## Annex A - Proposal Cover Sheet: Federated Tools and Services Call

| **Cover Sheet for Proposals**<br>***(All sections must be completed)*** | JISC **Capital Programme** |
|---|---|

| | |
|---|---|
| **Name of Capital Programme:** | **e-Research : e-Infrastructure** |
| | |
| **Name of Lead Institution:** | King's College London, Centre for e-Research |
| **Name of Proposed Project:** | ASPiS (Architecture for a Shibboleth-Protected iRODS System) |
| **Name(s) of Project Partner(s):** | Science and Technology Facilities Council |

**Full Contact Details for Primary Contact:**
**Name: Mark Hedges**
**Position: Deputy Director, Centre for e-Research, King's College London**
**Email: mark.hedges@kcl.ac.uk**
**Address: King's College London**
**26-29 Drury Lane (Room 303),**
**London , WC2B 5RL.**
**Tel: 020-7848-1970**
**Fax: 020-7848-1989**

| | | | |
|---|---|---|---|
| **Length of Project:** | **15 months** | | |
| **Project Start Date:** | 1st January 2008 | **Project End Date:** | 31st March 2009 |

| | |
|---|---|
| | |
| **Total Funding Requested from JISC:** | **£178,898.28** |

**Funding Broken Down over Financial Years (Apr–Mar):**

| **Apr07 – Mar08** | **Apr08 – Mar09** |
|---|---|
| **£34,923.5** | **£143,974.78** |

| | |
|---|---|
| | |
| **Total Institutional Contributions:** | **£95,779.15** |

**Outline Project Description**

The ASPiS project will enhance the access management functionality provided by iRODS (Rule-Oriented Data Management System), an open source data grid middleware system developed at the San Diego Supercomputer Center (SDSC) as the successor to the widely-used Storage Resource Broker (SRB). The project will integrate Shibboleth access with iRODS, allowing iRODS-based data grids to be deployed within the UK Access Management Federation for Education and Research, and providing a more scaleable, flexible and user-friendly means of implementing authentication and authorisation in iRODS-based data grids.

The project will focus on two complementary aspects of managing access to data:

- Control of access to iRODS data resources, allowing iRODS to capture user attributes from Shibboleth, and to use those attributes to make authorisation decisions.
- Capture of provenance metadata for iRODS data resources.

A prototype service will be made available for evaluation to the user community, and in particular to National Grid Service users.

| | | |
|---|---|---|
| **I have looked at the example FOI form at Appendix A and included an FOI form in the attached bid (Tick Box)** | **YES** ✓ | **NO** |
| **I have read the Circular and associated Terms and Conditions of Grant at Appendix B (Tick Box)** | **YES** ✓ | **NO** |

**FOI Withheld Information Form**

We would like JISC to consider withholding the following sections or paragraphs from disclosure, should the contents of this proposal be requested under the Freedom of Information Act, or if we are successful in our bid for funding and our project proposal is made available on JISC's website.

We acknowledge that the FOI Withheld Information Form is of indicative value only and that JISC may nevertheless be obliged to disclose this information in accordance with the requirements of the Act. We acknowledge that the final decision on disclosure rests with JISC.

| Section / Paragraph No. | Relevant exemption from disclosure under FOI | Justification |
|---|---|---|
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

Please see http://www.ico.gov.uk for further information on the Freedom of Information Act and the exemptions to disclosure it contains.

**Proposal: ASPiS (Architecture for a Shibboleth-Protected iRODS System)**

**Submitted under: JISC Circular 02/07: e-Infrastructure (Federated Tools and Services)**

**Submitted by: King's College London, Centre for e-Research**

## 1. Appropriateness, Fit to Programme Objectives, Overall Value to JISC Community

1.1 Project timescale: 15 months, from 1st January 2008 until 31st March 2009.

1.2 The proposed ASPiS project will be a collaboration between the Centre for e-Research (CeRch) at KCL, which incorporates the Arts and Humanities Data Service (AHDS), and the Science and Technology Facilities Council (STFC). Its broad aim is to enhance the access management functionality provided by iRODS (Rule-Oriented Data Management System), an open source data grid middleware system developed at the San Diego Supercomputer Center (SDSC) as the successor to Storage Resource Broker (SRB), which is widely used in the UK and elsewhere for the management of research data. In particular, we will enhance iRODS by integrating Shibboleth access and thus allowing iRODS-based data grids to be deployed within the UK Access Management Federation. Focussing on iRODS rather than SRB for this work is important for three reasons: i) iRODS is expected to replace SRB within the next few years; ii) the iRODS licence is more suitable for scientific data services; iii) iRODS micro-services permit us to add rule-driven data management features.

1.3 The project will focus on two complementary aspects of access management to data:

- Controlling access to iRODS data resources – *who* can access *which* files, and *how*.
- Capture of audit/provenance metadata for iRODS data resources – *who* accessed *which* files, *when*, *how*, and with *what* outcomes.

1.4 The first of these is the more familiar application of Shibboleth. In this part of the project we will develop software components allowing iRODS to request user attributes from Shibboleth, and to use those attributes to make authorisation decisions. This will enable authentication in an iRODS data grid to be devolved onto a user's home institution, and will allow data grid managers to implement finely grained role-based access control for data resources. The result will be a more scaleable, flexible and user-friendly means of implementing authentication and authorisation in iRODS-based data grids.

