseminated outside of the STFC security realm. In a collaboration crossing administrative domains, researchers may need to share background IP (e.g. proprietary data, copyright software and hardware) in scientific workflows from behind their organisational firewalls. For reasons of autonomy and performance, researchers may also opt to use data locally. The work performed locally would in turn generate further data or foreground IP that have to be managed. The main questions are how to enforce policies on data and secondary derived data dispersed outside the control of the e-infrastructure?

Another issue is to what extend can usage be controlled? Research is a process of discovery and contains an element of serendipity. It is not feasible to nail down at the start of a project the precise approach and analytical applications to be used. Scientists regularly write computational programmes to analyse experimental results in specific ways in order to test crucial aspects of the models that they are developing. It is desirable that the Consequence enforcement mechanism is application independent. However, if we permit valuable data to be accessed by unknown applications, is it then possible to control how these applications use the data?

Due the prevalence of short-term employments which are often tied to research grants or Ph.D. projects, there is a high-level of mobility among research staff. Academic research is a highly specialised profession, so it is not unusual to find researchers moving between rival groups. In this context, it is useful for data sharing policies to evolve dynamically in line with the project ecology to ensure that access rights are revoked.

These are generic issues for the secure dissemination of scientific data in a cross-domain collaborative environment. The nature and basis of collaboration will prescribe the data sharing requirements and these could be extremely diverse. Different research domains also have their own culture and practices towards data sharing. For instance, raw data about space exploration is freely available from the National Aeronautics and Space Administration

(NASA). The emphasis in this case would be about protecting the integrity of disseminated data rather than restricting access. Pharmaceutical companies, on the other hand, are far more circumspect about their data. Not least because of IPR, their research may involve sensitive personal data which is subject to DPA. A key challenge for Consequence is to provide a flexible framework to accommodate these diverse data sharing landscapes.

2.2. Scenario Story

In contrast to the Crisis Management Test Bed which describes data sharing requirements in a highly dynamic emergency situation using a common XML structure – the Tactical Situation Object - our test bed focuses on a scientific research project involving researchers from different organisations sharing a variety of data in different contexts over the research lifecycle (Figure 1).

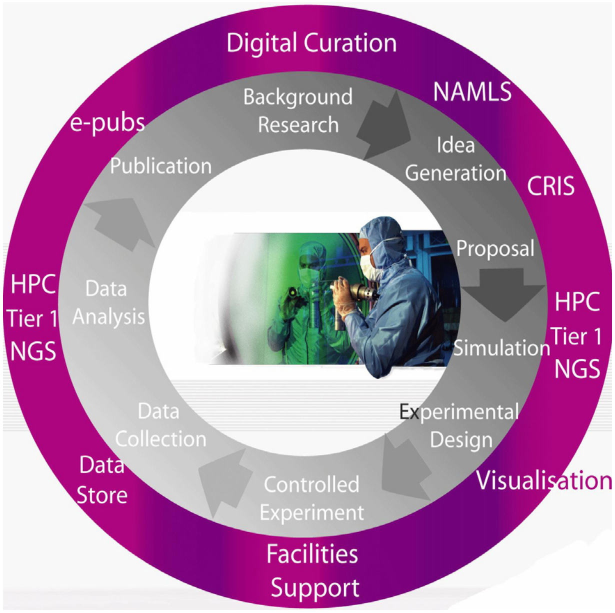


Figure 1. The scientific research lifecycle (inner circle with arrows) with the STFC e-Science support services available to support each stage (outermost circle).

The story centres on a 5-year public-private research collaboration co-funded by a key UK public funding agency in Bioscience and an SME biotechnology company, BioTech. The award is made under the agency's Connect Industry programme to promote research with commercial potentials and to encourage knowledge transfer between UK industries and the academic science base. The consortium includes academics from two UK Universities and industrial researchers from BioTech working in the niche market of supplying drug discovery software to the pharmaceutical industry.

The award covers capital and recurrent costs, a postgraduate studentship and industrial placements. The academic partners will benefit from the financial support, gain experience of the private sector's applied research and development environment and insights on commercialising research. The industrial partner, in turn, will gain access to cutting-edge research and technology to improve its products and the possibilities of recruiting appropriate trained staff at the end of the project. It will licence limited, time-bound access to its structure-based drug design algorithms and proprietary data to the partners. In line with the Bioscience agency's funding criteria, the partners have made agreements at the grant proposal stage on the licensing, ownership and exploitation of foreground and background IP and a protocol for the publication of research outputs. The agreement also includes a Schedule on

Good Data Management Practices. This requires the partners to use the STFC Data Portal and Information Catalogue (ICAT, see 2.4.1) to manage research outputs and to share sensitive data using the Consequence framework.

The collaboration's key scientific objective is to use X-ray crystallography to study HIV integrase, an enzyme that is essential to the HIV lifecycle. Insights into how the enzyme assists the virus to hijack the target human cell's survival machinery will contribute to the rational design of new viral inhibitor drugs. The project has two distinct phases: Phase 1 (years 1-3) covers pre-competitive research and Phase 2 (years 4-5) focuses on speculative research and the commercial exploitation of Phase 1 deliverables. A change of research personnel will take place between the two phases

During Phase 1, the academics will develop a method to synthesis integrase crystals of sufficient purity and size for X-ray analyses. These will be used in diffraction experiments to determine the integrase's 3D structure, both on its own and in complex with a host protein. The experiments will be carried out at the STFC synchrotron X-ray beamline. The solved structural information will then be used to study the enzyme's structure/function relationships. The student will use Biotech's background IP in this specific research. Phase 1 will produce generic project management and academic outputs ranging from laboratory procedures, experimental data, diffraction images, digital protein models, algorithms, documentations, progress reports and manuscripts etc. It is important that the Consequence framework can support a variety of proprietary and generic data formats. At the end of Phase 1, the postgraduate student is expected to have completed his Ph.D. and moves on to a placement in BioTech alongside a junior scientist from Structural Biology Department. The role changes will entail changing access contexts which the enforcement mechanism must support.

In Phase 2, the remaining scientists will study the modes of inhibitor binding of the enzyme and its utility as an alternative biological target for drug design. The out-placed scientists will work alongside BioTech colleagues to develop algorithms using information on the enzyme active sites to improve the quality of automatic function prediction used in the company's flagship drug discovery informatics software. Apart from the outputs outlined above, Phase 2 deliverables will include copyright materials and IP of potential interest to the pharmaceutical industry. Consequently, the partners agreed that data must be shared in a secure, fully audited framework in Phase 2.

2.3. Stakeholders

We present the test bed's principal Organisational (Section 2.3.1) and Individual (Section 2.3.2) actors. As the Individual actors are potential end-users, we also outline their IT profile and generic data management roles to help inform on their interactions with the Consequence framework.

2.3.1. Organisations

The main actors in this category are the financial sponsors and host institutions.

- The public funding agency – the UK BioScience Funding Council.
- The industrial partner – the BioTech Company.
- The academic hosts –
    - University A (Structural Biologists).
    - University B (Protein Crystallographer) .
- The facility provider - STFC Synchrotron X-ray Facility.

- The e-infrastructure owner – STFC Synchrotron X-ray Facility.

The organisation-level data sharing protocols or guidelines of these entities apply to the collaboration and need to be referenced by the master collaboration agreement as described in 2.5.1 and the miscellaneous bi-lateral project-specific data sharing/non-disclosure agreements between partners (see Appendix 3).

2.3.2. Individuals

Table 1 describes the typical actors in a research project. Only some will play a part in the selected use cases, we include the additional actors to provide a fuller picture of the typical data sharing requirements in a research collaboration. The primary Consequence users will be the Research Personnel as they are *the* research project. These users are expected to be highly IT literate and conversant with scripting and scientific programming languages like shell scripts, Python, FORTRAN and C. On the other hand, they are unlikely to possess similar level of competency in mainstream programming languages such as JAVA or DotNet, nor will they be familiar with formal policy languages. The university-based researchers will most likely be responsible for administering local security policies to manage access permissions and account policies on the computers in their laboratories.

Actor	General Description
Research Personnel	
Principal Investigator (PI)	Lead researchers with management responsibilities.
Co-investigator (CI)	Researchers contributing specific expertise to the collaboration.
Ph.D. Student	Research students undergoing a programme of academic training to develop their original research through working on the project. The work will comprise a key part of their theses submitted for their Ph.D. examination.
Support Personnel	
Research or Contract Manager	Administrators whose role is to facilitate the scientists in formulating, bidding and managing a research grant.
Beamline Scientist	STFC Scientists responsible for hosting users to carry out experiments at their beamlines.
IT System Administrator/ Developer	Administrators responsible for configuring and maintaining the local IT system. In the Universities, the academics may be responsible locally for their own machines, although they will not have the power to alter domain or organisational level policies, such as firewall configuration.
Review Personnel	
Thesis Examiner	Academics appointed to examine the Ph.D. thesis with regard to the award of the degree. The principal examiner will not be a member of the research project.
Journal Reviewer	Academics appointed by a Journal to review scholarly manuscripts submitted for publication by scientists.
Funding Agency Reviewer	Academics appointed by the Funding Agency to review the quality and progress of the research undertaken by funded researchers.

Table 1. Scenario Actors.

The Beamline Scientist and Review Personnel will be practicing researchers. They would have a similar level of IT competence to the Research Personnel. The Beamline Scientist will also administer local security policies on the beamline computers. In line with the Facility Data Sharing Policy, he/she has right to administer access on ALL raw data managed by

ICAT which originates from his/her beamline station. In contrast, the Review Personnel will have a more peripheral role in the collaboration and will mainly be data consumer.

The Research or Contract Managers representing partners from each organisation have the responsibilities to help develop and support the collaboration. They will play an active role in capturing the partners' data/IP sharing requirements at the proposal stage and, acting on the advice of legal experts, assist the partners in formulating the collaboration and data sharing agreements. The Research Managers are likely to be a competent IT user and are expected to author and manage the data sharing agreements on behalf of the project. As described above, it is envisaged that some actors will have the power to author local security policies for the computers under their control but they will not be able to alter domain or organisation level security policies. The implementation and administration of the Consequence components will be left to the IT specialists – the Administrator/Developer, from each organisation. Table 2 summarises the actor's generic role in managing information in the collaboration.

Actor/Role	**DSA Author**	**Define Local Security Policy**	**Consequence Developer**	**Doc/Data author**	**Doc/Data Consumer**
Principal Investigator		X		X	X
Co-Investigator		X		X	X
Ph.D. Student		X		X	X
Research Manager	X			X	X
Beamline scientist		X			X
IT Admin/Developer		X	X		
Ph.D. Examiner					X
Journal Reviewer					X
Funding Agency Reviewer					X

Table 2. An Actor's data management role in the collaboration.

2.4. Scenario Assumptions

The scenario describes typical data sharing behaviour in a cross-domain scientific project involving the use of large research facilities. In line with current research practices, we assume that the scenario actors will share scientific data in particular manners via particular transmission mechanisms. These assumptions represent constrains on the Consequence framework and are described below. Although the STFC data management framework is highlighted here, we wish to emphasise that this framework is not atypical of those adopted by similar large research facilities worldwide. In fact, the US Oakridge National Laboratory Spallation Neutron Facility (SNS), the national Synchrotron and Neutron Facilities in Australia, and the Canadian Lightsource facility are among those who have adopted the ICAT system (2.4.1). Furthermore, STFC is a member of an alliance of large research facilities seeking the development of a cross-facility data sharing and access infrastructure based on the STFC scientific metadata model (2.4.2).

2.4.1. Infrastructure

The scenario assumes that the researchers will be using a variety of technology to carry out their research and consume data. Their organisations will be connected via the SuperJanet or proprietary broadband network. Researcher within the same administrative domain will be able to communicate via LAN. Given the different institutional IT strategies and the academics' freedom to run their research units, we cannot expect the partners to use the same IT systems or have similar application architecture. Organisational firewall policies will also limit the extent the different groups can share digital resources. Cross-domain data sharing would likely involve the use of file transfer protocol, SSH, VNC, resource broker and centrally managed services like community data portals or virtual on-line collaboration tools. Researchers will also be using a variety of computing technology and OS platforms in their scholarly activities, from purpose-written scientific software to open-source community and proprietary software and officeware running on miscellaneous flavours of HPC, Linux, UNIX, MS , Apple ® Mac, etc.

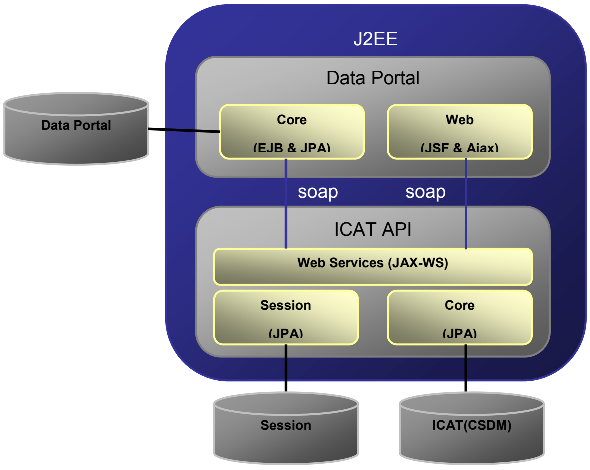


Figure 2. Data and ICAT Architecture.

In the scenario, the STFC e-infrastructure (or to be more specific, the Data Portal and ICAT, Figure 2) will be the main but not the only mechanism for researchers to share data. The Data Portal and ICAT together form an integrated information system. ICAT is a suite of software tools which provides access to specific information within large distributed data collections, large both in terms of complexity and size (into the PetaByte range). The Data Portal is an https web client that exposes ICAT data search and retrieval facilities. It can be linked to more than one ICAT instances in cross-facilities searches. ICAT uses the CSMD, which is described in the next section, to annotate and enrich the experimental data and maximise the value of this data for current and future users. ICAT can also be configured for direct access via web services by trusted SOAP clients. SOAP (Simple Object Access Protocol) is a platform independent specification for distributed applications to exchange information using standard XML-based messaging protocol over HTTP or other common network protocols.

