Semantic Knowledge Underpinning Astronomy (SKUA) Case for support

Norman Gray and Tony Linde (University of Leicester)

21 June 2007

1 Executive summary

1 We propose the creation of a semantic infrastructure for astronomy based on the organisation of assertion services with relatively simple interfaces. Astronomy has been part of the UK's e-Science effort since its inception, the majority of this under the AstroGrid project. The focus of this effort, in the UK and within projects in at least 15 other countries, is the creation of a worldwide Virtual Observatory (VO), making astronomical data and applications easily available to astronomers regardless of their location and affiliation. The VO will, by defining and implementing standard interfaces, make it possible to access common resources from multiple applications. These resources are located via a globally distributed resource registry, which has been defined and working for over two years now.

2 To date, relatively little work has been done within the VO effort on semantic systems development. An ontology of object types has been developed by VO-France and one of us (Gray) has developed an access control system based on OWL inferencing and a mechanism for converting the standard VO registry XML format to RDF triples. Our project will provide a semantic infrastructure with toolkit and API which will make it possible for many more VO developers to engage with Semantic Knowledge Organisation Systems (SKOS). The key benefit of this proposal is that it engages with an existing vibrant development and user community, and builds upon working infrastructure, making it possible to demonstrate and prove both concepts and tools as we develop them. In doing so, we engage with key outcomes of the Capital Programme and its e-infrastructure programme.

3 The core concept of SKUA is that of a Semantic Assertion Collection (SAC). A SAC is a service combining an RDF triple store with an interface providing the ability to • store, modify and delete assertions (RDF triples); • return the result of SPARQL queries; and • optionally federate its queries to one or more other SACs. This simple extension to proven tools forms the basis of an infrastructure which supports federating tags and queries across multiple collections, covering perhaps a user's personal collection, that of a project they are working on, the department they belong to, and the worldwide VO. This allows for the construction of very personalised queries.

4 On top of this layer of capability, we will construct a few sample applications to demonstrate some of the additional functionality that it might provide. We expect other developers to build many more such examples. This layer and the SAC components will be packaged as a toolkit for these developers. In addition we will take part in JISC and astronomy meetings to promote the technology.

2 Background

5 The Semantic Web has, with startling speed, graduated from wild-eyed vision to deployable engineering. The goal of letting computers 'understand' has solidified into established practice and competing implementations, so that now, with the bleeding edge moving off into yet more exotic directions, is the ideal time to bring the core technologies to practical application. Europe has a world-leading role in the world-wide Semantic Web (SW) community, the fruit of years of heavy EC investment in the technology. The SKUA project will embed this expertise in a UK project, thus disseminating it from the UK to the worldwide VO community, and within the UK to the other metadata projects supported by the JISC.

6 The SKUA Project (Semantic Knowledge Underpinning Astronomy) will implement a distributed architecture of semantically aware RDF stores. This 'semantic layer' will support a cluster of applications which will either directly support users in finding and recovering useful resources, or indirectly support them by supporting user-facing applications. We describe the architecture and an initial set of applications below. Although the system we build will be specialised to astronomy, and proved by its interaction with, and eventual embedding within, the Virtual Observatory, the bulk of the semantic knowledge is localised in the RDF store, with the design goal that it could be replaced if desired by the analogous semantic knowledge of a different domain.

2.1 The astronomy knowledge cycle

7 Astronomy has a well-developed electronically-supported knowledge cycle (shown in Fig. 1), analogous in many ways to the JISC one. The astronomical cycle has data originating from multiple telescopes and archives (with sizes ranging from the very large to the informally maintained), and has long made use of institutional and community repositories, which are now being tied together with help from the international VO. The SKUA project aims to help close the circle, by creating a flexible infrastructure, allowing applications and users to annotate resources all over the VO, in a semantically meaningful fashion.