1.5 Provenance of data is a key issue in most sciences, and particularly in a dynamic grid environment, where numerous users and virtual organisations can execute a wide variety of services and workflows that create and modify data. Provenance metadata may be regarded as representing the steps by which a particular piece of data was derived, and it is fundamental when assessing the quality and accuracy of information, and also provides added value to grid users who subsequently publish, cite or further process the data. Provenance metadata typically incorporates a variety of elements, but an important component will be the identities or roles of the users that played a part the data's derivation, information that Shibboleth is well placed to contribute. In this part of the project, we will implement software components that allow iRODS to create and record provenance metadata for data when it is created or modified, including but not limited to the Shibboleth attributes.

1.6 We will provide a prototype service to the National Grid Service (NGS) for evaluation among the wider user communities. This will combine well with existing proposals to provide production-ready Shibboleth access with integrated VO role management for the NGS. We expect the NGS will appear as a Shibboleth SP in the UK Access Management Federation by the time this project starts wider testing, but there are other ways of involving NGS communities in evaluations if the NGS has no SP.

1.7 Our design approach throughout will be to develop modular service-based components that encapsulate key functionality and conform to common interfaces. The project represents the community's first steps in integrating Shibboleth and provenance with the iRODS system, and we consider that our approach will greatly facilitate subsequent changes and extensions to the software, as well as allowing the future integration of different service implementations within the same framework. All software will be founded on detailed and validated use cases, user-centric iterative development, rigorous testing, and close liaison with actual and potential users throughout the project.

1.8 A key benefit of this modular approach is that projects needing Shibboleth access management but not provenance can use just the Shibboleth components, others that need provenance can take this and adapt it to their needs, e.g. to take advantage of other authentication systems. The work will be made available under an OSI-approved open source licence to the wider iRODS communities. For example, we expect members of the US InCommon federation to be interested.

1.9 The project will make several contributions to the Capital Programme and the wider community:

a) It will provide a set of use cases, which can be leveraged by other projects, concerning the management of access to research data in virtualised or grid environments.

b) It will provide a rule-based access management system for iRODS data grids, integrated within the Shibboleth-based UK Access Management Federation for Education and Research.
c) It will make available a prototype Shibboleth-enabled iRODS data grid for user evaluation.
d) It will demonstrate the utility of iRODS to the numerous SRB-based data grid users in the UK, both NGS users and other users supported by STFC.
e) It will initiate the creation of a sustainable user community for iRODS in the UK.
f) The project will provide a valuable case study into the application of Shibboleth in a data grid environment, both for access management and provenance capture.

## 2. Quality of Proposal and Robustness of Workplan

### *2.1 Project Background*

2.1.1 Research across the academic disciplines is increasingly both driven by and a generator of data on a large scale, the so-called data deluge. In the “big” sciences, the primary emphasis has been the management of very large data sets, but the issue frequently arises in other disciplines as well, as it is increasingly necessary to retain both the original data and processed versions of the data. These requirements arise partly from the need to prove that analysis was carried out correctly, partly because computations can be expensive and time consuming so cannot easily be repeated, and partly to enable future analysis of intermediate datasets. These complex workflows are often supported by data grid middleware, for example the widely used Storage Resource Broker[1] (SRB). STFC support numerous production SRB systems on behalf of researchers in a variety of disciplines.

2.1.2 iRODS[2] is an open source project developed by SDSC as the successor to SRB, with significantly enhanced functionality. A particular feature of iRODS is the ability to represent data management policies in terms of rules. The system incorporates a Rule Engine, which interprets these rules and allows pre-defined sequences of actions (represented as "micro-services") to be executed in response to certain triggers, e.g. when an data object is stored, read, or updated. These rules have great potential for implementing data management strategies that are to take place "under the hood".

2.1.3 With its Rule Engine, iRODS addresses a particular limitation of the SRB, by providing a simple, flexible and integrated means of implementing specific application requirements that are not supported by the core software, and in particular providing a mechanism for accessing external systems. These could include, for example, access management systems, or systems that support more complex metadata than is provided natively by SRB or iRODS. This enables integrated iRODS-based systems that incorporate the best of both (or several) worlds, combining the virtualised storage of the data grid with external services that encapsulating specialised processing and knowledge.

2.1.4 This is of crucial importance when we consider that in many disciplines the “data deluge” concerns not only the size of the data, but also its complexity and diversity, which is reflected in the complexity of the metadata required to manage both the data and access to the data. For example, a humanities researcher may work with digital resources that combine of textual objects (enriched with various degrees of XML mark-up), databases and multi-media objects; a medical researcher may use large images with complex, detailed annotations and links to other resources.

2.1.5 When making digital resources more widely available, access management issues are of key importance, particularly in a highly disaggregated grid environment where access may depend on dynamic and ad hoc arrangements between users, institutions and virtual organisations. Much of the research community is simply not going to engage with grids and make resources available in this way, if security issues are not addressed to their satisfaction.

2.1.6 SRB addresses user authentication by means of X.509 certificates or centrally managed user lists and passwords, in combination with access control lists and groups to handle authorisation of access to individual grid resources. Natively, iRODS follows a similar approach. However, identity-based authorisation does not scale well, and does not easily support role-based access in dynamic environments involving virtual organisations working on complex data. In addition, the technical complexities of setting up and using certificates will not encourage a wide uptake of grid use among users who are less likely to be technically knowledgeable, e.g. in arts and humanities disciplines.

[1] http://www.sdsc.edu/srb
[2] http://irods.sdsc.edu

## *2.2 Overview of Proposed Work*

2.2.1 We propose to address two complementary aspects of access management for virtualised data:

- access control that is based on user roles (including, e.g., VO membership) allowing access rights to be defined for individual files, modes of access (e.g. get, put, update, copy), and user roles (not just the users' identities).
- generation of audit and provenance information that tracks access to resources.