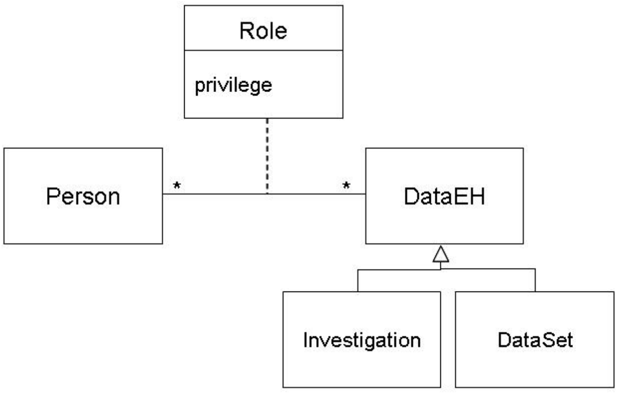


Figure 3. ICAT Authorisation Class Diagram

The current ICAT implementation has a standard role-based access control security model. The authorisation structure maps each system-defined role to a set of operation privileges, for example, download, select. A user is then given a role for a particular data container object or element hierarchy (Figure 3). ICAT currently recognises two data container objects – Investigation and DataSet (Figure 4). Authorisation policies are applied to these two key nodes and these policies are inherited by their child objects. Under this design, a DataSet is the smallest unit for applying authorisation policies, although it may only contain one single DataFile.

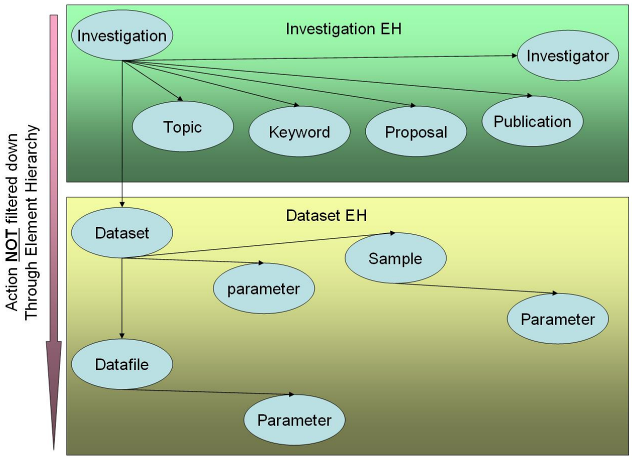


Figure 4. ICAT Data Object Container Structure.

2.4.2. The Core Scientific Metadata Model

The model of metadata adopted in STFC facilities, including the e-Science Centre, is the *Core Scientific MetaData (CSMD)* model defined originally in [6] and further enhanced in [7,8]. The model captures the general metadata required for cataloguing scientific data generated from experiments carried out by scientists over STFC facilities. In this way, it provides a standard format for sharing scientific data and eases citation, collaboration, exploitation and integration of the data into their applications.

The CSMD model of a metadata object is shown in Figure 5 below.

The metadata object is composed from the following entities, which provide information about the data being described by the object:

- *The Study*, which holds information about the name of the study, institutions involved, the investigators and their roles, the start and end times of the study, its source of funding, resources needed, its status and any investigations carried out under it.

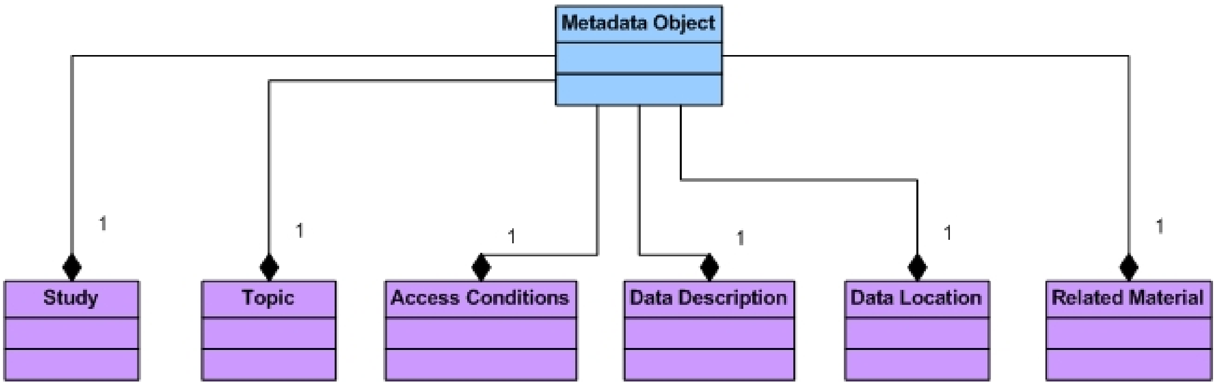


Figure 5. The CSMD Model of a Metadata Object.

- *The Topic*, which is an indexing entity that contains a set of keywords and subjects related to the study. It also contains other information about the keywords/subjects such as the discipline, i.e. the area of science, from which the keyword/subject are derived and a link to a dictionary/vocabulary/ontology describing the meaning of the keyword/subject.
- *Access Conditions*, which express the conditions that must be satisfied if access to the metadata/data is to be granted. These could include, for example, Access Control Lists (ACLs) for users, their IP addresses, statements on policies such as embargo-until-a-fixed-date policy, conditions of use and information on pricing and payment. This entity may also hold a link to an external authentication and authorisation service, which controls access to the metadata/data.
- *Data Description*, which contains information about the nature of the data themselves, such as the logical name of the set or the file, its type (format), status and topic. It also includes a logical description of the data such as values of parameters used in deriving the data, the data lifetime and the facilities used in its generation.
- *Data Location*, which refers to the physical location in which the data is stored and may include information about the retrieval (e.g. communication protocols).
- *Related Material*, which may be a link or a textual description of any material related to the data, such as other data generated by related investigations under the same study or indeed other related studies past or running in parallel with the current one.

2.4.3. Collaboration Format

The scenario describes a five year research project involving different researchers and support personnel from several organisations. Research is a complex activity: it comprises individual scholarly activity, collaborative activities and consultancy activities that occur in random order. The activities can take place both inside and outside the research institutions. For example, a researcher may work from home on a standalone laptop, or conduct experiments with colleagues in the laboratory. We envisage that researchers from each organisation work in the same research unit and share IT facilities. We further assume that the different groups will hold formal meeting periodically but will communicate mainly via email, phone and video. They will also share data centrally on ICAT, and exchange data by file transfer over SSL connection, email and via physical storage media like HP memory spot, flash drive or DVD. Data without contextual information is considered low risk and may be exchanged in printed formats, for example, images of diffraction patterns. Security considerations will ultimately determine the transmission method. For example, proprietary data must not be sent as an email attachment.

2.5. Scenario Use Cases

We describe use cases from two key aspects of the chosen scenario story - the specification of data sharing agreements that set the framework for data sharing within the consortium and the actual data sharing behaviour in specific contexts:

- Server-based – this involve data disseminated by a centrally managed service, the ICAT Information System. The system server will be responsible for enforcing data policies before data is disseminated.
- Peer to peer – data will be shared outside of a centrally managed but in a networked environment. End-point client applications will be responsible for enforcing data policies on the distributed data.
- Off-line – disseminated data will be used on a standalone machine without network connection. The properties of the data policies will determine if and how the data may be used.

For the first two contexts, we provide mini use cases to highlight specific data sharing requirements from different usage perspectives. Figure 6 gives a schematic representation of the application architecture for the data sharing use cases.

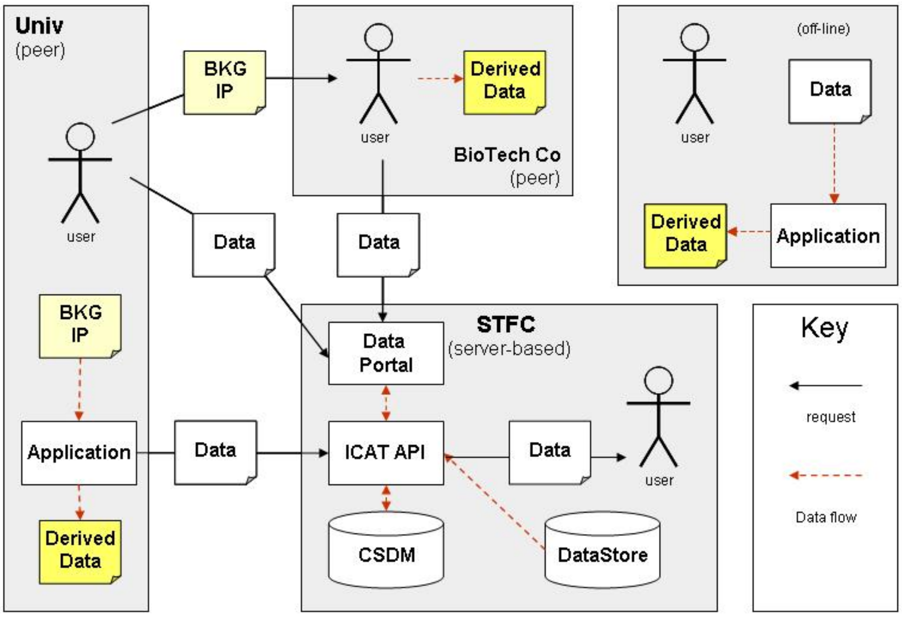


Figure 6. General Architecture for Data Sharing.

2.5.1. Data Sharing Agreements

In line with existing RCUK funding criteria, we envisage that the generic data sharing requirements between the project partners are captured in a collaboration agreement. This agreement constitutes the blue print for the collaboration and is normally negotiated and agreed at the proposal stage. It formalises in legal language the relationship between project participants and sets out their respective rights and obligations. It also contains crucial clauses on data sharing, including exemptions from obligations imposed by organisational level data sharing policies and government legislation. Participants may also agree separate, data sharing policies between themselves for specific research data or for new items arising from the research but which have not been prospectively assigned in the collaboration agreement. For example, a separate studentship agreement may be used to cover data relating to the student's thesis research work. It should be noted that in the academic domain, data policies are usually described in these high-level documents instead of self-contained, explicit data sharing agreements as used in the Crisis Management Test Bed. An analysis of some data sharing agreements indicates that they follow the board principles set out in the collaboration agreement but are more precise about:

- What data will be shared.
- The delivery/transmission mechanism.
- The processing and security framework.
- Disposal policies.
- Liabilities and sanctions.

Appendix 3 provides an analysis of the different agreements relevant to data sharing in research collaborations.

In the scenario, we assume that the organisation hosts' data sharing policies are specified independently of the collaboration. The participants are bound by default to these higher level policies through their contractual relationships to the host organisations, unless they have obtained an exemption from the policy owner. The collaboration, studentship agreement and data sharing policies should contain references to the pertinent parent policies and government legislation. The current version of the referenced policies as at the agreement date will be used. It should be noted that these agreements and policies are often written in the form of generic good practices or guidelines rather than prescriptive statements. For example, the NERC Data Policy Handbook [5] states that data from funded project '…must be offered to data centres after a reasonable time reserved for exclusive use…' It may be open to interpretation as to what constitutes 'reasonable'.

An analysis of sample UK collaboration agreements and a survey of the Research Councils' data policies (Appendix 4) indicate that the following legislation may apply to the scenario:

- Copyright, Designs and Patents Act 1988.
- Data Protection Act 1998.
- Freedom of Information Act 2000.
- Environmental Information Regulations 2004.

The relevance of the above legislation is summarised in Appendix 5. Other legislation may also apply to discipline-specific research involving restricted data. For instance, the use of identifiable patient data is subject to the Health and Social Care Act 2001 (England and Wales) plus approval from the appropriate research ethics committees and governing bodies like the Caldicott Guardian. Given the numerous permutation of research data and governing legislation, it is beyond the scope of this particular scenario to detail all their possible combination. We would emphasise that the data in scope for the scenario is not sensitive personal data but valuable data with sensitive academic and commercial IP. The Crisis

Management Test Bed scenario showcases an emergency with human life at risk. It will cover in greater details issues relating to sensitive personal information. Our test bed focuses on scientific research that gives rise to IP. In this context, an ability for the system to track derived data and provides a 'provenance trail' would be invaluable. This trail provides evidence for the thorough conduct of the research and demonstrates the quality of the results as well as ownership. These are important elements for patent application, journal review and research assessment processes.

2.5.2. Test Bed Use Cases

The chosen use cases cover the initial phase of establishing the collaboration when agreements are set up and the active research phases when the partners enact data sharing in different contexts.

2.5.2.1. Agreements Specification and Policy Administration

For the scenario, we envisage that administrators, such as a Research or Contract Manager, from the organisations hosting the researchers will be jointly responsible for drafting and maintaining the collaboration and related agreements. Although formal data sharing agreements as such are not commonly used in the academic environment, the data sharing clauses in the collaboration and confidential agreements as well as generic data sharing policies could be projected into controlled language DSAs for the purpose of formal policy analysis and refinement. Figure 7 shows use cases related to the specification and administration of these agreements. In line with current practice, the collaboration or similar agreements will be drafted using templates provided by the relevant organisations. The administrators acting as DSA Author, will build the data sharing clauses of the agreement using appropriate terms from a library of controlled vocabulary. It is essential that the controlled vocabulary can be used to define unambiguous data sharing conditions and obligations for specific dataset/s. The agreement definition process will also cover the identification and resolution of formal policy conflicts. This is to ensure that a consistent set of enforceable policies arising from this and other STFC data sharing agreements can be achieved for the same subject dataset/s. The DSA Author will not be experienced in formal policy language and will need tools to help them map a formal policy to the originating controlled language DSA clause and vice versa.