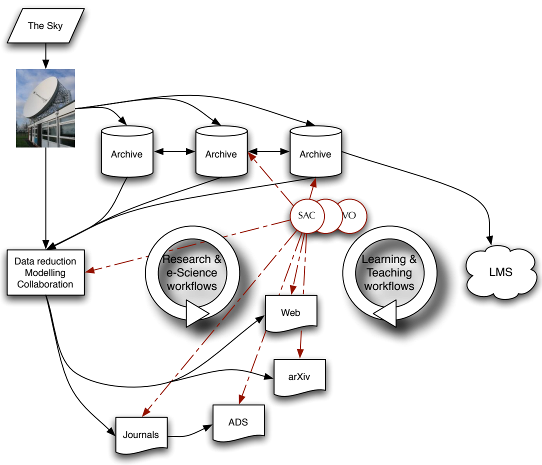

Figure 1: Astronomy knowledge cycle

2.2 The Virtual Observatory

8 The Virtual Observatory (VO) is a world-wide collaboration, supporting astronomical research through a network of projects to support data management, interoperability, portable workflows and common services. It is managed at the international level by the International Virtual Observatory Alliance (IVOA), acting as a standards body closely modelled on the W3C. The UK has a long-term leading role in the VO through the UK e-Science AstroGrid project, AstroGrid participation in the European VO Project, and the substantial UK investment in the European Southern Observatory (ESO), another Euro-VO partner. A primary focus of the various international VO projects is the continuing definition and maintainance of practical and internationally supported metadata describing archive data and web services; and one focus of the SKUA project is to add semantic value to the deployed VO metadata registries, aligned with ongoing VO efforts to develop ways of making these registries useful to astronomical applications.

9 The network of SACs described here interacts with the wider VO context by providing high-value services to VO client applications. The PIs are strongly connected with the VO's development plans,

and are in a position to react quickly to, and support, the needs of VO application authors.

10 The VO has an existing distributed registry service, containing metadata about large numbers of resources, from organisations and institutions, to large-scale data archive services. This registry is deployed already, in the form of a network of database-backed services. The global resource illustrated in Fig. 2 is seen as an RDF mirror of this registry.

11 The global VO has long recognised both the necessity and the complexity of shared metadata, and has made substantial time and software investments in the VO registry network described above. It recognises, however, that the problem is not yet completely solved, and is moving towards semantic solutions compatible with the solutions in this JISC proposal. This proposal therefore represents an opportunity to give a JISC-funded project a leadership role in the design of a component crucial to the infrastructure of the UK, European and world-wide VO efforts.

3 The SKUA architecture

3.1 Project architecture

12 The core component is a network of Semantic Assertion Collections (SAC) providing rather generic semantic Web Services. For performance reasons, we expect the semantic reasoning within the SACs to be rather simple, with more elaborate reasoning either performed in the background and separately asserted, or simply retained within value-adding clients. The optimal level of integration with, or even replacement of, the VO registries, will become clear during the course of the project.

13 This structure integrates with e-Infrastructure outcomes by supporting **new ways of retrieving data**, and by **integrating with key initiatives in the wider research community**.

14 We conceive the semantic layer as a directed acyclic graph (DAG) of SACs, each of which can store a greater or smaller number of RDF triples and, crucially, federate queries to a configurable list of partner stores, in such a way that a query against one SAC is effectively made against the RDF triples stored in that SAC and all the SACs that it federates to (Fig. 2). Thus the personal SAC, which may be a local desktop service or a personal section of a remote service, will typically store user-specific annotations or notes, and the global SAC will store VO-wide information such as an RDF mirror of the VO Registry. Information is transparently shared by being copied from a local SAC to an appropriate one of the SACs shared within a research group, or an ad-hoc group of collaborators, with this copy process being managed, directly by the user, using a small UI, or as a part of a separate user-facing application's functionality.

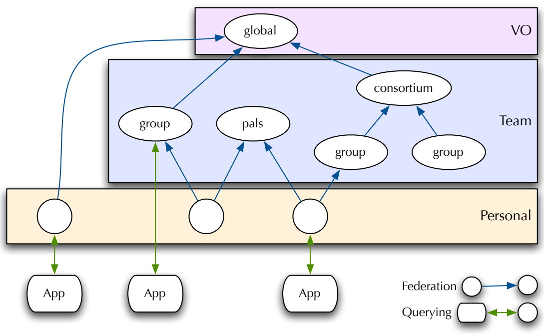

Figure 2: Architecture

15 Each SAC has a (standard) SPARQL endpoint which will respond to queries both from clients and from other SACs which federate to this one. Each SAC will also support a simple RESTful API for managing its RDF data.

16 A SAC must not respond to queries indiscriminately, since to do so would expose possibly private annotations; each SAC will keep a list of those SACs to which it has permitted federation. The topology of federations is specified exclusively by the SACs which do the federation; the permission to query or to write to a SAC is the responsibility of the SAC being federated to. The VO is deploying a SSO/Security infrastructure which this project would make use of. This infrastructure would handle the authentication issues involved, but we anticipate leaving the SAC access-control as the responsibility of the SACs themselves (either internally, or at the HTTP layer if appropriate).

17 We believe these three functions – querying, updating and sharing RDF information – will support a flexible and open-ended array of user-supporting client applications, and we will validate this assertion by developing an initial set of such applications, as described below.

18 The SKUA project uses standard standard technologies and protocols, composed in an innovative way. The SACs will build on one of multiple available triplestore implementations; they will be queried using the W3C-standardised SPARQL query language (`http://www.w3.org/TR/rdf-sparql-query/` – CR as at June 2007). The VO security infrastructure realises JISC investments by building on the Shibboleth infrastructure. The simple SAC management interface will be specific to the SACs, but there will be no requirement for this to go beyond the standard REST interaction pattern. Our goal is to produce a simple, open-source, and easily composable, Web Service, proved by applications. This builds on the PIs' experience with generations of application/service deployments in the VO and other projects.