2.2.2 To address the first of these we plan to integrate Shibboleth with iRODS, enabling the authentication of a user to be devolved onto the user's home institution. Middleware will be implemented to capture a user's Shibboleth attributes (which can include a unique user identifier and multiple roles) and make them available within iRODS, allowing finely grained access control to data resources. We will investigate how these attributes/roles can be used to make authorisation decisions within the iRODS context. A major line of investigation will be the use of iRODS' rule architecture, whereby an attempt to access a resource results in the execution of a "micro-service", which could be used to determine access rights. We will also investigate the use of external systems such as PERMIS, an authorisation system implementing "Role Based Access Control" authorisation.

2.2.3 As well as using the Shibboleth attributes to control access to grid-based data, we intend to investigate the complementary use of these attributes in the generation and recording of provenance metadata. The concept of provenance is of key importance when assessing the quality and accuracy of information and services available in a grid. Typically, in a grid-based research environment many people can create data, and apply services and workflows to analyse and process this data to produce new results. The assessment of the validity or reliability of information will be dependent on the entire sequence of actions (and actors) that culminated in its current state[3]. Broadly, provenance metadata represents the series of steps by which a particular piece of data was derived, and can include such information as the ancestral data from which it was derived, workflows and services that were executed to derive it, the inputs and outputs of those workflows and services, as well as information about users who initiated them. Such information represents added value for e-Science users who subsequently make use of the data, whether by publishing, citation or further processing.

2.2.4 An important part of the provenance of a data object relates to the agent that produced the data. Some data may result from entirely automated workflows, but often a user is the ultimate cause. A user's Shibboleth attributes may contain several items of relevance for recording provenance: in terms of the eduPerson attributes recognised by JISC as providing the necessary core functionality for the UK academic community, these include *eduPersonPrincipalName* (a persistent identifier for the user[4]) and *eduPersonEntitlement* (indicating, e.g., membership of a number of research groups)[5]. We will implement a service (or services) to generate provenance information, incorporating Shibboleth attributes among other metadata, for data objects held in iRODS data grids, when they are created, read or modified. We realise of course that the UK Access Management Federation encourages IdPs to not publish personal data, so we will encourage our user communities not to require this in their provenance data, in the spirit of the federation policies. However, since some communities may genuinely require personal data to be stored in the provenance metadata, we will provide means to associate personal identifiers with the *eduPersonTargetedID*, with the users' consent. This process of association may also involve formal authorisation by the user's project PI for additional security and role assignment.

2.2.5 Note of the current status of iRODS: Development of iRODS began in mid-2005, and since then it has undergone extensive testing. The current version (0.9.2) contains a significant subset of the intended functionality (including those components required for this project), and has a demonstrable level of reliability. CeRch & STFC have been collaborating on applying iRODS rules to the curation of research data (see Paragraph 5.1). IN2P3 have been carrying out extensive trials of iRODS, in particular performance testing.

## *2.3 Workpackage Breakdown*

**Workpackage 1: Project Management**

[3] http://www.cs.indiana.edu/~ysimmhan/l/pubs/simmhan-record-2005.pdf
[4] However, there are issues to be addressed with *eduPersonPrincipalName*: an institution will not in general publish this attribute, and also may choose to recycle the value after a user leaves.
[5] http://www.jisc.ac.uk/uploaded_documents/JISC_Fed_doc_full.doc

2.3.1 This workpackage includes all management activities, including planning, coordination with partners and external bodies (e.g. the NGS), reporting, and assess of risks and opportunities as the project progresses. It also covers project advocacy and dissemination.

2.3.2 The project will be managed in accordance with the JISC project management guidelines. The inter-related nature of the work being carried out by the two partner institutions will necessitate continuous and close liaison between all parties. There will be monthly meetings between the Project Director and Senior Technical Officer, together with a meeting of the whole team every three months. In addition, there will be more informal communication on a continuous and *ad hoc* basis.

2.3.3 Deliverables will include:

- Detailed project plan; progress and risk assessment reports; website & dissemination.

**Workpackage 2: Review of existing research and technology**

2.3.4 To ensure that the development has a firm foundation, it will be preceded by a brief review of:

- the main technologies relevant to the project: iRODS, Shibboleth and others;
- existing grid-Shibboleth integration projects;
- projects investigating provenance metadata, and tools for generating, browsing and otherwise using provenance metadata (e.g. Taverna, PASOA, Provenance Explorer).

2.3.5 Although the proposed personnel have experience of each of iRODS, its predecessor SRB, and Shibboleth, this is not evenly spread throughout the project; this WP will enable staff members to fill in the gaps. Additionally, the prior experience of other relevant projects will be invaluable for subsequent workpackages, to avoid either re-inventing the wheel or repeating the mistakes of others. As part of the JISC-funded SERAPIS project, the AHDS produced a report surveying existing grid-Shibboleth integration projects; this will be useful input, and will be updated to reflect more recent developments.

2.3.6 Deliverables:

- Survey document
- Training event for technical personnel.

**Workpackage 3: Use Cases**

2.3.7 To ensure that the implementation is grounded in authentic user requirements, we will produce detailed use cases relating to managing access to data in a virtualised data grid environment, specifically looking at:

- Role-based, fine-grained control of access to data;
- Generation, capture and use of audit/provenance metadata.