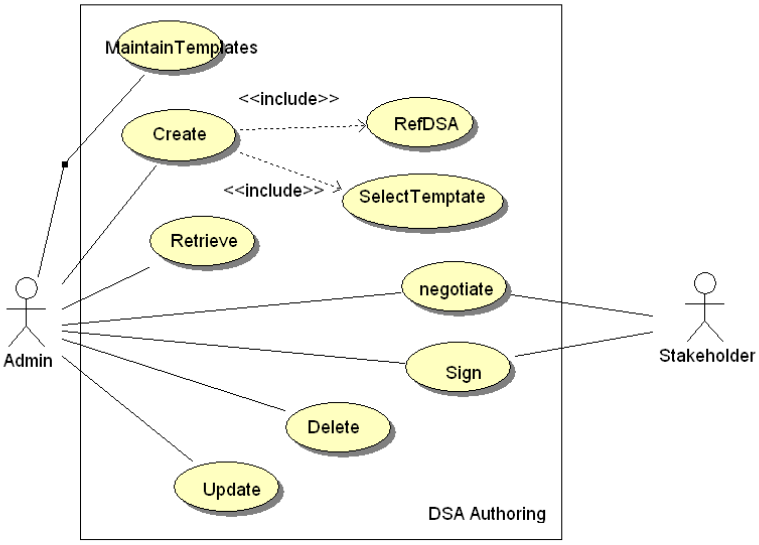


Figure 7. Agreement Definition Use Cases.

The completed draft will then be negotiated by stakeholders, which include the grant applicants, their host organisations, legal experts representing the stakeholders and, in some circumstances, the funding agency if it needs to authorise specific policy concession. As the collaboration/confidential agreement may be created at the proposal stage well ahead of the commencement of research activities, we envisage that the creation and deployment of enforceable low-level data policies will be triggered by a manual process. For other scenarios that do not involve a time lag between making the agreement and implementing the policies, creation and deployment of the enforceable data policies could be triggered automatically on completion of the signing process.

In view of the relatively long timescale of collaboration, it is expected that the initial set of policies refined from data sharing clauses in the different agreements will need to evolve in line with the project ecology. For instance, change of personnel, new data sharing requirements for emerging results not covered by existing agreements. It is envisaged that minor changes to the agreements which do not conflict with existing enforceable data policies will only require the consent of the participants but not a full-scale negotiation. After consent has been obtained, new low level policies will be generated and deployed.

It is expected that each organisation will have its own house rules and technical administrators for maintaining and managing its Consequence infrastructure, its portfolio of low level policies and related audit logs. The administrators will have the role of 'superuser' which include an ability to override low level policies within the remit of in-house policies. They will be provided with suitable tools to help them diagnose policy conflicts, trace policy provenance and 'debug' policies to validate policy properties and run-time enforcement behaviour.

To set the scene, Table 3 below lists the main data sharing policies relevant to the scenario story. It is envisaged that the controlled language DSAs will be able to unambiguously express data sharing requirements. If consent or external trigger is a policy property, it is further assumed that the agreement will provide sufficient information on how to obtain the appropriate information.

Policy (P)	**Agreement**	**Subject**	**Data Shared**	**Conditions (C)**	**Purpose**
1	Facility Data Policy	STFC Facility ↔ Facility users	1. Experimental Raw Data 2. Metadata	1. 3 years embargo from completion of experiment. 2. Investigators named on the beamline application can access their data and metadata during the embargo period. 3. No one can update or delete experimental raw data files, except the ICAT superuser. These files are contained in DataEHs with type set to 'expr_raw'. 4.An ICAT Administrator can insert, download and select authorised DataEHs managed by an ICAT. It can also update, logical and physical delete DataEHs that are new or it owns, as long as these Data EHs are not set as 'Facility Acquired'. This attribute denotes data and metadata imported from other systems. Experimental raw data is owned by the data distribution system. 5. Beamline Scientist is an ICAT Administrator of all experimental raw data generated by his/her station, but not other types of DataEH.	1-2. Provide data producers a time-limited period to exploit their data. 3-6. To protect data and data integrity. 5. To enable facility maintenance/ development.. 7-8. To incentivise data producers and to foster collaborative behaviour. 9. For usage and quality management.

				6. Actions in 4-5 can only be carried out on a machine within the Facility domain. 7. Researchers planning to use publicly available data in their analyses should, where possible, contact the original PI to suggest a collaboration. 8. Researchers who carried out analyses using publicly available data on ICAT are encouraged to link their published results back to the raw data on ICAT. 9. The facility would like researchers who have downloaded publicly available data to permit the periodic collection of usage information for data management purposes.	
2	Facility Data Policy	STFC Facility ↔ Facility Users	Refined Data	Refined data is limited at *all* time to users authorised by the data owner/administrator.	Protect IPR.
3	Funding Agency Data Sharing Policy	BioScience Funding Agency ↔ Funded Researchers	Research outputs arising from funded project. This includes all data that *can* be shared.	1. Data must be released in a timely fashion for sharing via deposition to an appropriate data bank or disseminated directly to others. 2. Data, where appropriate, should be accompanied by contextual information or metadata. 3. Licensing of IP generated from funded research should include a provision for research use by other BioScience supported scientists. 4. Any resulting articles that are published in journals or conference proceedings must be deposited at the earliest opportunity in an appropriate e-print repository. Subject to compliance with publisher's copyright and licensing policies.	1-4 to promote open access and maximise exploitation.
4	Data Sharing Schedule in Collaborat-ion Agreement	BioTech Company ↔ SB Partners (University A)	Proprietary data (background IP)	1. access limited to researcher partners from University A. They will be granted a royalty-free, non-exclusive and time-limited licence to use its background IP for the *purpose* of carrying out the project. 2. data can only be communicated via a secure network connection. . 3. an audit trail must be kept. 4. downloaded data must be deleted 30 days after last access. 5. user cannot update the data.	1-5. Protect IPR 5. Ensure data integrity
5	Data Sharing Schedule in Collaborat-ion Agreement	BioTech Company ↔ Consortium members	All Data (foreground IP)	1. data/academic paper using the background IP must: - acknowledge this work - can only be released with their explicit consent 2. all work related to the development of its drug discovery software in Phase 2 of the project must be: - carried out in its laboratory - using approved software only - supported by an audit trail. 3. No data, document should be disseminated without its explicit consent.	1-3 Protect IPR
6	IP Schedule in Collabor-ation Agreement	BioTech Company ↔ Consortium members	All Data (foreground IP)	Will exclusively own the IP in the results from the work related to its drug discovery software.	IP attribution
7	University IPR Policy	University B ↔ All employee	IPR	Retain IPR of works created by its employee (the Protein Crystallographer)	Default attribution of IPR.

8	IP Schedule in Collaboration Agreement	University A ↔ consortium	IPR	1. Concede IPR rising from the research carried out in Phase 2: - drug discovery software to the Biotech Company - characterisation of integrase to the CI and PI in the SB department. 2. Assign IPR arising from Phase 2 work that cannot be prospectively assigned to the respective parties as and when they are created. The parties will undertake to notify others when this occurs.	1-2 attribution of known and yet to be created IP.
9	Data Sharing Schedule in Collaboration Agreement	SB Researchers (University A) ↔ Consortium partners	IPR	1. Access limited to researchers working on the project. They will be granted a royalty-free, non-exclusive and time-limited licence to use its background IP for the *purpose* of carrying out the Phase 1 of the project 2. data/academic paper using this background IP must: - acknowledge this work - can only be released with their explicit agreement	1-2 Protect IPR
10	Data Sharing Schedule in Collaboration Agreement	Protein Crystallographer (University B) ↔ Consortium partners	Data, other research results	1. Data contributed to the project must only be used by members of the project and for the *purpose* of this project 2. data/academic paper using data and results arising from work carried out for this project must: - provide proper acknowledgement - can only be released subject to consent.	1-2 Protect IPR
11	Data Sharing Section in Studentship Agreement	Ph.D. Student ↔ Consortium	All his data.	Nobody should access his private data without his permission until the data or the thesis is published, whichever is the earliest.	Protect the IP potential of his work and avoid compromising background IP.

Table 3. Summary of Data Sharing Policies.

2.5.2.2. Server-based Data sharing

The following use cases consider the generic requirements for sharing data on a centrally managed data server, such as the STFC Facility ICAT Information System. We envisage that the various data sharing agreements have already been refined into local enforceable policies, deployed and ready for enforcement. In these use cases, we will consider requirements from the perspective of *both* the data users and the data service providers.

In the academic research domain, community data banks are custodian of research asset. Their data sharing policies will be influenced by the funding agencies and community best practices. A community data bank's holding is normally divided into private and publicly released data. Each data bank will have its own policies towards releasing private data to the public domain. RCUK suggests a guideline of three years within generation for data from publicly funded research. During the embargo period, access and usage of private data will be regulated according to the user and facility-defined access policies. These restrictions will be lifted automatically at the end of the embargo period unless an exemption exists. For the scenario, we assume that these access and public release policies originate from pre-defined data sharing agreements. Although there is no concern about access once the data is made public, a data bank may still wish to monitor data usage for management purposes. For example, to protect the integrity of data disseminated outside the server against accidental or deliberate modifications. Or to monitor usage for reporting or resource planning purposes.

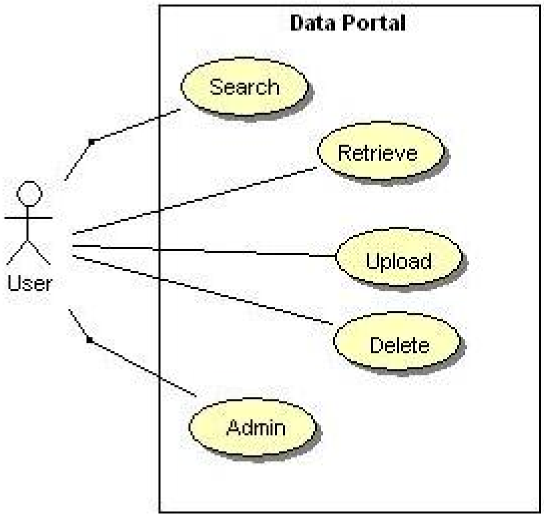


Figure 8. Data Portal Use Case

We present here a specific use case (**mini use case 1.0**) on sharing data on ICAT which is governed by the facility data sharing agreement with the facility and facility users as end parties. The Ph.D. Student visits STFC during Phase 1 of the project to carry out diffraction experiments at the X-ray Facility. After a few runs, he refines the raw data to try and solve the crystal structures using the CCP4 suite of software at the beamline. The results do not give him a clear view of how to proceed with the experiments in order to achieve optimal results from his limited beam time allocation. He seeks advice from the SB CI back at University A, who agrees to review the data. The CI tries to download both the experimental raw and refined data sets using Data Portal. (Figure 8 shows the typical use cases for Data Portal.) As a named CI on the beamline application, he is authorised by the facility data sharing policy to access the experimental raw data but not the refined one. The Student remedies this by giving his consent and enables the colleague to download the refined data set from ICAT. It should be mentioned that, in line with the data hub model described in Section 2.1, this use case focuses mainly on how the facility data sharing policy controls data sharing between end parties, which are the facility as the data hub/provider and facility users as data owners/users. As described in Table 3, the Student also has a separate policy in the Studentship Agreement unbeknown to ICAT which denies unauthorised access to his private data. If ICAT has not prohibited the download in the first place, it is envisaged that the student's own data sharing policy will be enforced on the client side as the CI is an end party to this second agreement. Client side policy enforcement will be discussed in more details in the next section.

Figure 9 shows a user downloading a requested data object from the Data Portal. The authentication process is omitted for clarity purpose.

1. User requests data object from Data Portal.
2. Data Portal passes on request to ICAT.
3. ICAT after checking the user permission against the request attributes, resolves the physical location and returns the metadata to Data Portal.
4. Data Portal renders a hotlink on the web page and presents this to the user.
5. The user clicks on the hotlink to download the object. Data Portal re-directs the request to the http Server which streams the object back to a servlet in Data Portal.

6. The user is prompted to display or save the object.

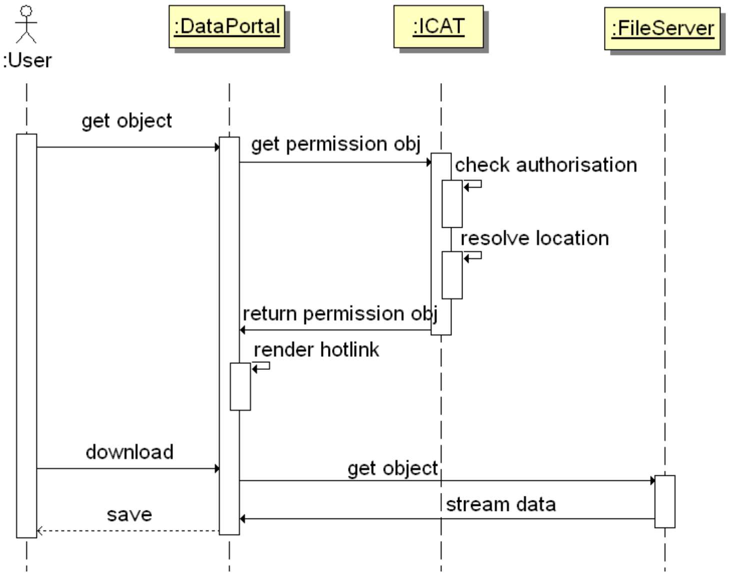


Figure 9. **Download a File on Data Portal**

In the above mini use case, the SB CI is permitted to view but not modify the experimental raw data as mandated by the facility data sharing policy. We introduce here an alternative use case (**mini use case 1.1**) whereby usage control may be waived on *publicly released* data only. A community data bank is the custodian of data contributed by stakeholders. It is entrusted with the task to curate, maximise the dissemination and exploitation of this data. Under this remit, a data bank may need a mechanism to waive usage control on unrestricted data. For example, STFC has data sharing agreements with other Research Councils to manage and disseminate raw data generated by their funded researchers on our facilities. Although STFC would prefer to protect disseminated data, this concern may be superseded by a business need to maximise public dissemination. If a user refuses to download the relevant policy component with the public data, ICAT may simply log the decision but waive the requirement rather than refusing the download request. This alternative use case shows that the policy to protect data integrity is not enforced on publicly released data. During the initial implementation phase when the Consequence framework is not yet pervasive, the scenario may result in different versions of the same data exist outside ICAT. However, we envisage these unlicensed versions will be less ‘trustworthy’ than an authenticated version protected by metadata. Meticulous scientists who are conscientious in maintaining quality research will be motivated to use only properly provenanced data in their work. Thus, we envisage that through a process of natural selection and the incremental adoption of the Consequence framework by the academic community, the need for a waiver mechanism would be phased out.