3.2 Case studies and completion scenarios

19 The core of our proposal is the SAC architecture described in Sect. 3.1. The SAC servers will comprise a relatively thin layer on top of currently available triplestore technology, and so we do not expect the server implementation to be challenging.

20 Deployment and user buy-in will be at least as large a problem. The PIs have a long and continuing involvement in the VO community, and so can lead this deployment and react quickly to user requirements. However, user acceptance can be encouraged by producing exemplar applications, which illustrate how the architecture can be used, and which are independently useful. We describe two such applications here, which we will implement during the course of the project.

3.2.1 Tagging resources and sharing bookmarks

21 The most basic use of the SAC network, used by both of the applications below, and most immediately usable by existing user-facing applications, is to allow users to tag and bookmark resources on the web or within the VO (since tags and bookmarks are technically identical, and differ only in how they are used, we will talk only of tags below), and share those tags with other users. Web-2.0 services such as del.icio.us and Flickr have shown how very successful simple tagging can be, both to let users re-find resources they have found useful, and to be told of resources they had not found before. We can do better than simple tagging, however, since a tagging application can make use of the semantic context available from the SAC to suggest and interpret tags both when tagging and when querying. At least one existing VO application uses a private tagging framework, demonstrating that the demand is present.

3.2.2 Application: Spacebook – semantic VRE

22 As the name suggests, Spacebook has an interface and (liberal) sharing model styled on the very successful social software application, Facebook. In the case of Spacebook, though, individuals will be able to create and share queries, workflows and assertions about VO resources, in addition to supporting a professional/social network. In this, Spacebook will be a type of Virtual Research Environment (VRE) with additional semantic functionality. The VRE aspect will include portlets which embed components from the AstroGrid VO project including: query construction and submission, workflow construction and submission, virtual storage and jobs status; all these components are available now. Analogously with Facebook, Spacebook will have the concepts of Person, Institute, Group and Project, with Institute membership keyed to a user's institutional email address. Individual users may create Groups, and Spacebook administrators may create Projects.

23 **Scenario**: Claire logs into Spacebook and sees a summary of activities in all the areas to which she belongs including current status of long-running jobs that she has submitted. One such job was a complex workflow which has completed. She verifies the results are valid, tags the workflow script in order to describe it and then pushes the script into her project area [Spacebook will transfer

the script from Claire's virtual storage area to the project's, it will then pick up all the assertions in Claire's SAC associated with this workflow and push them to the project's SAC, with her agreement; this will also move assertions relating to the workflow's components]. In a blog she reads about a new paper published in her field so tags that for later reading [Spacebook adds assertions about the paper (via an arXiv URL), and passes the paper to a text mining tool which parses the paper for terms in the VO astro-ontology, her SAC and federated SACs, adding them to the assertions for that paper – Claire can review and change them when she later reads the paper]. She then moves into her Project area in Spacebook (where the workflow appears as a new item added). One of her colleagues has created a new version of a data analysis tool that implements an algorithm the project has developed. She makes this tool accessible to 'friends' in a Group specially created to test the tool [Spacebook copies assertions about the tool to the Group SAC]. Finally, Claire wants to execute a query that a colleague has placed in the Project area but over a different set of data sources. She begins typing into a search box; as she types each term, a graphical representation of associated terms appears with tags often cited together appearing closer. One term in the tag-cloud catches her eye as crucial and she clicks and adds this to her list of terms. In a window separate to the tag representation, a list of data resources appears and is refreshed as she enters each term [as she types, Spacebook conducts searches on each term or set of terms through Claire's personal SAC, the project SAC and all SACs to which these are federated; data resources associated with highly cited tags will appear on the resource list]. Claire picks the data sources she wants to use, submits her query and heads off for a coffee.

3.2.3 Application: Suggestions server

24 A continuing problem within the VO is that of browsing or searching the existing registries for resources of interest, since the obvious ways of doing this produce either too few, or far too many hits. The situation is improving with the arrival of better interfaces, but the semantically rich information available within the SAC network (the user's local SAC plus those it delegates to) would allow for richer query support. We have preliminary designs for a 'suggestions server', acting as a web service, which would take a list of one or more resources of interest, and return other sets of resources related to the initial ones by an open-ended set of algorithms, using semantic relationships, connections to existing astronomical controlled vocabularies, and statistical cluster analysis, implemented as plugins to the server.

