Sharing Scientific Data: Scenarios and Challenges

Shirley Crompton[⋆], Benjamin Aziz[+], Michael Wilson[+]

[⋆]e-Science Centre, STFC Daresbury Laboratory, UK
[+]e-Science Centre, STFC therford Appleton Laboratory, UK

1 Introduction

Inter-disciplinary study and collaborative research with industry is changing the way researchers interact, share data and manage intellectual properties. 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.

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 or regulatory legislation.

In this position paper, we outline typical use case sccenarios associated with the scientific data sharing process and the challenges these scenarios raise.

2 The Scientific Collaborations Scenario

The main scenario we present here is based on a public-private research collaboration with limited lifetime, such as five years, co-funded by a key *public funding agency* in some scientific field, such as Biosciences, and a *small-to-medium* *enterprise* technology company, for example, with specialisation in bioinformatics. such collaborations typically involve academic researchers from *universities* or *industrial research departments* with relevant expertise, a *scientific facility provider*, whose facility provides the infrastructure for performing scientific experiments and finally, an *e-Science infrastructure owner* to provide the data management and dissemination technologies.

The award from the funding agency 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 design algorithms and proprietary data to the partners. In line with the agency's funding criteria, the partners will make agreements at the grant proposal stage on the licensing, ownership and exploitation of foreground and background intellectual property rights and the publication of research outputs. The agreement also includes a schedule on good data management practices, demnading the use of data portals and information cataloguing to manage research outputs and to share sensitive data.

3 Data Sharing Use Cases

The scenario involves typical data sharing behaviour in a cross-domain scientific project that uses 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. For example, researchers within the same university will be able to communicate via a local-area network. Cross-domain data sharing traversing administrative boundaries would likely involve the use of Secure SHells (SSHs), Virtual Network Computing (VNC), resource brokering and centrally managed services like community data portals or virtual on-line collaboration tools.

Data generated from experiements will be held on a *cataloguing database*, which is exposed via a web service API, for example, to a *data portal*. The database provides information about the scientific data, i.e. the scientific metadata based on some model, which could be anything from a description of the conditions under which the data were generated, to more high-level associations with dissemination policies. Such server-based catalogue systems typically incorporate simple role-based access control mechanisms, however they usually do not enforce policies for wide-scale data sharing and usage, as we shall demonstrate in the following use cases.

3.1 Agreements Specification and Policy Administration

Data sharing requirements are usually captured by means of formal collaboration agreements among the partners, typically established during the negotiation phase of the proposal preparation. An analysis of data sharing agreements (DSAs) indicates that they are usually concerned with questions regarding what data will be shared, the delivery/transmission mechanism, the processing and security framework, disposal policies and liabilities and sanctions. As Figure 1 illustrates, our scenario depicts a situation where a central scientific facility establishes DSAs individually with each partner in the collaboration.

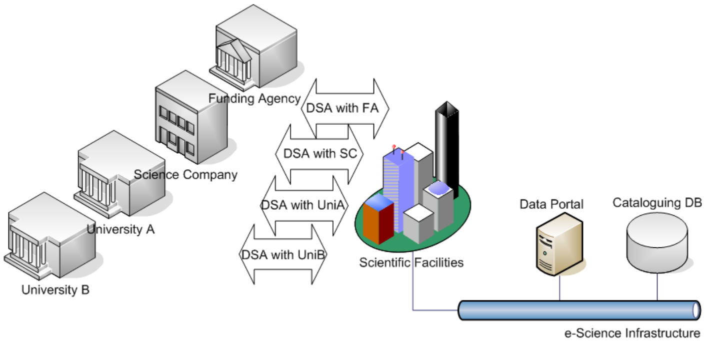

Figure 1: DSAs between the Scientific Facilities and the Collaborators.

The setting-up of the DSAs requires at least two technologies: A) a DSA authoring tool, which may or may not provide some reasoning on the DSA and B) a library of controlled natural language vocabulary, which can define unambiguously the DSAs conditions and obligations. Furthermore, a DSA reasoning tool is required in case some formal verification of DSA properties, such as proving the lack of conflicts, is to be established. Finally, automatic translation of DSAs to enforceable policies is also desireable to lessen the burden of DSA-to-policy translation. This last step leads to another important step; how to manage the composition of the resulting policies with pre-existing policies currently present in each partner's domain. The relatively long timescale of a collaboration introduces additional challenge: where it may be the case 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, new data sharing requirements for emerging results not covered by existing agreements.