We also present an extended use case (**mini use case 1.2**) on server-based policy enforcement. ICAT is compatible with other information system implementing the CSMD and the ICAT API (see Figure 10). In this use case, the Student queries an integrated ICAT Information System for all data on a particular molecule. He will not be required to separately authenticate against each ICAT in the cross-search. Each participating ICAT will have its own set of facility and user policies which will need to be evaluated and enforced. We assume that the ICAT servers will be able to collaborate in resolving and enforcing the full set of data sharing policies including any side effects arising from the interactions of the separate sets of policies.

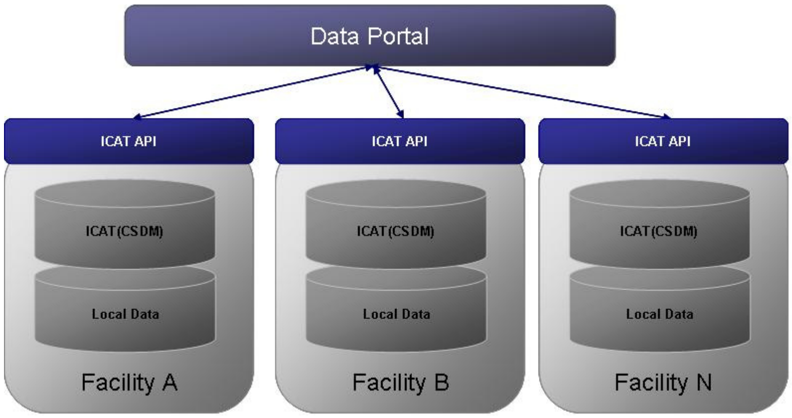


Figure 10. An Integrated ICAT Information System.

2.5.2.3. Peer-to-Peer Data Sharing

In cross-domain collaboration, partners may share data on-line using email, sFTP/FTPs or accessing data programmatically using server side script, wrapper or resource broker middleware. This use case (**mini use case 2.0**) considers client-side enforcement of data policies defined in the relevant project-specific agreements, which the partners are end parties to.

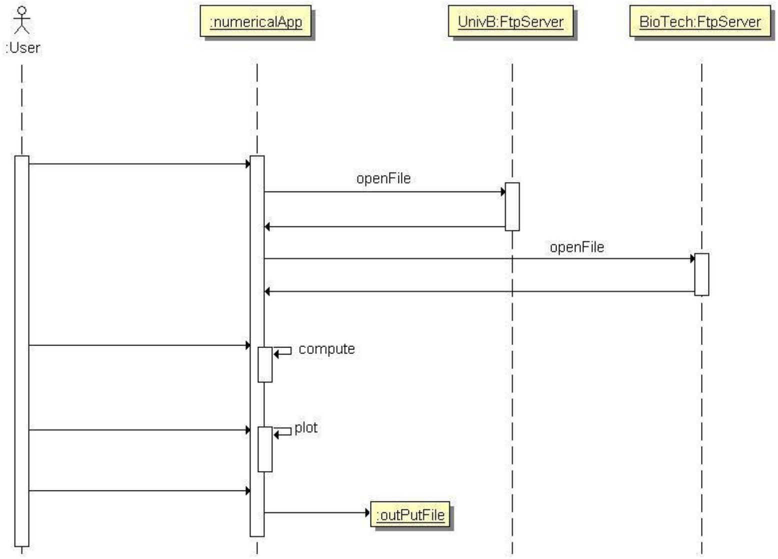


Figure 11. Sharing Data over FTP.

Our student, now back at the University, wants to explore the relationship between temperature and cell dimension in a protein structure. University B and BioTech have suitable data that he could use. The Student uses a numerical application to access the data via sFTP, performs some transformations and saves the results in a new file (Figure 11). This use case touches on the use of background IP which have various conditional access and obligation policies attached. For example, the data can only be used by partners for the *purpose* of the project, must be transferred over a secure network, and an audit trail must be kept on its usage (see Table 3). In particular, the creation of derived data brings out issues relating to the propagation of the parent data policies. For example, the derived data cannot be published without the explicit consent nor proper acknowledgement of the data providers, i.e. BioTech and University B.

It should be emphasised here that given the diversity of scientific research, it is not feasible to quantify all the software scientists use in research as they are as likely to write their own algorithms in scientific programming languages. These custom applications typically have a procedural design and perform some operations while iterating through a dataset. Ideally, the framework should have a platform independent architecture and support security over an application session rather than a single access. In addition, it may also be helpful to provide software libraries or plug-ins for popular scientific software or programming languages to facilitate the development of Consequence-aware scientific applications to widen adoption of the framework.

As an extension to this use case, we may consider the Student attempting to access partners' data in an automated, grid-based scientific workflow (Figure 12) to simulate the molecular structure of integrase in complex with a host protein (**mini use case 2.1**). This scenario would involve the Student delegating a form of credential to the workflow system. The workflow resource broker or a sub-process then uses the delegated credential to access and transfer proprietary data, possibly via an intermediate staging server, to third party computation nodes allocated at runtime. The physical usage contexts (e.g. time, geographical location) would evolve dynamically over the workflow enactment, the proposed solution must therefore be capable of monitoring environmental parameters to ensure the correct enforcement of context-related low-level data policies at read time.

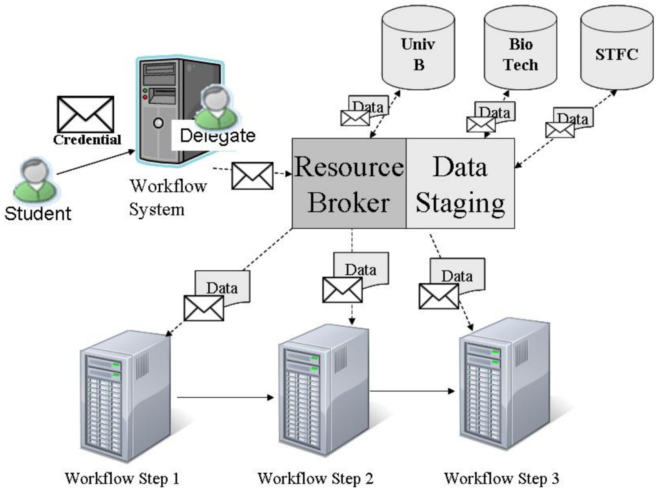


Figure 12. A Schematic Example of Data Sharing in a Scientific Workflow.

2.5.2.4. Off-line Data Sharing

Research is a creative activity and it is common research culture for researchers to carry out their activities when and where they feel appropriate. To accommodate this practice, it is desirable to permit occasional secure data sharing in an environment without network access for some data sharing activities, i.e. if not expressly forbidden by the properties in the low-level data policies. For example, the policy may only permit data to be accessed in a networked environment. The objectives of this extreme use case are to test policy properties and to stretch the capabilities of the proposed solution. It is assumed that the end-user's computer will have the appropriate Consequence policy components installed.

In this use case, our student is working from the student hostel and the broadband is broken. The PI has reviewed the digital protein model that he derived from the experimental results. The model and comments are stored on a physical medium. Our Student is keen to access the information as the model is a major output of his research. We envisage that he uses his desktop on a standalone mode to access the protected files. These digital documents will be protected by security policies and metadata or licence. The security policies will be evaluated against the security metadata or licence and enforced where appropriate. The Student will access the file and his actions are logged locally. When the network communication is resumed, we further envisage that an audit will be triggered. The policy enforcement component will review the local log to determine if breaches in pertinent data sharing policies have taken place and raise events as obligated by the policies. The log may also be used to support the resolution of conflicts over liability if a breach is detected.

3. Requirements Arising from the Scenario

Three main categories of test bed requirements: business, technical and administrative, are outlined below. Appendix 2 provides a mapping between these requirements and the mini use cases described in Section 2.5.2.

3.1. Business Requirements

BR1. A suitable agreement must be in force before research activities commence. The agreement, as described in Section 2.5.1, could be a Data Sharing Agreement, Collaboration Agreement, Confidential Agreement, Studentship Agreement, or Data Sharing Policy, etc.

BR2. Under the hub and spoke model, a single party may have multiple agreements with different end parties covering the same dataset. (Subject to all related DSA consistency check.)

BR3. The agreement should contain references to all pertinent external agreements, legislation and government law/s.

BR4. Template agreements with placeholders for the insertion of controlled vocabularies to express data sharing policies shall be used to facilitate the processing of the formal data sharing policies.

BR5. The controlled vocabularies must be understandable to human administrators or legal experts who are responsible for preparing and negotiating the agreement.

BR6. The controlled vocabularies must be suitable for the purpose of expressing formal, machine readable data sharing agreements.

BR7. It must be possible to identify the party/parties, data/datasets covered by the formal language DSA, although the description may be identity- or attribute- based.

BR8. The formal language DSAs should clearly define the conditions of use and obligations for the target data/datasets.

BR9. It should be possible to trace the provenance of a formal DSA policy back to the formal language DSA that gives rise to it.

BR10. It should be possible to propagate data sharing restrictions from a source dataset to the derived data.

BR11. It should be possible to identify the source of an inherited enforceable policy of a derived dataset.

BR12. If the access condition requires consent from a relevant party, the agreement should contain precise instructions on the procedure for getting consent.

BR13. If the access condition depends on an external trigger event that is not environmental (e.g. publication of the data) in nature, the agreement should state how the information can be obtained and who is responsible for monitoring the event. It may be that the data owner must inform the DSA administrator when the event occurs, who will then update the system.

BR14. The DSA authoring tool should provide human readable error messages to facilitate the diagnosis of conflicts or errors arising from the refinement of formal language DSAs into machine readable data sharing policies.

BR15. The formal data sharing policies applicable to a particular dataset must be unambiguously resolved between the different agreements with various end parties, such as facility users, projects, host organisations, funding bodies and third party services.

BR16. An agreement may be updated by the appointed owner if the amendment does not conflict with other current machine readable data sharing policies applicable to the target dataset/s. All parties must be informed and consent to the amendments.

BR17. An agreement must state the retention period and disposal process of the audit trail.

BR18. The researchers will retain administrator rights to machines under their management.

BR19. It is desirable for the solution to provide libraries for popular scientific programming languages and programming environments or software plug-ins to facilitate the development of Consequence-aware applications for the scientific domain.

BR20. It is desirable for the solution to support different application levels to cater to different security models. For instance, to allow optional usage control in a low risk, low security system as described in the use case for the public dissemination of public data by a data bank (Section 2.5.2); to the mandatory use of the full Consequence infrastructure and Consequence-aware applications in a high risk, high security system.

BR21. Each organisation host must put in place a clear organisational and reporting structure for the management and support of the agreements and the technical infrastructure.

BR22. Each organisation must establish business policies and procedures in line with domain best practices for the management, administration and security of agreements, data sharing policies and audit logs.

BR23. The proposed framework should not bring about a detriment of the RCUK goal to achieve open access to publicly funded research information. The proposed solution should be reliable, secure and easy to operate; thereby facilitates rather than hinder the timely release of controlled information and helps to encourage the early deposition of sensitive scientific information into managed repositories for long-term curation.

BR24. The solution should demonstrate cost-benefit effectiveness to community data providers such as the STFC large facilities.

3.2. Technical Requirements

TR1. The proposed architecture must be compatible with the STFC data management framework.

TR2. The data sharing policies must be compatible with the enhanced CSMD (see section 3.4).

TR3. The proposed architecture should be platform independent.

TR4. The proposed architecture should strive to support interoperation with a variety of software platforms including .NET, JAVA, Python and supportable on different flavours of operating systems including LINUX, Windows^{TM}, UNIX.

TR5. The proposed architecture should provide well-defined interfaces that specify the operations and message formats, supported protocol bindings to facilitate integration into existing application architecture such as the STFC ICAT Information System which uses SOAP.

TR6. The solution should provide a mechanism for analysing and resolving policy conflicts.

TR7. The solution should provide a mechanism to trace the provenance of a low level data sharing policy back to the originating controlled vocabulary DSA/s. Or, for an inherited policy, the parent low level data sharing policy.

TR8. The proposed solution should support session. Many scientific algorithms are procedural in design and require iterative read and write access to a dataset during an application session.

TR9. The data sharing policies are enforceable over varying types and volumes of scientific data generated throughout the science lifecycle.

TR10. The proposed policy infrastructure has the capability to actively monitor on-going environmental parameters to support context-aware data usage.

TR11. The data sharing policies are enforceable not just over data held centrally by the facility ICAT, but also when the protected data is disseminated and analysed on third party locations.

TR12. An audit trail shall be available to support the resolution of conflicts and liability if a breach in security occurs.

TR13. The solution should provide both manual and automatic low-level policy deployment mechanisms. Automatic deployment is desirable as this will minimise the security gap between the parties signing/consenting the agreement/update and the enforcement of the low level data policies. This is particular important in situations where a data sharing policy cannot be post-dated, for example, to revoke access due to personnel changes.