2.3.8 We emphasise the *use* of provenance metadata here, as surveys of data provenance in e-Science have indicated that the nature and format of the metadata recorded has been strongly dependent on the requirements for use[6], and diverse requirements have led to an equal diversity of representations. In order to integrate provenance metadata creation with iRODS in a way (or ways) useful to the community, we must determine how the metadata will be used, which may include browsing, citation, automated analysis and reasoning. We aim to develop a machine-interpretable representation that is sufficiently flexible to accommodate a diverse subset of the requirements and applications that may be expected in an iRODS data grid environment. It will also be extensible, to allow additional information to be built in as new requirements are encountered.

2.3.9 The use cases will be written in consultation with a variety of stakeholders (see Section 3.1). Note that engagement with the stakeholders will not be limited to this phase, but will involve a continual process of evaluation throughout the project. In addition to this user engagement, the use cases will take into account the more abstract and wide-ranging requirements for access produced by the ESP-GRID project[7], and the provenance use cases produced by the PASOA project[8].

2.3.10 Deliverables:

- Use cases; requirements survey, in particular addressing the requirements for personal data in provenance.

**Workpackage 4: Software Architecture**

[6] http://www.cs.indiana.edu/pub/techreports/TR618.pdf
[7] http://wiki.oucs.ox.ac.uk/esp-grid/UseCases; http://wiki.oucs.ox.ac.uk/esp-grid/GridRequirements
[8] http://eprints.ecs.soton.ac.uk/10269/01/pasoa04requirements.pdf; http://www.pasoa.org

2.3.11 This workpackage covers the specification of the overall architecture for the iRODS enhancements. Our general approach will be to develop additional modules that can be incorporated into iRODS rules without changes to the core software, following the design approach of iRODS that uses a standard "micro-service" interface. In some cases, changes to the core software may be necessary, e.g. when capturing and storing Shibboleth attributes; however in any case a modular approach will be taken and changes to the core will be minimised. A key design aim is to decouple the services from the iRODS architecture to enable different implementations of authorisation or provenance services to be used by different iRODS systems. For example, some systems may wish to use PERMIS for access authorisation, others a different service; different systems may require provenance metadata to be recorded in different ways. In each case the architecture will facilitate the substitution of one implementation for another by using a common interface layer.

2.3.12 Within this framework, we will address a number of possible implementation architectures, from which our preferred solution will be selected. The Software Architecture document will describe this solution in detail, along with a higher-level description of the rejected approaches. At the current stage, it is of course not possible to describe this architecture in any detail. However, it is likely to include the following components[9]:

- a Policy Information Point (PIP), which captures and stores user attributes from Shibboleth.
- a Policy Decision Point (PDP), which processes each access request, receives user attributes from the PIP, applies the access control policy, and determines whether to allow or reject the request. To accommodate the use of third party authorisation modules, this component is likely to be divided into two parts, a component that interfaces to iRODS, and a component (which may for example be PERMIS-based) that makes the decision.
- a Policy Enforcement Point (PEP), which enforces the decision made by the PDP.
- components for generating provenance metadata. Again, like the PDP, we will develop at least two components, one acting as a generic interface, another generating the metadata, so as to accommodate future modules satisfying different metadata requirements.

2.3.13 Deliverables:

- Software architecture specification.

**Workpackage 5: Software Development**

2.3.14 This workpackage includes the detailed design, coding, unit testing and initial integration of the software components. To ensure that the software meets the needs of the community as a whole, we will follow a user-driven, evolutionary approach, involving incremental cycles of implementation and assessment in collaboration with potential users and other stakeholders (see Section 3.2), and in particular the NGS, of which STFC are significant partners. This iterative approach implies that WP 5 overlaps with WP 6 (Testing and Evaluation) to a significant degree.

2.3.15 The workpackage also includes the production of user documentation:

- a technical guide to enable subsequent developers to further enhance the software.
- an installation, configuration and administrative guide, for grid managers using the software.
- documentation for project administrators, explaining how to manage the system.
- documentation for end users, taking into account their potential lack of technical expertise.

2.3.16 Deliverables:

- Detailed technical specification of software components.
- Module tested and integrated software components.
- Documentation: technical specification; test descriptions and logs; user guides.

**Workpackage 6: Testing and Evaluation**

2.3.17 We intend to produce software that is usable in other iRODS systems, at least as a beta version, and consequently rigorous testing is of key importance. All formal testing will be fully documented: test specifications will be written; the tests will executed and the results logged, together with any issues (such as test failures) that arise; software will be updated and rebuilt to resolve these issues where possible. Given the incremental approach to development, testing may be repetitive, so we will create sets of automated test scenarios that can be executed against the software without human interaction, to validate the software after any significant changes have been made.

2.3.18 Testing will take place in several environments:

[9] The terminology is explained in http://www.terena.nl/events/tnc2006/core/getfile.php?file_id=753.

a) Initially, the software components will be tested both individually and in combination by the developers in localised test environments at STFC and CeRch.
b) We will set up an iRODS data grid involving servers and storage at both STFC/RAL and CeRch, for executing more realistic test scenarios, internally to STFC and CeRch.
c) We will set up a prototype iRODS data grid based at STFC/RAL, which is a core NGS site, with additional storage at CeRch, for user testing and evaluation. This will be a key aspect of our user-driven approach to development. The grid will be populated with initial test data, although users will be able to add their own (subject to space limits). Potential evaluators are listed in Section 3.2

2.3.19 The aim is that a given "release" of the components will pass though these three test environments in succession, becoming increasingly robust as it does so. There may be several "releases" of the software being tested at any one time in different environments, so close attention will be paid to issue tracking and version control: all software changes will be tracked using the Subversion, and bugs/issues will be tracked using an appropriate tool, such as TRAC[10].