3.2 Server-based Data Accessing

Once high-level data sharing agreements have been drafted and agreed-upon by all partners, they are refined to enforceable policies, which are in turn deployed on a central server accessible by the catalogue API to provide controlled accesses to the scientific information that it manages and also to the actual datastores hosting the data.

For example, a typical use case for accessing scientific data and their metadata could follow along the lines: 1) a user requests some data object from the data portal, 2) the data portal passes the request to the cataloguing database, 3) the cataloguing database then, after checking the user's credentials against the request attributes and the local policy, resolves the physical location and returns a download url to the data portal, 4) the data portal then renders a hotlink on the web page and presents this to the user, 5) the user clicks on the hotlink to download the data object. The data portal re-directs the request to the HTTP server, which verifies the user request with the cataloguing database and streams the object back to the client browser, then finally, 6) the browser prompts user to display or save the object.

Figure 2 represents an example of possible accesses from the science company and one of the universities in our scenario.

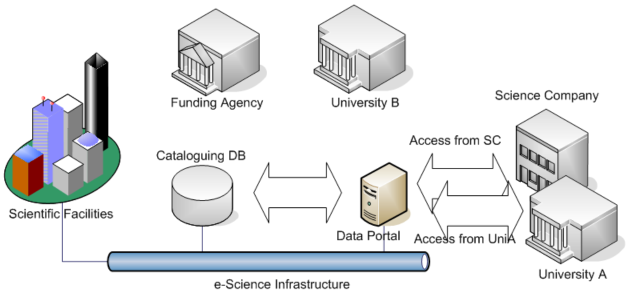

Figure 2: Metadata and Data Accesses.

3.3 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 scripts, wrappers or resource brokering middleware. This online sharing and using of data must also be controlled via usage policies that enforce what can be done on the data and to whom the data may be sent. Figure 3 represents an example of possible metadata and data sharing in our scenario.

At a more complex level, sharing may take the form of an automated scientific workflow. The workflow resource broker or subprocess then uses delegated credentials from the user 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. Therefore, the proposed solution must be capable of monitoring environmental parameters to ensure the correct enforcement of context-related low-level data policies at sharing time.

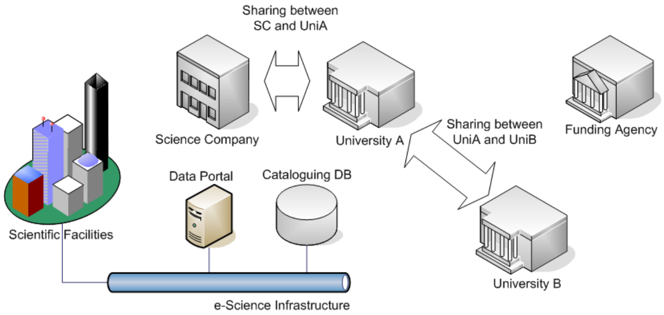

Figure 3: Metadata and Data Peer-to-Peer Sharing.

Finally, data may be analysed or altered in transition, which leads to the creation of new data bringing out issues related to the propagation of the parent data policies and whether these should also be subject to derivation or not. This is a realistic issue since scientists are quite likely to write their own algorithms in scientific programming languages, which typically have a procedural design and perform some operations while iterating through a dataset.

3.4 Offline Data Sharing

Research is a creative activity and it is common practice for researchers to carry out their activities when and where they feel appropriate. To accommodate this practice, it is desirable to permit secure data usage in an environment without network connection. This raises the requirement that digital documents be protected by security policies, metadata and licences. The security policies will be evaluated against the security metadata or licence and enforced where appropriate. Users will access the protected file and their actions on the file will be logged locally. When the network communication is resumed, an auditing processing 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.

4 Conclusion

In this paper, we outline the various challenges that need to be addressed by an end-to-end secure data sharing framework for the scientific research domain, including enforcement across administrative domains where network connection may not be available and the capabilities for runtime evaluation of dynamic usage conditions and obligation execution.