TR14. It is useful for the solution to provide an efficient and reliable mechanism to obtain consent if this is an access condition.

TR15. It is desirable for the solution to support secure data sharing in an environment where network connection is not always available.

3.3. Administration Requirements

AR1. A data file will be the smallest unit of a *data object* being shared. This does not apply to metadata.

AR2. Each organisation shall have a designated group of DSA administrators empowered to implement, update and override enforceable policies according to pre-defined organisational best practices.

AR3. The data sharing policy administrators will be a part of the organisation structure defined in BR13.

AR4. The organisational best practices could be represented by an organisational level data sharing policy referenced by individual agreements.

AR5. When a collaboration/data sharing agreement refers to a third party agreement/policy, it will adopt the version current on the date of the agreement.

3.4. CSMD Requirements

The CSMD model in its current format provides a single entity, called Access Conditions, for describing the right conditions under which access and usage of the scientific data is granted. Such information constitutes a form of *security metadata*, i.e. metadata that can be understood by access and usage enforcement mechanisms when evaluating security policies.

Naturally, there is a relationship between the security (and possibly the scientific) metadata of some data, the class of security policies used to express access/usage of that data, and the enforcement mechanism needed to enforce the policies, as shown in Figure 13.

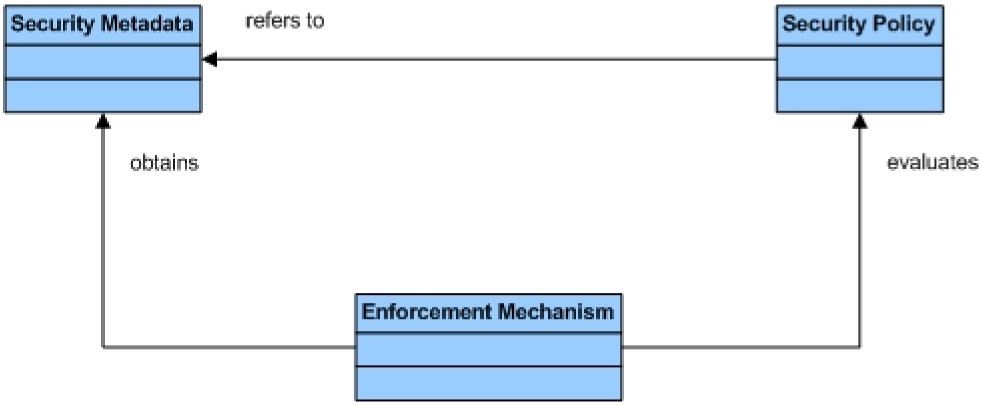


Figure 13. Security Metadata, Policies and Enforcement Mechanisms.

Security policies usually *refer to* certain attributes of the data and users requesting access to that data. The type of attributes required by a policy depends on the class of security models to which that policy belongs. For example, policies expressed in the Mandatory Access Control (MAC) model (e.g. [9]) require a data object to have an attribute representing its security classification, which is normally derived from a lattice [10].

Similarly, a Role-Based Access Control (RBAC) [11] policy would require the data object to have an attribute representing the set of rights permitted on that object, where those rights are assigned to roles, which in turn are assigned to users.

The enforcement mechanisms *evaluate* and enforce the security policies, and in order to achieve that, they may need to *obtain* some of the object's attributes. In our case, object attributes are captured by the security metadata information attached to the scientific data.

The current Access Conditions entity does not provide much information on the data security. Therefore, we suggest defining a new security metadata model, which will take into consideration the following expressivity requirements:

MD1. The model must be able to express data classification information, such as security levels and classes of conflict.

MD2. The model must be able to express information related to the permissible contexts in which the data may be used or accessed.

MD3. The model must be able to express information on the permissible domains and networks routes that the data is allowed to exist in.

MD4. The model must allow cryptographic information on the data, such as digital signatures and hash functions, to be expressed.

MD5. The model must be able to express data freshness.

MD6. The model must be able to allow for log information on the access, usage and routing of data to be included.

MD7. The model should include provenance information.

MD8. The model must include redundancy information about the data. This is particularly important to preserve the data integrity.

MD9. The model must include information about the permissible operations (access, usage and routing) that can be applied to the data.

The new model, which represents a new taxonomy of data security, will be developed as part of Consequence D3.1, and it will be used in combination with the CSMD model when generating and managing the metadata attached to data.

4. Risk Assessment and Threat Analysis

The purpose of DSA management in the Sharing Sensitive Scientific Data test bed is to ensure that all those who should be able to use the data for the stated purpose can do so, and that those who should not have use of the data are prevented from doing so.

The highest level risks are that:

1. It is unknown whether an Agent should be permitted to use data or not.
2. Agents who should have usage cannot use the data.
3. Agents who should not have access are permitted usage.

The first risk gives rise to the other two, so it can be judged as a secondary, causal, risk, but it is treated at this highest level since it has a different organisational status in the test bed environment.

At present, for the test bed environment, it is unclear what data sharing policies apply to any data set, because no common analysis exists of all agreements between STFC and all actors which include policies on data sharing. A major requirement on the Consequence system for this test bed is to include an analysis of policies deriving from different agreements to identify conflicts, and address the first risk, so that it is clear whether an agent should be able to use the data or not.

Once policies are consistent and in control of data access/usage the second two risks become significant. This stage in the process can be broken into three secondary risks given the organisational deployment environment:

1. Authentication of agent identity.
2. Authorisation of server side access controls.
3. Enforcement of client side usage controls – on and off-line.

The test bed environment enforces policies for issuing agent identity credentials that comply with the International Grid Trust Federation (IGTF) identity vetting policies [15,16] in order to manage the risks on issuing credentials, leaving the first authentication risk above to be addressed in the enforcement environment by Consequence.

The risks associated with authorisation of server side access controls are classic risks of failure to enforce the set of policies that apply to a dataset. There are no novel detailed risks from the test bed in this area.

The client side usage controls use cryptography and fine-grained policies to limit what a user can do with data using an information-centric security approach. The consumer counterparts of this technology have suffered one successful attack after another, demonstrating that they only provide a low-surety solution. The threat model for consumer Digital Rights Management (DRM) includes:

1. Single expert developing a tool to make one copy of the data.
2. One infringing data copy propagating across the world.
3. Copying tool propagating across the world.
4. Casual-copying – prevent a majority of users from copying data.

Most consumer DRM solutions (e.g. Apple iTunes, DVD encryption) address the fourth threat, but while the encryption can be removed from the data with an effort that justifies doing so for one piece of data, they do not address the first three threats. Once a decryption tool has been developed it can quickly propagate as freeware (e.g. JHymn), and be used by a significant proportion of users until the decryption method changes, thereby nullifying the DRM solution. In the sensitive scientific data scenario the same threat models exist. Solutions based on "trusted computing" platforms which require specialised hardware would not be acceptable to the user community. Therefore it is necessary to accept the risks posed by the threat model and manage them. The balance of the argument for consumer DRM is whether people who illegally copy music would buy it anyway, so the copying is a loss of income. This trade-off does not apply to the scientific data use case, or Enterprise-DRM in general. In the scientific data test bed the loss of revenue may come a long way down stream through a competitor beating a company to the market, or a patent on a new drug, or a scientist being beaten to the discovery of a new planet[1].

The risk assessment must be in terms of the effort required to design a decryption tool for one piece of data compared to the benefits to be gained. Since some of the drug discovery related data considered in the scenario could be worth millions of euros, it must be assumed that the motivation exists to design a decryption tool, and the research community is certainly one through which open source software propagates very quickly, so once developed such a tool would quickly become widely used.

For the identified risks, it is necessary to establish where vulnerabilities exist within the Web Service system which could result in incidents. The following Vulnerability-Incident life-cycle model provides illustration how vulnerability may become a potential security threat and further develop to an Incident [12, 13]:

Vulnerability => Exploit => Threat => Attack/Intrusion => Incident

Vulnerability is a flaw or weakness in a system's design, implementation, or operation and management that could be exploited to violate the system's security policy.

Exploit is a known way to take advantage of a specific software vulnerability.

Threat is a potential for violation of security, which exists when there is a circumstance, capability, action, or event that could breach security and cause harm.

[1] In July 2005 a meeting abstract posted online stated that a team of scientists lead by Mike Brown of Caltech believed that they had discovered a new planet from images taken in 2003. A second group of scientists led by Jose-Luis Ortiz of the Sierra Nevada Observatory in Spain identified the planet in the same data available to the community, and claimed the discovery of Haumea for themselves on 28th July 2005, having accessed the on-line observing logs of Brown's team two days earlier - no evidence has surfaced showing that Ortiz's team found the object using Brown's observing logs.

Attack is an assault on system security that derives from an intelligent threat.

Incident is a result of successful Attack, and the risk has proven to occur.

The main classes of attack identified for web service based data services, utilising vulnerabilities in web service design are summarized below [14]:

1. *User Credentials Attacks* (UCA) are attacks originating from user credential theft or compromise.
2. *"Wire" Intelligence Attacks* (WIA) include a wide spectrum of attacks that can happen if service-level communication is not protected enough against eavesdropping and interception. Most threats in WIA group come from potentially uncontrolled environment messages may pass, especially if end-to-end service communication involves SOAP messages routing and intermediate processing. Communication and messages compromise and manipulation may lead to such classes of attacks as "Man in the middle" (MITM), credentials compromise and/or replay, session hijack, SOAP routing detour, and as well as attributes/credentials probing and brute force attacks.
3. *Malefactor Initiated Attacks* (MIA). This group of attacks can be undertaken by a potential attacker using both traditional and Web Services specific techniques that include WSDL probing, malicious XML content, brute force and dictionary attacks to by-pass site security services, and traditional Denial of Service (DoS) attacks that may target all components of the site services stack. It is even more difficult to avoid this type of attacks against Web Services because traditional network and host protection tools, like Firewalls, are transparent to SOAP communications.
4. *Site Management Attacks* (SMA) include possible attacks that can be caused by improper site security services configuration and management: insufficient AuthN and AuthZ credentials verification including security context verification, improper key and privileges management and control, improper error handling that may disclose internal information about service operation, and also insufficient or insecure logging that may allow an attacker to hide or forge its activity.
5. *End Service Attacks* (ESA) target known vulnerabilities in the end-service. They use different techniques to construct malicious input content, e.g. XML/SQL injection, external references in XML schema and XML documents, internal and external cross-references with XPath and XSLT instructions. Attacker may intend to violate suggested quota or acceptable use of the resource what may be prevented by proper access control and accounting. End service application can be a target and a mediator of viruses and worms carried over some types of unchecked input, and therefore antivirus protection should also be considered for Web Services applications.

There is a requirement on the design of the Consequence system that it does not include vulnerabilities, or provides mechanisms to manage attacks on those vulnerabilities to prevent successful attacks, that remain threats from these identified classes of attack.

5. Evaluation

5.1. Provisional Plan for the Demonstration

We describe a provisional plan for the review demonstration. We propose a set of four tasks to demonstrate selected mini use cases as described in Section 2.5. The plan will be updated and fleshed out in future iterations of the requirements to reflect development of the Consequence architecture and the test bed.

5.1.1. Specific demonstration requirements

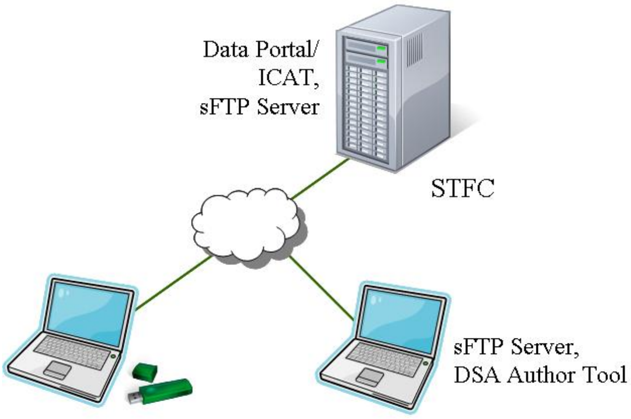


Figure 15. Proposed Setup for the Demonstration.

It is desirable to have a wireless network connection that permits remote access to the Data Portal and ICAT test bed in STFC Daresbury Laboratory (Figure 15). We will use two laptops that have the appropriate policy component/s installed to demonstrate the tasks outlined in section 5.2. If it is not feasible to have an internet connection, we will use a third laptop as the Data Portal/ICAT and sFTP server.

5.1.1.1. Agreement Specification

This task maps to the use case described in Section 2.5.2.1. A DSA Administrator will use the DSA Authoring Tool on one laptop to complete a draft collaboration agreement. He will use the tool to retrieve the agreement from the connected repository and add a new data sharing clause in the IP Section. The clause specifies usage conditions on data derived from the use of BioTech background IP. The usage conditions include context and purpose awareness as described in Table 3 Policy 4. He will build this clause by selecting the correct fragment from a pre-defined list of controlled vocabulary terms available to the tool. He will then use the tool to carry out DSA analysis and map the data sharing clauses in the agreement to a set of formal policies. The analysis will check these policies for consistency and conflicts.

This task is designed to show case the functionalities of the DSA authoring tool in building agreements which can then be used to auto-generate enforceable policies applicable to data sharing in the scientific research domain.

5.1.1.2. Secure Data Sharing in a Server-controlled Environment

This task maps to the first mini use case (1.0) described in Section 2.5.2.2. We will log on via https to the test bed as the Co-Investigator to download a restricted experimental raw dataset and a refined dataset. The Student has a policy on the latter dataset to deny unauthorised access and he will need to amend this policy or give his consent to permit access by the CI. (It is not clear which is the optimal approach at this early stage of specifying the framework architecture.)

This task is designed to demonstrate the correct enforcement of facility data sharing policies, the ability to make and implement dynamic changes to an enforcement data policy on ICAT, the data server.