2.3.20 The prototype in (c) will be protected by a Shibboleth SP located at STFC/RAL and registered by STFC as an SP with the UK Access Management Federation for Education and Research, of which both STFC and KCL are members and have IdPs. The prototype will be made available as a test bed to NGS users that are affiliated to an institution within the UK federation. Currently, this covers approximately 75% of the UK-based NGS users, and this percentage will increase as more institutions join the federation, and more members set up IdPs.

2.3.21 To enable non-federation users to access the service, depending on the feasibility and the needs of the user communities, we will investigate setting up a non-federation IdP that is not coupled to an institution's authentication service, which we can use to manage the users. Naturally, such an IdP cannot join the UK federation. We will thus need to set up an independent WAYF that can redirect users both to the federation IdPs and to the test IdP. Alternatively (or additionally) the prototype iRODS service will also support the standard iRODS authentication and authorisation mechanisms for non-Shibboleth users, thus enabling federation and non-federation authentication to co-exist. We will use the Internet2 implementations of Shibboleth, and test both with Version 1.2, which is the version currently used by the UK Federation, and Version 1.3, which is the most recent group of releases.

2.3.22 Deliverables:

- Test specifications, automated test scenarios, test logs, issue logs.
- Modified software.
- Working prototype iRODS data grid incorporating the enhancements produced by the project.

**Workpackage 7: Final reports**

2.3.23 Produce final release of software and final versions of all documents. Many documents will be produced in the first instance as drafts, and will be updated during the project as a result of feedback from user evaluation and other sources. In addition, we will produce a Final Report that incorporates a case study addressing the issues raised by the project, evaluation results, and recommendations for follow-up work based on the experiences of the project.

### *2.4 Summary of Main Deliverables*

a) A set of use cases and a requirements survey (see WP 3).
b) Modular software components and rules for iRODS (and possibly for Shibboleth), implementing the following functionality: capture of Shibboleth attributes, use of Shibboleth attributes for determining access to iRODS data resources, capture of audit/provenance metadata (see WP 4).
c) Comprehensive documentation, including: detailed technical specifications and test documents; survey document from the review of related research and technologies; user guides.
d) A prototype Shibboleth-enabled iRODS data grid, based at STFC/RAL with additional storage at CeRch, and using the software from (b), available as a test bed for NGS and other users.
e) Final report, incorporating case study.

### *2.5 IPR*

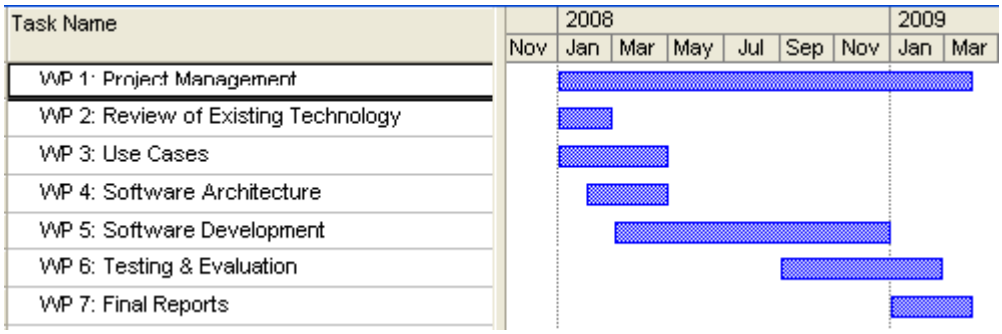
2.5.1 IPR in all reports and other documents produced by the project will be retained by the authors and host institutions but made freely available on a non-exclusive licence as required by JISC. All software created during the project will be made available to the community on an OSI-approved open-source basis on the GPL licence. We will respect the licence model of all third party software used during the project, most of which is made available under open source licences.

[10] http://subversion.tigris.org/; http://trac.edgewall.org/

*2.6 Risks*

| Risk | Probability (1-5) | Severity (1-5) | Score (P x S) | Action to Prevent/Manage Risk |
|---|---|---|---|---|
| Difficulties recruiting and retaining staff. | 1 | 4 | 4 | Most of the nominated staff are already employed at CeRch & STFC. There is a broad knowledge of the technologies involved within the partner organisations.<br>Spread expertise throughout the project, document work so that knowledge is not lost. |
| A complete solution cannot be implemented within the project time constraints | 2 | 3 | 6 | The absence of a complete solution is not an indication of failure, as one aspect of the project is to investigate potential problems. The project report will address the issues that could not be resolved. |
| Dependencies on other projects (specifically, the iRODS development project). | 1 | 2 | 2 | A production-quality release (1.0) of iRODS is due in Oct. 2007. The current release (0.9.2) has been used by the partners and is a sufficiently sound basis for the project.<br>Expertise in iRODS is available at STFC and CeRch. STFC liaises closely with the SDSC iRODS team. |
| Failure to meet project milestones. | 2 | 3 | 6 | Produce project plan with clear objectives.<br>Continuous project assessment and rescheduling when necessary. |
| NGS does not have Shibboleth access when testing starts | 2 | 2 | 4 | NGS has deployed prototype Shibboleth portals, e.g. in the ShibGrid project. We can reuse prototypes from this project. |
| Lack of engagement from user communities | 1 | 4 | 4 | Identify relevant communities early. Both STFC & CeRch work closely with diverse communities and have existing SRB users. |

*2.7 Outline Project Timetable*




## 3. Engagement with the Community

*3.1 Engagement with project stakeholders*

3.1.1 The primary stakeholders of the project are:

- Research projects and staff who generate, manage and consume research data, in particular in distributed and virtualised data grid environments.
- Research projects and staff who would do this if their requirements were better supported.
- Funding bodies, including JISC and the research councils.