5.1.1.3. Peer-to-Peer Data Sharing

This task maps to the first mini use case (2.0) described in Section 2.5.2.3. The Student actor will use an EXCEL spreadsheet on the laptop to access restricted datasets via sFTP over wireless connection to two sFTP servers representing University B and BioTech. The Student will use EXCEL function to transform the data and plot the results. He will then save the results in a new file. We could then inspect the security metadata and low level data policy/ies attached to the new file. We will look for evidence that the derived dataset has inherited the appropriate data sharing policies from the source datasets. These include requirements to maintain an audit trail and to obtain consent from the data owners before the derived data can be published.

This task is designed to demonstrate endpoint, client-side enforcement of controlled datasets at read time using a Consequence-aware application. The task will also test different policy properties which include usage, context and purpose awareness as described in Table 3 Policy 4 and P10. With regard to purpose awareness, we envisage that the policy enforcement mechanism is able to evaluate access purpose using attribute/s describing the Student's role in the project.

5.1.1.4. Off-line Data Sharing

This task maps to the use case described in Section 2.5.2.4. We will turn off the wireless switch on the laptop at this stage. The Student actor will attempt to access protected data on a physical storage media, use the data and save the file. Then we will turn the wireless back on the laptop. We envisage that this will trigger the policy component/s to connect and update with the central policy server. The local component/s should evaluate if any security breaches have taken place when the network connection was unavailable and perform the appropriate obligation tasks. We could also review the audit log to examine if the expected sequence of events were carried out.

The task is designed to demonstrate enforcement of usage policies associated with disseminated data in a non-networked environment. We envisage that the enforcement mechanism/s will evaluate relevant properties in the security metadata or licence against low level policy/ies on the data to determine if off-line access is permitted.

5.2. Evaluation Criteria

We are currently gathering evaluation criteria through interviews with potential Consequence end users and adopters. These include contract managers, research scientists and the owners and developers of the STFC large facility ICAT Information systems.

The feedback we have received so far from the facilities relate to the cost-benefits of implementing and supporting the Consequence framework and the possible impact on ICAT if users can specify their own data sharing policies. For example, will the scenario lead to an explosion of both textural and formal data sharing policies that require significant administration efforts? And will the framework require considerable development efforts to integrate and will it affect interoperability between the different ICAT and CSMD implementations by other facility data providers? Stakeholders also expressed concerns regarding system performance and usability when additional overheads for evaluating and enforcing policies as well as maintaining the trust environment are factored in.

The scientists raised issues regarding supports for specialised scientific software tool such as MatLAB, SAS and domain specific tools like CCP4, openGENIE, not to mention their own purpose-written algorithms. Will they only be able to use Consequence-aware applications for accessing controlled dataset/s and how much extra efforts it would involve to make their algorithms Consequence-aware? A few also expressed concern about the impact of usage control on current research culture. Researchers have considerable freedom on how and where they carry out their activities. University-based researchers often administer computers in their research unit. Will adopting the Consequence framework mean that they will no longer be able to do so as it may break some Consequence trust assumptions? For example, end users should not have access to the operating system or that all data access must be through Consequence-aware applications? And will it significantly change they way they do research? For instance, will the system be restricted to specific technology platform? Will they have to install and maintain different software components on their machines and only access controlled data in a networked environment? And will there be significant bureaucracy associated with using the Consequence framework?

Perhaps the most crucial aspect towards measuring success is what benefits will Consequence confer over and above the efforts required? For example, from RCUK's point of view, will Consequence encourage funded researchers to entrust their research data to a managed community repository early in the research lifecycle thereby minimising the risk of data lost? For the data banks, will Consequence improve security in protecting their private data holding while ensuring the integrity of disseminated data? For the research managers, could the framework facilitate reporting and resource planning? For example, through obligation policies to collect usage statistics to assist resource planning and to track derived data to validate research output and quality? From the data producers' perspective, will the security model be sufficiently robust to reassure industrial users to let the facility manage their raw data and to encourage industry to leverage existing government investment in grid computing infrastructure and increase productivity? For the generic researchers, will the Consequence framework be pervasive but non-invasive? Will they have to sacrifice a certain degree of academic freedom to satisfy the Consequence trust assumptions? And will it facilitate rather than hinder collaborative research activities?

With these different concerns in mind, we may measure the success of the Consequence framework along the following lines:

1. Will the framework deliver its objectives given its trust assumption?
2. Will the framework be effective given the prevalent research culture and practice?
3. Will the framework be usable, easy to implement and maintain?
4. Will the framework add values? Both in terms of physical performance and business benefits.

We propose to obtain community feedback to the above questions through conducting a series of interviews with selected representatives from the pertinent end user communities. We will make available a test bed implementation of a Consequence-aware ICAT information system and a DSA Authoring Tool to appropriate users and obtain feedback on their usability, performance and business fits. We will also carry out internal technical assessments to evaluate the performance of the framework implementation using purpose written test cases to test the refinement and enforcement of low level policies derived from sample textural agreements in data sharing.

6. References

1. OECD Principles and Guidelines for Access to Research Data from Public Funding. OECD 2007. Available at http://www.oecd.org/document/55/0,3343,en_2649_34293_38500791_1_1_1_1,00.html
2. Research Funders' Policies for the Management of Information Outputs. RIN 2007. Available at http://www.rin.ac.uk/files/Funders'%20Policy%20&%20Practice%20-%20Final%20Report.pdf
3. BBSRC's Data Sharing Policy. http://www.bbsrc.ac.uk/publications/policy/data_sharing_policy.pdf
4. MRC Policy on Data Sharing and Preservation. Available at http://www.mrc.ac.uk/PolicyGuidance/EthicsAndGovernance/DataSharing/PolicyonDataSharingandPreservation/index.htm
5. NERC Data Policy Handbook V2.2 (December 2002). Available at http://badc.nerc.ac.uk/data/NERC_Handbookv2.2.pdf
6. Shoaib Sufi and Brian Matthews, eds. Kerstin Kleese van Dam, *The CLRC Scientific Metadata Model*, DL-TR-2002-001 (2001), CCLRC, 2001.
7. Shoaib Sufi, Brian Matthews and Kerstin Kleese van Dam, *An Interdisciplinary Model for the Representation of Scientific Studies and Associated Data Holdings*, In proceedings of the 2003 UK e-Science All Hands Meeting (AHM2003), Nottingham, UK, 2003.
8. Shoaib Sufi and Brian Matthews, *The CCLRC Scientific Metadata Mode: Version 2*, DL-TR-2004-001 (2004), CCLRC, 2004.
9. Bell, D.E. and L.J. La Padula, *Secure Computer Systems: Unified Exposition and Multics Interpretation*, Mitre Corporation, USA, 1975.
10. Dorothy Denning, *A Lattice Model of Secure Information Flow*, Toplas:19(5):236-243, 1976.
11. Ravi S. Sandhu, Edward J. Coyne, Hal L. Feinstein and Charles E. Youman, *Role-based Access Control Models*, Computer:29(20):38-47, 1996.
12. Shirey, R., "Internet Security Glossary", RFC2828. May 2000. Available at http://www.faqs.org/rfc/rfc2828.txt
13. Demchenko, Y., "Grid Security Incident definition and exchange format", EGEE MJRA3.4 Deliverable document. Available at https://edms.cern.ch/document/501422/
14. Demchenko, Y., "Web Services and Grid Security Vulnerabilities and Threats Analysis", EGEE JRA3 Technical document. Available at https://edms.cern.ch/document/632020/1
15. EUGridPMA, Policy on vetting identity in a face to face meeting. Available at https://www.eugridpma.org/guidelines/1scp/1SCP-vetting-f2f-0.2.pdf
16. EUGridPMA, Policy on vetting identity by a trusted third party. Available at https://www.eugridpma.org/guidelines/1scp/1SCP-vetting-ttp-0.1.pdf
17. EPSRC Guidance for Collaboration Agreements. Available at http://www.epsrc.ac.uk/Business/Guidance/CollaborationAgreements.htm
18. Lambert Tool Kit for Collaborative Research (2008). *The Lambert Working Group on Intellectual Property.* Available at http://www.innovation.gov.uk/lambertagreements/index.asp.
19. Northamptonshire Young People's Education & Learning Data Exchange Agreement. 2005. Available at https://www.northamptonshire.gov.uk/NR/rdonlyres/53AFDA54-

F981-4BB4-A6E3-F52CB18B1F88/0/DataExchangeAgreementConnexionsfinalversion.doc

20. The Russell Group Studentship Agreement Template. Available at http://www.innovation.gov.uk/lambertagreements/files/Russell_Studentship_Agreement.DOC
21. The Uniform Biological Material Transfer Agreement. Available at http://ott.od.nih.gov/NewPages/UBMTA.pdf
22. World Intellectual Property Organization Contracts Database. Available at http://www.wipo.int/tk/en/databases/contracts/
23. Freedom of Information and Intellectual Property Rights. *Dundas & Wilson.* 2004. Available at http://www.jisclegal.ac.uk/publications/foidundaswilsonipr.htm
24. Johnson, Claire, *Freedom of Information (Scotland) Act (FOISA) and Research Information.* University of Glasgow. 2004. Available at http://www.recordsmanagement.ed.ac.uk/InfoStaff/FOIstaff/SHEIPResearch/resissues9.pdf
25. Theo, Andrews, *Intellectual Property and Electronic Theses*. 2004. Available at http://www.jisclegal.ac.uk/publications/ethesesandrew.htm

Appendix 1. Glossary

ADS (http://ads.ahds.ac.uk/)	Archaeology Data Services.
AHRC (http://www.ahrc.ac.uk/Pages/default.aspx)	Arts and Humanities Research Council.
BBSRC (http://www.bbsrc.ac.uk/)	Biotechnology and Biological Sciences Research Council.
Background IP	IP contributed to the collaboration by the participants.
CCP4 (http://www.ccp4.ac.uk/main.html)	Collaborative Computational Project No. 4 - Software for Macromolecular X-ray Crystallography.
CPDA (http://www.opsi.gov.uk/acts/acts1988 /Ukpga_19880048_en_1.htm)	Copyright, Designs and Patents Act 1988.
CSMD	Common Scientific Metadata Model.
Cumulative research	Collaborative research that could be interdisciplinary in nature. Research that build on previous efforts and discoveries.
DFID (http://www.dfid.gov.uk/)	Department of International Development.
DH (http://www.dh.gov.uk/en/index.htm)	Department of Health.
DIUS (http://www.dius.gov.uk/)	Department for Innovation, Universities & Skills.
DLS (http://www.diamond.ac.uk/default.htm)	Diamond Light Source third generation synchrotron research facility.
DPA	Data Protection Act 1998.
DSA	Data Sharing Agreement.
EIR	Environmental Information Regulations 2004.
EPSRC (http://www.epsrc.ac.uk/default.htm)	Engineering and Physical Sciences Research Council.
ESRC (http://www.esrcsocietytoday.ac.uk/ ESRCInfoCentre/index.aspx)	Economic and Social Research Council.
feC	Full economic costs.
FOIA	Freedom of Information Act 2000.
Foreground IP	IP arising from a research collaboration.
HEI	Higher Education Institutions.
HPC	High Performance Computing.
ICAT (http://code.google.com/p/icatproject/)	Information Catalogue. ICAT is a suite of software tools which provides access to specific information within large distributed data collections, large both in terms of complexity and size (into the PetaByte range).
ICO (http://www.ico.gov.uk/)	Office of the Information Commissioner.

IGTF (http://www.igtf.net/)	International Grid Trust Federation.
IPR	Intellectual Property Right.
ISIS (http://www.isis.rl.ac.uk/)	STFC ISIS pulsed neutron and muon science facility.
Integrase (http://www.mrc.ac.uk/index.htm)	An enzyme produced by a retrovirus (including HIV) that enables its genetic material to be integrated into the DNA of the infected cell. It is also produced by viruses containing double stranded DNAs for the same purpose.
LAN	Local area network.
MRC (http://www.mrc.ac.uk/index.htm)	Medical Research Council.
NASA (http://www.nasa.gov/)	National Aeronautics and Space Administration.
NERC (http://www.nerc.ac.uk/)	Natural Environment Research Council.
NHS (http://www.nhs.uk/Pages/homepage.aspx)	UK National Health Services.
OECD (http://www.oecd.org/home/)	Organisation for Economic Co-operation and Development.
Open-GENIE (http://www.isis.rl.ac.uk/openGenie/)	A data access programme for analysing neutron scattering data.
Pre-competitive research	Fundamental research work that is not aimed at producing products, but at providing the tools, information, and data that enables others to develop future products and services. It is work that: • confers equal benefits to all competitors; • industry is willing to have fully published.
RCUK (http://www.rcuk.ac.uk/)	Research Councils UK. A strategic partnership of the seven UK research councils – AHRC, BBSRC, EPSRC, ESRC, MRC, NERC and STFC.
SB	Structural Biology. Studies of the structure and function of proteins.
SOAP	Simple Object Access Protocol.
SSH	Secure Shell.
STFC (http://www.stfc.ac.uk/)	Science and Technology Facilities Council.
Structure-based Drug Design	The design of novel compounds based on the three-dimensional structure of a protein, e.g. a receptor protein.
SuperJanet (http://www.ja.net/)	SuperJanet (Joint Academic NETwork) is a private British government-funded high-speed fibre optic computer network linking education and research institutions across the UK.
Synchrotron light	An electromagnetic radiation produced by bending magnets and insertion devices (undulators or wigglers) in storage rings and free electron lasers.
UKDA (http://www.data-archive.ac.uk/)	UK Data Archive.

UKPMC (http://ukpmc.ac.uk/)	UK PubMed Central.
VNC	Virtual Network Computing.

Appendix 2. Requirements Matrix

This table maps use cases or relevant sections of the test bed description to specific business, technical and administrative requirements from Section 3. The table includes some advance and desirable requirements which have lower priority than the main requirements. Metadata requirements are discussed separately in D3.1. The baseline expectation is that the Data Security taxonomy model should meet the requirement of the missing security metadata.