3.1.2 All of these stakeholders have an interest in making the results of their research (or the research that they funded) available, while both controlling access to these results (whether for reasons of sensitivity, privacy, commercial value) and tracking or monitoring access to the results and the uses that are subsequently made of them (e.g. in publications or further research).

3.1.3 The project partners have close connections with a wide variety of users and potential users in various disciplines. STFC currently manages data on a very large scale for researchers in a variety of (mainly scientific) disciplines, using SRB-based data grids, both on behalf of the NGS and on their own initiative. It is expected that the future users of iRODS will include this community, which will form a substantial part of the stakeholders consulted. Consequently STFC has an extensive knowledge of these users' requirements for access management, which we shall build on in WP 3.

3.1.4 While arts and humanities researchers are at present less engaged with grids, such technologies have great potential for them. Simplification of access management is particularly important in these disciplines, as technical knowledge cannot be assumed and use of certificates is not widespread. In order to achieve engagement with this community, the project will liaise with Tobias Blanke of the Arts and Humanities e-Science Support Centre (AHeSSC), which is hosted by CeRch. AHeSSC is a partner in the JISC-funded Enabling Uptake of e-Infrastructures project, collaborates in the e-Infrastructure Use Cases and Service Usage Models projects on arts and humanities issues, and has contacts with the burgeoning community of potential users in this area (e.g. those funded by the AHRC-EPSRC-JISC e-Science programme). ASPiS will also address the user needs expressed in the Arts and Humanities e-Science Scoping Study[11] produced by the AHDS.

### *3.2 Evaluation*

3.2.1 To ensure that the project outputs meet the needs of the research community, we will follow throughout the project a user-driven approach to development, involving incremental cycles of implementation and evaluation in collaboration with targeted evaluation partners. These will include:

- Projects/researchers using the SRB services provided by STFC, both within and outside the NGS.
- Other NGS users, possibly including the UK e-Science Engineering Task Force.
- The JISC e-Infrastructure team, and projects arising from the JISC e-Infrastructure Community Engagement and Support Programme (see Section 3.1).
- Grid Technology Group at the STFC Daresbury Laboratory
- In liaison with AHeSSC, current and potential data grid applications in the arts and humanities.
- The iRODS development team at SDSC.

### *3.3 Liaison with other initiatives*

3.3.1 We will liaise closely with other related projects and groups (such as standards bodies), both within the JISC e-Infrastructure programme and elsewhere. In particular:

- We will participate in JISC programme-wide evaluation and support activities, in particular with a view to engaging the research communities from which potential users will be drawn.
- We will liaise with OMII-UK and the Grid Tools Liaison Group (comprising members of JISC, NGS and OMII-UK) regarding technical and other standards for middleware enhancements.
- Results and issues will be shared with NeSC, AHeSSC and the JISC e-Infrastructure team.

3.3.2 In particular we will collaborate with the NGS, which we have contacted about the proposed project. We will set up a prototype iRODS data grid at RAL, which is a core NGS node, as a test bed for user evaluation. A Shibboleth SP will also be installed at RAL. As this will be a non-production system, it will not be possible to make it available directly as an NGS service. However, we will make the test-bed available to targeted NGS users, who will be contacted and encouraged to use the system. In particular, we can make NGS users aware of it via the NGS wiki, which provides a placeholder for such development activities. Initially a small number of users will be targeted; as the project progresses the system will be opened up to a wider user base, and we will liaise with the UK e-Science Engineering Task Force (ETF) regarding pre-production evaluation.

### *3.4 Sustainability*

a) All specific software to be used (e.g. iRODS, Shibboleth, Fedora) is open source. All additional software developed will be made available to the wider community on an open-source basis, in accordance with the *Policy on Open Source Software for JISC Projects and Services*.
b) We will liaise with OSS Watch as regards the appropriate way of making the source available. Any iRODS enhancements and rules will be made available to the core iRODS project.
c) After the end of the project, STFC (at RAL) and CeRch (at KCL) will maintain the prototype for at least 12 months after project completion, with resources taken from their core budget. Indeed, it is very likely that they will continue to develop the infrastructure further after project completion.

[11] http://ahds.ac.uk/e-science/e-science-scoping-study.htm

*3.5 Dissemination and take-up*

a) Carry out all work in close coordination with the NGS and other bodies (see Section 3.3). NGS users, and other data grid users among the research community supported by STFC, will be actively encouraged to evaluate and comment on the work throughout and after the project.
b) Participate in related JISC workshops/programme-level activities.
c) Make all documentation publicly available, e.g. on the JISC, CeRch, STFC and project websites.
d) Present papers and posters at conferences and workshops, and submit papers to journals.
e) Keep in close contact with the core iRODS development team at SDSC.
f) Contribute to the iROD-Chat discussion list.

## 4. Budget

| **Directly Incurred Staff** | **Year <07-08>** | **Year <08-09>** | | **TOTAL £** |
|---|---|---|---|---|
| [redacted] | [redacted] | [redacted] | [redacted] | [redacted] |
| **Total Directly Incurred Staff (A)** | **£ 23,303.05** | **£ 94,231.99** | | **£ 117,535.04** |
| | | | | |
| **Directly Incurred Non-Staff** | | | | |
| Travel and expenses | £ 3,000 | £ 14,000 | £ | £ 17,000.00 |
| Hardware/software (new servers at CeRch – not charged to project) | £ 12,000 | £ | £ | £12,000.00 |
| Dissemination (not including travel) | £ | £ 1,000 | £ | £1,000.00 |
| Evaluation | £ | £ | £ | £ |
| Other | £ | £ | £ | £ |
| **Total Directly Incurred Non-Staff (B)** | **£ 15,000** | **£ 15,000** | **£** | **£ 30,000.00** |
| | | | | |
| **Directly Incurred Total (A+B=C) (C)** | **£38,303.05** | **£109,231.99** | **£** | **£147,535.04** |
| | | | | |
| **Directly Allocated** | | | | |
| Staff | £ | £ | £ | £ |
| [redacted] | [redacted] | [redacted] | [redacted] | [redacted] |
| Other | £ | £ | £ | £ |
| **Directly Allocated Total (D)** | **£5,111.26** | **£ 20,702.83** | **£** | **£ 25,814.09** |
| | | | | |
| Indirect Costs [redacted] | [redacted] | [redacted] | [redacted] | [redacted] |
| **(E) total** | **£20,088** | **£81,240.30** | | **£101,328.30** |
| **Total Project Cost (C+D+E)** | **£63,502.31** | **£211,175.12** | **£** | **274,677.43** |
| **Amount Requested from JISC** | **£34,923.5** | **£143,974.78** | **£** | **£178,898.28** |
| **Institutional Contributions** | **£28,578.81** | **67,200.34** | **£** | **£95,779.15** |
| | | | | |
| **Percentage Contributions over the life of the project** | **JISC 65.13%** | **Partners 34.87 %** | | **Total 100%** |

**Nature of Institutional Contributions**

| **Directly Incurred Staff** | | | | |
|---|---|---|---|---|
| [redacted] | [redacted] | [redacted] | [redacted] | [redacted] |
| **Directly Incurred Non Staff** | | | | |
| Hardware/Software etc. | £12,000 | £0 | | £12,000 |
| **Directly Allocated** | | | | |
| Staff, Estates etc. | £248.76 | £995.24 | | £1,244 |
| **Indirect Costs** | | | | |

| | | | | |
|---|---|---|---|---|
| Indirect Costs | £14,637.60 | £59,436.40 | | £74,074 |
| **Total Institutional Contributions** | **£28,564.76** | **£67,145.12** | | **£95,779.15** |

4.1 The estimated travel costs are relatively high because they cover 6 week-long visits to the iRODS development team at San Diego Supercomputer Centre, at an estimated cost of £10,000. The remaining £7,000 of the travel costs cover visits to IN2P3 in Lyon, travel between the project partners' sites for face-to-face meetings, and all travel related to dissemination activities, in particular visits to workshops and conferences (which are likely to be international).

## 5. Previous Experience of Project Team

5.1 STFC is contributors to the iRODS core development, in collaboration with SDSC. CeRch and STFC are working together closely on iRODS implementations and applications, particularly with regard to rules, integration with external systems, and preservation of complex digital objects (a paper describing their work has been accepted for the *3rd IEEE International Conference on e-Science and Grid Computing* in December 2007). They have been working closely with SDSC and with other early adopters of iRODS, e.g. IN2P3[12], and also have significant experience of other technologies related to the project, e.g. Shibboleth. Both STFC and CeRch are involved with production-level data grids based on SRB, and have a strong interest in facilitating the take-up of its successor iRODS.

5.2 ***Project Director**, 0.25 FTE, 15 months, based at CeRch, King's College London*
*Project management; technical management; production of use cases; acting as project advocate.*
Dr Mark Hedges is Deputy Director of CeRch & Technical Manager of the AHDS. He has extensive experience of software development, technical R&D, technical management & project management, gained from 17 years work in the IT industry. For the last 2.5 years, he has been manager of the AHDS technical infrastructure and software development, managing several projects, in particular involving data grids (SRB/iRODS), Shibboleth (from the JISC Core Middleware Programme) and digital repositories.

5.3 ***Senior Technical Officer**, 0.1 FTE, 15 months, based at STFC*
*Supervision of STFC Technical Officer; contribution to technical architecture; production of use cases.*
Dr Adil Hasan obtained a PhD in High Energy Particle Physics from Queen Mary College, London and has concentrated on data management and data distribution for the past 8 years. He has worked in the area of data grids for the past 6 years, has more than 5 years of experience using SRB, and is closely involved with the iRODS development at STFC.

5.4 ***Two Technical Officers**, 15 months. 1.0 FTE at CeRch/KCL, and 0.6 FTE at STFC.*
*The Technical Officers will be responsible for the technical/research review, software development and testing, production of technical reports and other documentation. These activities will be broadly divided into: (i) access control, (ii) access audit/provenance. However, this is unlikely to be an even split between the 2 partners, and a more granular division of work will be made as the project progresses. In any event, close collaboration between the CeRch and STFC staff will be necessary.*
CeRch is currently carrying out recruitment to complete its portfolio of core staff expertise. One of these new staff will be a Technical Officer/Software Developer whose remit will include data and access management issues. This staff member will be line managed by Mark Hedges in his role as Deputy Director of CeRch, and will be deployed to the project should it be awarded. The Technical Officer at STFC will be Roger Downing. Roger has worked in IT for over 7 years. He has considerable experience in distributed systems and has spent the last 3 years in the STFC e-Science department, working with the SRB. Roger has an extensive technical knowledge of the SRB software, and is currently involved in iRODS development and testing.