Requirement ID	Description	Use Case	Section
Main Requirements			
BR1	A valid agreement must be in place before start of research.	Agreement Specification	2.5.2.1; 5.1.1.1; 2.5.1
BR2	A single party may have multiple agreements with different parties on the same dataset.	Agreement Specification; Server-based Data Sharing mini use case 1.0	2.5.2.1 (Table 3); 2.5.2.2
BR15	The formal data sharing policies applicable to a particular dataset must be unambiguously resolved between the different agreements with different end parties.	Agreement Specification	2.5.2.1; 5.1.1.1
BR3; AR5	An agreement should include references to relevant external agreement/s and legislation. The current version as at the agreement date will be used.	Agreement Specification	2.5.2.1; 5.1.1.1; 2.5.1
BR4	Template agreement shall be used.	Agreement Specification	2.5.2.1; 5.1.1.1
BR5	The controlled vocabularies for building DSA clauses must be understandable to human users.	Agreement Specification	2.5.2.1; 5.1.1.1
BR6	The controlled vocabularies must be suitable for expressing policy for analysis and refinement purposes.	Agreement Specification	2.5.2.1; 5.1.1.1
TR6	The solution should include a mechanism for analyzing and resolving formal policy conflicts.	Agreement Specification	2.5.2.1; 5.1.1.1;
BR14	The mechanism referred to in TR6 should provide human readable error messages to facilitate policy analysis and refinement.	Agreement Specification	2.5.2.1; 5.1.1.1;
BR7	Formal language DSA may use identity or attribute-based description.	Agreement Specification	2.5.2.1; 5.1.1.1; 2.5.1
BR8	Formal language DSA should clearly define conditions of use and obligations for target dataset/s.	Agreement Specification	2.5.2.1; 5.1.1.1
BR9; TR7	Capability to resolve the provenance of a formal policy back to the high level DSA that gives rise to it.	Agreement Specification	2.5.2.1
BR10	Capability to propagate data sharing policies from parent to derived dataset/s.	Agreement Specification; Peer-to-Peer Data Sharing mini use case 2.0	2.5.2.1; 5.1.1.1; 2.5.2.3; 5.1.1.3
BR11, TR7	Capability to trace the origin of inherited policies in a derived dataset back to the parent dataset.	Agreement Specification; Peer-to-Peer Data Sharing mini use case 2.0	2.5.2.1; 5.1.1.1; 2.5.2.3; 5.1.1.3
	An agreement should contain precise instructions on:	Agreement Specification	2.5.2.1; 5.1.1.1

BR12 BR13	- the procedure to obtain consent if this is a policy condition; or - obtaining intelligence about external events that trigger changes in policy states.		
BR16	An agreement may be updated by the owner if the resultant formal policy does not conflict with other formal policies in force on the same dataset. The agreement end parties will be informed and their consent sought.	Agreement Specification	2.5.2.1
TR11	An audit trail available to support resolution of conflicts and liability.	Agreement Specification; Off-line Data Sharing Use Case	2.5.2.1 (Table 3); 2.5.2.4; 5.1.1.3
BR17	An agreement must state the retention period and disposal process of the resultant audit trails.	Agreement Specification	2.5.2.1
BR21;BR22; AR2; AR3; AR4	Each organization to establish clear organizational and reporting structure, guidelines and procedures for the administration and support of: - its portfolio of agreements and formal data sharing policies; - the related audit logs; - the technical infrastructure.	Agreement Specification	2.5.2.1
BR23	The proposed framework should not bring about a detriment of the RCUK goal to achieve open access to publicly funded research data.	Server-based Data Sharing mini use case 1.1	2.5.2.2; 1
BR24	Demonstrate cost-benefit effectiveness to community data providers.	Server-based Data Sharing mini use case 1.1	2.5.2.2;
TR1; TR2 TR5	The proposed architecture: - must be compatible with STFC data management framework including the Data Portal, ICAT3 and CSMD; - should provide well-defined interface to facilitate integration with existing application architecture.	Server-based Data Sharing mini use cases 1.0,1.1,1.2	2.5.2.2; 2.4
BR18	The researchers will retain administrator rights to machines under their management.		2.3.2
TR3	The proposed architecture should be platform independent.	Peer-to-Peer Data Sharing mini use case 2.0	2.5.2.3
TR4	The proposed architecture could interoperate with a variety of software platforms.	Server-based Data Sharing mini use cases 1.0,1.1,1.2	2.5.2.2; 2.4.1
TR8	The solution should support the notion of session.	Peer-to-Peer Data Sharing mini use case 2.0	2.5.2.3
TR9	The policies are enforceable over varying types and volumes of scientific data.		2.2
TR11	The data sharing policies are enforceable not just over data held centrally by the facility ICAT, but on disseminated data analysed in third party locations.	Server-based Data Sharing mini use case 1.0 Peer-to-Peer Data Sharing mini use case 2.1	2.5.2.2 2.5.2.3; 5.1.1.2
TR10	The policy infrastructure could actively monitor on-going environmental parameters to ensure the correct	Peer-to-Peer Data Sharing mini use case 2.1	2.5.2.3

	enforcement of context-aware data sharing policies.		
AR1	A data file will be the smallest unit of data objects being shared. This does not apply to metadata.		2.4.1
TR13	The solution could support both manual and automatic deployment of low level policies.	Agreement Specification	2.5.2.1
Advanced Requirement			
TR15	It is desirable for the solution to support secure data sharing in an environment where network connection is not always available.	Off-line Data Sharing Use Case	2.5.2.4; 5.1.1.3
Desirable Requirements			
TR14	The proposed solution may include an efficient and reliable mechanism to obtain consent if this is an access condition.	Server-based Data Sharing mini use case 1.0	2.5.2.2; 5.1.1.2
BR19	Provision of software libraries or plug-ins to promote the development of Consequence-aware scientific applications.	Peer-to-Peer Data Sharing mini use case 2.0	2.5.2.3; 2.3.2
BR20	Capability for the framework to support different application levels.	Server-based Data Sharing mini use case 1.1	2.5.2.2

Table A2. Requirements Matrix.

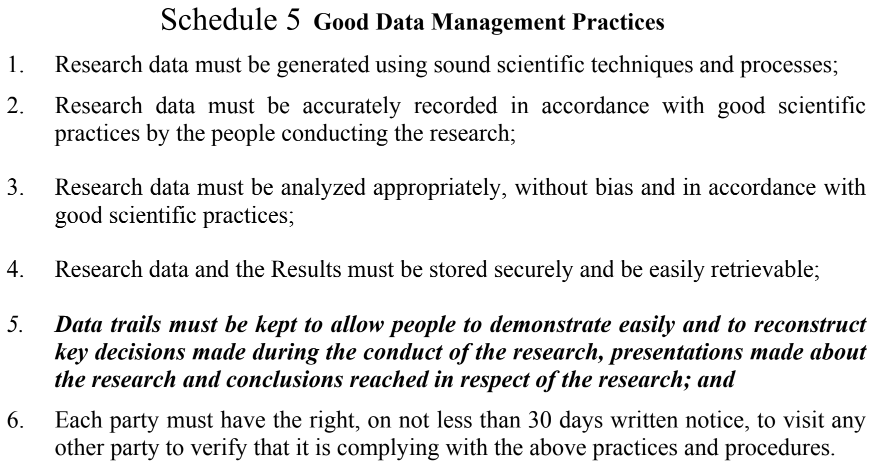
Appendix 3. Collaboration and Related Agreements

A Collaboration Agreement formalises the relationship between project participants and sets out the rights and obligations of each participant. RCUK terms and conditions (http://www.so.stfc.ac.uk/jes/TCfECv1.0.pdf) for research council feC grants stipulate that research project involving more than one Research Organisation must have a formal collaboration agreement in place before the research begins. Some Research Councils (see Appendix 4) further require that an agreement must be in place at the application stage to demonstrate the viability of the collaboration and an effective roadmap for exploiting project IP, which includes the sharing and dissemination of data arising from the project.

EPSRC guidelines [17] indicate that a minimum collaboration agreement should cover the following key topics:

1. Project management - Arrangements for the management and co-ordination of the project.
2. Partners' Responsibilities and liabilities – the rights and obligations (including financial) of each project partner.
3. IP arrangements – exploitation routes and agreements for the ownership, licensing and other arrangements for background and foreground IP.
4. Reporting and publication arrangements, access to results and confidentiality provisions – project report requirements and the protocol relating to timescale and restrictions on disseminating scientific findings – including primary and refined data.
5. Consequences of termination/default and ways of handling disputes – contingency arrangements in the event of partnership changes, actions that can be taken against under-performing partner, and arbitration procedures, etc.

The list illustrates that IP sharing and management are important elements of a collaboration agreement (see 3 and 4 above). Research IP is used here as an embracive term covering all results and outcomes of research, which include data, academic publications, patents and know-how. DIUS offers a toolkit [18] developed by the Lambert Working Group on Intellectual Property for drafting collaborative research agreements. The toolkit includes a step-to-step decision guide and sample agreements on different collaboration models. The decision guide also highlights arrangements for research outputs as key considerations in drafting the collaboration agreement. Indeed, one of its model agreements includes a Good Data Management Schedule which contains a clause (item 5) for maintaining data trails:



Schedule 5 **Good Data Management Practices**

1. Research data must be generated using sound scientific techniques and processes;
2. Research data must be accurately recorded in accordance with good scientific practices by the people conducting the research;
3. Research data must be analyzed appropriately, without bias and in accordance with good scientific practices;
4. Research data and the Results must be stored securely and be easily retrievable;
5. ***Data trails must be kept to allow people to demonstrate easily and to reconstruct key decisions made during the conduct of the research, presentations made about the research and conclusions reached in respect of the research; and***
6. Each party must have the right, on not less than 30 days written notice, to visit any other party to verify that it is complying with the above practices and procedures.

Figure A3. An Example Schedule on Good Data Management Practices [18].

The characteristics of a collaboration will govern the attribution of research outputs and partners' rights to publish. In a commercial contract collaboration, the academic partner and academic host normally relinquish these rights in exchange for a fee. In a research partnership, especially those involving public funds or background IP, the IP sharing requirements would be far more complex. For instance, the academic partners naturally prefer rapid publication in order to claim the glory of the discovery. But premature publication may impair the competitive position of the industrial partners and damage the chance of patenting foreground IP. It is a statutory condition that a patent, or any information about a patent, cannot previously have been published. Even in pre-competitive research collaborations, the release of research outputs to the public domain would still need to be carefully controlled. Industrial partners often have clauses in the collaboration agreement to allow them to read manuscripts before they are submitted, to delay publication if necessary and ask for changes if the manuscripts endanger background IP or damage the possibility of patenting foreground IP [2]. This arrangement permits the early controlled release of research outputs without compromising sensitive information.

Confidentiality or non-disclosure agreement is also commonly used to define a framework for sharing proprietary and/or confidential information. In contrast, a data sharing agreement is more precise as it contains explicit clauses stating what data is shared, the access conditions and how data will be handled. For example, the Northamptonshire Young People's Education & Learning Data Exchange Agreement [19] contains these sections:

1. Agreement date and signatories – information about the parties involved and date of the agreement, including signatures.
2. Purpose – objectives of the agreement.
3. Extent and type of information to be shared – the data sets/items to be shared.
4. Usage and handling – how the data will be used and processed.
5. Security and Data Management - details of the security and data management policies, including how the data will be stored, accessed and disposed of.
6. Complaints and Breaches of Confidentiality – details of the complaint process, disciplinary procedures applicable and how data discrepancies will be handled.
7. Indemnity – liabilities and compensation.
8. General Operational Guidance – arrangements for suspension or termination of the agreement.
9. Designated Officers – official contacts for each participating organisation.

The confidential agreement, on the other hand, is less precise and used generally to cover all shared data in scope to the project. Like the collaboration agreement, it also specifically describes IP arrangements. A confidential agreement could be implemented as standalone bi-lateral agreements between selected parties, or could form a section in the main, multi-lateral collaboration agreement. The STFC CLliK Confidentiality Agreement template has these sections:

1. Agreement date and signatories – information about the parties involved and date of the agreement.
2. Background - information on the research covered by the agreement.
3. Definitions - describe the precise meaning of the terms employed.

4. Confidentiality – describes what are covered, the parties' obligations including how information will be disposed of and exemptions permitted under the agreement.
5. Intellectual Property – describes what are covered, the license conditions and the parties' obligations.
6. Term – conditions of the agreement.
7. General – caveats and governing law pertaining to the agreement.
8. Signatures.

A studentship agreement [an example is given in 20] is a variation of the collaboration agreement as described above. It defines the rights and obligations of the participants, which include the student, the academic supervisor and the organisation hosts. It has two additional sections entitled Thesis and Materials. These cover IP ownership, usage conditions, arrangements for thesis submission and the disposal of unused Materials.

The sharing and protection of IP is a recurring theme in research collaboration, confidential and studentship agreements. Different academic disciplines also have their own domain specific IP sharing agreements. For instance, the Uniform Biological Material Transfer Agreement [21] is used in the US for the transfer of physical biological materials such as DNA, antibodies, model animals, etc. among non-profit institutions as well as between these and commercial institutions. The agreement covers IPR of the participants, caveats and usage conditions of the materials. Since Consequence is concerned with data sharing, we will confine our scope to IP of digital data, e.g. experimental data, metadata, digital documents etc. Readers interested in further examples of IP-related contracts may wish to consult the contracts database (http://www.wipo.int/tk/en/databases/contracts/) provided by the World Intellectual Property Organization [22].