5.5 The project will be able to draw on the expertise of Dr Jens Jensen (STFC), Dr Tobias Blanke (AHeSSC), and Dr Simon Miles (KCL), who will provide *ad hoc* consultancy, Jens on Shibboleth, Tobias on arts and humanities applications, and Simon on provenance. Jens is STFC Management Liaison in the UK Federation, is a member of the STFC petabyte datastore group, which provides SRB services, and has provided steering on Shibboleth and other security projects. Simon is a Lecturer in the Department of Computer Science at KCL, has previously worked on the Provenance-Aware Service-Oriented Architecture (PASOA) project, and has numerous publications relating to data provenance.

[12] http://www.in2p3.fr/

**Information Services & Systems**

**Karen Stanton**
Chief Information Officer
& College Librarian

4th Floor
East Wing
Strand
London WC2R 2LS
Tel 020 7848 2737
Fax 020 7848 1777
Email karen.stanton@kcl.ac.uk

KING'S College LONDON

**University of London**

1 October 2007

To Whom It May Concern

I write to confirm, on behalf of King's College London, our full support for the proposal submitted by the Centre for e-Research (CeRch) entitled '**ASPiS (Architecture for a Shibboleth-Protected iRODS System)** ', under e-Infrastructure programme, Call: Federated Tools and Services, of JISC Circular 2/07 Call for Projects under the JISC Capital Programme.

This project is a significant initiative in managing access to research data and we recognise its importance to the HE community.

Yours sincerely,

Karen Stanton

Karen Stanton
Chief Information Officer & College Librarian
Information Services and Systems
King's College London

Science & Technology Facilities Council
Rutherford Appleton Laboratory

**Science and Technology Facilities Council**
Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Didcot OX11 0QX United Kingdom

Tel: +44 (0)1235 445000 Fax: +44 (0)1235 445808
www.scitech.ac.uk

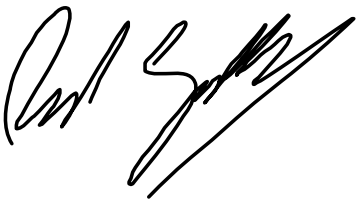
JISC
Northavon House,
Coldharbour Lane,
Bristol, BS16 1QD

October 1st, 2007

Dear Sir,

I am writing this letter of support for the **ASPIS (Architecture for a Shibboleth-Protected iRODS System)** proposal to the JISC Circular 02/07: e-Infrastructure (Federated Tools and Services). The proposed project to integrate the Rule Oriented Data management System (iRODS) with Shibboleth and storage of metadata information on the history of data objects (provenance) is of considerable interest within the e-science community. The iRODS system allows data management policies to be defined as rules that are executed on the data management servers at periodic intervals. In addition iRODS is flexible enough to allow new project specific policies to be incorperated into the system. Incorperation of Shibboleth will allow finer-grained role-based access of the data to be performed. The ability to store provenance information will greatly enrich the metadata information which will allow users to keep track of transformations applied to the data.

Yours sincerely



Neil Geddes
e-Science Director, Science and Technology Facilities Council

*SAN DIEGO SUPERCOMPUTER CENTER*

*9500 GILMAN DRIVE, MC0505*
*LA JOLLA, CALIFORNIA 92093-0505*
*TELEPHONE: (858) 534 5073*
*FAX: (858) 534 5077*

Centre for e-Research at King's College London
Arts and Humanities Data Service

**Letter of Support for the Proposal: ASPIS (Architecture for a Shibboleth-Protected iRODS System**
**JISC Circular 02/07: e-Infrastructure (Federated Tools and Services)**

Dear Sir:

I write in my capacity as Director of the Data and Knowledge Systems group of the San Diego Supercomputer Center (SDSC), and as principal investigator for development of the integrated Rule-Oriented Data Grid (iRODS) technology. I base my response on SDSC collaborations with the UK e-Science Data Grid, the Rutherford Appleton Laboratory, and the National Grid Service, all of which use the Storage Resource Broker data grid. Our interactions with these projects have been highly successful, and have enabled the use of data grid technology to broaden access to shared collections and promote collaborative research.

The current research and development activities at SDSC are focused on the execution of management policies that are applied to distributed shared collections. The iRODS system automates the application of rules that control the execution of remote micro-services. Rules can be defined that implement each management policy, whether related to time-dependent access controls, trustworthiness assessment criteria validation, integrity validation, data distribution, minimization of the risk of data loss, application of Institutional Research Board approval criteria or patient confidentiality mechanisms, and application of data analysis processes. The management processes are expressed as sets of micro-services applied at each remote storage location. This approach ensures that the management policies are enforced no matter which client is used to access the distributed shared collection.

The proposed integration of a Shibboleth authentication mechanism with iRODS, and the capture of audit and provenance information are highly desirable. These capabilities will enable integration of data management systems between the digital library, preservation, and grid communities. A central tenet in the development of the iRODS technology has been the creation of generic infrastructure that provides distributed data management for all applications. We strongly endorse the ASPIS project goals and the expertise of the members of the ASPIS team. The proposed development will benefit the broader community that plans to use the iRODS data grid.

Yours,

Reagan W. Moore
Director, Data and Knowledge Systems group
San Diego Supercomputer Center