Appendix 4. Data Policies of Research Funding Agencies

The seven UK research councils, in strategic partnership as RCUK, award grants on the basis of a single set of core terms and conditions which include a strong commitment to the open access of research outputs. Under this remit, all grant applications must include a data dissemination plan or roadmap which will be reviewed as part of the assessment process. Funded researchers are currently required to deposit peer-reviewed, published articles in a suitable e-prints repository. Individual council may add extra conditions to their awards to reflect the particular circumstances and requirements of the organisation, or the nature of a specific grant. For example, the ESRC and BBSRC recommend that research datasets are deposited in a suitable community repository such as the UKDA or Protein Data Bank.

A survey of the RCs data sharing and preservation policies show that these policies are more in the form of guidelines or best practices that could be interpreted rather than precise statements typical of a data sharing agreement as described in Appendix 3. For instance, there are liberal uses of qualifying phrases like ‘expect’,’preferrably’,’where possible’,’reasonable’,’appropriate’, etc. The following table summarises the main guidelines of the Research Councils’ own data policies alongside selected non-governmental funding agencies. The information on the last two sectors is based on an RIN report [2] on a survey of key research funding agencies across the public, private and voluntary sectors.

Organisation	Data Management Requirements	Data Publication/Sharing Requirements	Data Curation and Preservation
AHRC	For projects in the archaeology funding area, the ADS must be consulted within three months of the start of a project to discuss and agree the form and extent of electronic materials to be deposited with ADS.	Funded researchers should deposit of a copy of any resultant articles in an appropriate repository, and where possible, also deposit the bibliographical metadata relating to such articles including a link to the publisher’s website. Subject to the compliance with publisher’s copyright and licensing policies. For projects in archaeology, the materials and documentation should be offered for deposit at the ADS within three months of the end of the project. The Research Organisation must obtain a waiver from ADS if this is not possible. AHRC caters to diverse research disciplines that produce different types of research output (eg. physical artifacts, performance acts, oral histories) and accepts that not all outputs are appropriate for on-line sharing. It indicates that the models and mechanisms for publication and access to research results must be both efficient and cost-effective.	Significant electronic resources or datasets arising from the project must be made available in an accessible depository for at least three years after the end of the grant.
BBSRC	All applications must submit a statement on data sharing. This should include concise plans for data management and sharing. In the case of industrial partnership, the applicants must make agreements on the ownership and exploitation of IP arising from the project and highlight the	BBSRC covers communities with different needs. It expects data to be released in a *timely* fashion which would generally be no later than the publication of the main findings and in-line with established best practice in the field. Where best practices do not exist, release should be within 3 years of generation.	BBSRC expects data to be retained for ten years after the completion of a project.

	arrangement in the data sharing statement. BBSRC recognises the need for periods of exclusive use of data but considers that the commercialisation of research does not preclude or unduly delay data sharing.	Data sharing should, where applicable, be via deposition in an existing database or community research repository. It expects researchers using the data to preserve data confidentiality and to observe the ethical and legal obligations pertaining to the data.	
EPSRC	Encourages researchers to manage primary data as the basis for publications securely and for an appropriate time, in a durable form under the control of their host institutions. It requires a statement of how the research outputs will be disseminated. Collaborative projects must have a formal collaboration agreement in place before the project begins, to make sure that the IP arising from the research can be managed effectively.	Supports RCUK open access to research outputs.	Data to be managed securely by institution of origin for an appropriate time.
ESRC	Must carry out data review to ensure funding is not requested for data that already exists.	Data must be offered to UKDA/ESDS for deposit within three months of end of project.	As per UKDA/ESDS policies.
Leverhulme Trust	Applicants are required to submit in their application a dissemination plan outlining the intended distribution of research activities and the publication of research findings.	Recommend deposit of digital resources with community repository where appropriate.	No.
MRC	All funding proposals must include a strategy for data preservation and sharing and, if applicable, a special case for excluding data sharing. Must provide an end of grant report on the data management and sharing activities undertaken by the funded researchers. In the case of academic/industrial partnership, the application must be accompanied by a signed copy of the collaboration agreement detailing how the research outputs are to be shared and disseminated. Funded researchers and their teams must register through their organisations with the CIO and are expected to comply with the DPA.	For medical research involving personal data, the appropriate regulatory permissions (ethical, legal and institutional) must be in place before data can be shared. Published results, where possible, should include link to the associated data. A limited, defined period of exclusive use of data for primary research is reasonable, according to the nature and value of the data and the way they are generated and used. Licensing of IP generated from MRC-funded research should include a provision for research use by other MRC supported scientists.	Data arising from MRC-funded research must be properly curated throughout its life-cycle and released with the appropriate high-quality metadata.
NERC	Programmes must have data management plans.	Must be offered to appropriate data centres after a reasonable time reserved for exclusive use.	As per designated Data Centre.
STFC	New projects must have plans that formalise ownership and agree distribution mechanisms for data before they are funded.	Data should be made available to all, where possible and economical (with an exclusive period where appropriate). The full text and, where possible, bibliography metadata of any articles resulting from the grant that are published in journals or conference proceedings must be deposited at the earliest	Data centres have been supported as projects. Funding for data curation to be reviewed every two years and priorities identified.

		opportunity in an appropriate e-print repository. Subject to the compliance with publisher's copyright and licensing policies.	
STFC ISIS Facility	No but funded researchers are required to comply with ISIS Data Policy.	Raw data and metadata derived from experiments at the facility by non-commercial research project will be released into the public domain three years after the completion of the experiment except special cases. References for publications related to experiments on ISIS must be deposited in the STFC ePubs system within 6 months of publication or during any new applications to ISIS, whichever is the earliest.	ISIS repository accessible via the ISIS Data Portal and permission is based on entitlement.
Wellcome Trust	Applicants are required to provide a data management and sharing plan as part of their application.	Electronic copies of any peer-reviewed research papers arising from funded research must be made available through PubMed Central and UKPMC as soon as possible and in any event within six months of official date of publication. The Trust supports open access to published research and would fund open access fees to authors and publishers to license free use of research papers. The open access policy does not require the deposit of data, although researchers can deposit it alongside their papers in PubMed Central/UKPMC.	Advocates use of PubMed Central/UKPMC to integrate data with the research paper; other public data repositories can also be used.
Commercial Organisations	Data may be shared between university researchers and companies for the purpose of a project.	n/a	Internal data curation is important to some companies, eg GSK, Vodafone.
Government Departments	n/a	n/a	The funding recipient and their institution are responsible for the curation of research data. DFID expects the Data to be retained and accessible. DH also states that data relevant to the findings of research should be accessible.

Table A4: Summary or research data policies of funding agencies [2].

Appendix 5. UK Legislation

Scientific research information spans a wide spectrum of data and may fall under the remit of different legislation. Table A5 summarises the key over-arching legislation followed by a brief discussion of their effects on the sharing of research information. Readers should bear in mind that this is a layperson's interpretation of a broad and extremely complex subject. Those interested in an in-depth and authoritative discussion of the legislation is recommended to consult the ICO (http://www.ico.gov.uk/) and the JISCLegal (http://www.jisclegal.ac.uk/) websites.

Title	Description
Data Protection Act 1998	The Act ensures that personal data about an individual is processed in accordance to legal requirements in order to protect the rights of the individuals. There are 8 data protection principles and it is unlawful to use personal data in a way that breaches any of the principles. The 8 principles require that personal data must be: 1. Processed fairly and lawfully. 2. Obtained for one or more specific and law purposes. 3. Adequate, relevant and not excessive relative to the purpose/s. 4. Accurate and, where necessary, kept up-to-date. 5. Retained no longer than is necessary for the purpose/s. 6. Processed in line with the data subjects' rights under DPA. 7. Secured with appropriate technical and organisation procedures against unauthorised or unlawful processing and against accidental lost, destruction or damage. 8. Kept within the European Economic Area. Personal data can only be transferred to countries/territories with adequate security measures. The Act does not cover information regarding deceased or anonymous individuals. It also distinguishes two categories of personal data: ordinary and sensitive. DPA S33 offers limited exemption to the requirements of Principles 2 and 5 with respect to the use of personal data for research, history and statistics purposes - if the purpose of the research processing is not measures or decisions targeted at particular individuals and it does not cause distress or damage to a data subject. S33 does not exempt researchers from the other requirements, for example, to keep data securely and to only publish results in a form that maintains the anonymity of the data subject/s, etc.
Freedom of Information Act 2000	The act permits any person, anywhere, to request any non-personal, *recorded* information, from any date, held by any public authority unless the information falls under one of the specific exemptions in the Act. Public authorities are required to adopt and maintain a publication scheme approved by the ICO. The publication scheme lists information routinely made available to the public. Any information provided in the publication scheme or has been published elsewhere is exempt from a request for information. The Act does not cover knowledge per se, if it is not in a recorded format. The FOI (Scotland) Act 2002 S.27 has specific provisions for research programme information. These exempt the disclosure of research information: 1. Containing commercially sensitive information or personal data. 2. That is obtained as part of an on-going programme of research. 3. Which is intended for publication within 12 weeks. 4. On Health and Safety ground. These exemptions may be subject to the Public Interest Test which balances the benefits of disclosure against non-disclosure to public interest. There are no equivalent provisions for research information in FOIA but it has similar exemptions. Requests for information under FOIA must be made in writing and is

	generally processed by a human being. The organisation has up to 20 working days to response and can charge a reasonable fee for providing the requesting information in accordance to the FOI Fees Regulations. Most public authorities have a dedicated Information Manager to handle these requests.
Environmental Information Regulations 2004	These incorporate the European directive 2003/4/EC for upholding access rights to any information relating to the environment and anything which may have a negative effect on it. EIR provides broadly similar rights to FOIA but relate specifically to information about the environment. The environment refers to: 1. State. 2. Water. 3. Air. 4. Fauna and flora. 5. Land. 6. Soil. 7. A natural site. EIR bears similarities to, but is not identical to the FOIA regime. There are differences in the exceptions, calculation of fees and in the circumstances whereby private companies may fall within its scope. And EIR applies over FOIA. All EIR requests for information must be made in writing or in a record-able format (eg. email) and would normally be processed by a human being. The organisation has up to 20 working days to response and may charge a fee for providing the requesting information. Most public authorities have a dedicated Information Manager to handle these requests.
Copyright, Designs and Patents Act 1988	CDPA and associated Primary Acts and Statutory Instruments are government legislation which allows those people who are inventors, writers, musicians, or film makers to protect their IPR from plagiarism or theft of their ideas. The four most common types of IPR in the UK are: 1. Patents for inventions. 2. Trade marks for brand identity. 3. Designs for product appearance. 4. Copyright for material, eg. software. Under the terms of CDPA, where intellectual property is created by an employee in the course of his/her employment, that IPR is owned by the employer.

Table A5: Legislation relevant to the management of research information.

DPA and CPDA protect a person's IPR and right to privacy. FOIA and EIR, on the other hand, uphold a person's right to access information. From the perspective of individual researchers occupying the roles of data producer and consumer, they need to observe DPA and CDPA. There would generally be a series of measures taken at different organisation levels to provide a governance framework. At a generic level, researchers would be bound by in-house and the funding agency's data sharing policies or good practices. At a more specific level, they are required to obtain clearances and ethnical approval from the relevant governing bodies if their research involves the use of restricted information such as patient data. For example, see the data publication/sharing requirements for MRC listed in Table A4. Depending on the nature of the data, it may also be necessary to register with the CIO via the organisation's Data Controller and put in place data protection procedures and practices to ensure adequate compliance. Following this assumption, we envisage that that the organisation generic data sharing policies would be sufficient for most purposes. Otherwise, the researchers must define suitable data policies for their data to ensure compliance.

On the other side of the coin, as data keepers, universities, HEI and research councils are considered public authorities and fall under FOIA and EIR. They are obliged to respond to requests for information and may be required to disclose information that they hold regardless of its origins or ownership. An organisation risks abusing DPA or CPDA if it discloses information whilst unaware of any existing or pending IPR and DPA issues associated with the requested information [23]. There are also general concerns that special interest groups with a private agenda will be keen to access research information, for instance, animal right activists seeking information about the research use of animals to target their campaigns; or companies scanning research outputs to further their own commercial interests [24].

Open access is a double-edged sword to public research organisations. Being users, producers and disseminators of information themselves, they will need to minimise their liability to infringement on the one hand and to maximise IP exploitation on the other. Any public disclosure of research materials, including the publication of thesis on the internet, could have an adverse effect on a patent application. In order to avoid prior publication and risk jeopardizing patent application, thesis authors are currently advised to limit access to their research by applying for a restriction order if they are considering applying for a patent [25]. Researchers may also seek similar protection to restrict sensitive data sets from the public domain while awaiting formal publication. Most community data banks permit a time-limited embargo period on submitted data. As described in Appendix 4, RUCK also recognises the need to reward data producers by offering a reasonable period of exclusive exploitation. For example, ISIS, the STFC Muon and Neutron X-ray facility, allows a three-year embargo on the release of experimental raw data generated on its beamlines by publicly funded research projects.

It should be noted that the FOIA and EIR give anyone the right to access any recorded information held by an organisation, regardless of who owns the IPR in the information, unless an exemption can be justified as permitted by the Acts (see Table A5). In other words, a restriction order by the author is not sufficient ground for a refusal. Consequently, when a FOIA request is received for restricted material, it will be critical for the organisation’s Information Officer to have the appropriate support information to decide whether or not the request could be lawfully refused. The support information could be security metadata, policies defined in data sharing, licensing, collaboration or confidentiality agreement. Thus, the data owners must be pro-active in the use of data sharing or similar agreements to prevent accidental disclosure. Another mitigation strategy would be to prevent requests for information from arising in the first place. Rather than holding their own copies, organisations may consider sharing sensitive digital information via a secure data sharing framework such as Consequence. Our test bed scenario describes a multi-organisation academic/industrial collaboration which does not involve sensitive personal data, but there are significant background and foreground IP at stake. The challenges to control the access and usage of restricted information will be parallel those facing peer organisations under the current legislation.