25 **Scenario**: Jules is writing an application to help users find new VO resources. His user has already identified a few useful resources, and Jules would like to find more similar ones. He makes a simple query to a suggestions server, listing the known resources, and asking for 'more like this'; the server responds with groups of resources which are 'like' the initial set in various more-or-less heuristic ways, leaving Jules to display these to the user in whatever way best fits with his UI.

3.2.4 Other use-cases

26 Using NaCTeM tools, and other specialised text-mining tools developed with the VO, we can conceive of one or more SAC client applications deriving information from text sources and adding it to a personal or group SAC.

27 Another value-adding client application would be an access-control service, managing role and group information asserted within, and distributed amongst, the SACs. We have outline designs for such a service, which would build on the distributed nature of our semantic layer, but do not intend to implement it in this project, simply using it as one of the potential use-cases to drive the design.

4 Appropriateness to this call

4.1 Fit to programme objectives

28 The SKUA project is closely aligned with key outcomes of the Capital programme, providing **tools to support collaborative research**, by **improving discovery of, and access to, resources**. In addition, by consolidating current Semantic Web technologies and establishing a sustainable community of use (in VO developers), it is aligned with the vision of the **Knowledge Organisation and Semantic Services** theme of the e-Infrastructure programme.

29 The infrastructure and tools we are proposing will help **automate the research lifecycle** by helping users and applications find the resources they need either through semantic associations (if a resource is useful for quasars, then it is also useful for the more general class of compact objects) and through recommendations and tagging by colleagues. Through this means, shared assertions **facilitate collaboration** by taking advantage of the emergent **virtual organisations** represented by group SACs, and layering on top of the VO, seen as a worldwide virtual organisation.

30 This proposal gains its strength from being initially focused on astronomy, but within this there are multiple subdisciplines, and a major goal of the VO is to support **cross-disciplinary** data access (a headline use-case for the VO is to allow an X-ray astronomer to make reasonable use of radio data: such cross-field sharing has long been a notoriously difficult problem). Astronomy has for decades been naturally **cross-institutional**. The SKUA project will support these features by adding astronomy-wide semantics to the infrastructure, and supporting both institution-based 'group' SACs and per-project 'pals' SACs.

31 Astronomy makes extensive use of semi-conventional **metadata**, within the FITS files which contain the bulk of astronomical data. This metadata is typically not semantically enriched: the VO movement is pushing in this direction with standardised vocabularies/ontologies, formats and **registry descriptions**, but it is the SKUA project which will build on this to **create a semantic layer** by using SACs to **organise knowledge** and provide **semantic services**.

32 The SKUA project is an integration project, pulling together many of the emerging threads, and addressing many of the problems, which the proposers have identified in their years working with VO science and emerging VO practice.

33 We anticipate using existing and well-understood SW tools and technologies to implement our SACs, and will build on existing work (done by Gray) in converting the VO registries to RDF form. The proposers, and their VOTech network partners, have experience with high-performance triplestores, as well as with the Jena and Sesame APIs, giving experience of SPARQL, RDF, OWL and other standard Semantic technologies.

34 The SKUA project is motivated by the the existing VO (which is a VRE in this context), and will be proved through its **integration with this VRE**. As such, it benefits from both an **existing community of research end-users**, plus a network of application authors, ready to use and drive the high-value services the SKUA project will provide, providing **requirements and usability testing** through the existing VOTech network (indeed, the idea for the suggestions server described above arose from a specific present requirement of a Astrogrid applications author). The AstroGrid project is strongly usability-driven (enough that it has thrown away more than one prototype), and we will benefit from these high expectations. This integration will naturally **sustain** the SKUA deliverables.

35 Astronomical data is generally well-supplied with **metadata** (it is interoperable metadata that is the problem), and has well-established **communities** of instrument specialists (generating data), data archivists, consortia (concerned with rights management and sharing of data), and the authors of both general and specialist applications (consuming data), all of these working alongside the scientific analysts of the data. Although these specialisms are all within the context of astronomy, they represent a mature ecosystem of diverse and sophisticated interests, well versed in **information modelling issues** which bridge the various domains. The AstroGrid and VOTech projects cut across these domains, employing experienced and sophisticated developers in all areas. These projects

	Deliverable / Milestone	Due		Deliverable / Milestone	Due
1	Project plan completed	T1	8	1.0 Spacebook site	T12
2	Initial web site	T1	9	Final (1.1) Spacebook service	T18
3	Collaborative tech. deployed	T3	10	Prototype suggestion service	T9
4	Prototype SAC service	T6	11	1.0 suggestion service	T12
5	SAC API	T6	12	Plugins for suggestion service	T16
6	Final SAC service & client	T12	13	SKUA conference	T15
7	Prototype Spacebook site	T6	14	Final report on SKUA project	T18

Table 1: Project milestones. Code: T0: project start date (assumed to be 01-Oct-2007); Tn: n months into the project (eg, T4 is 01-Feb-2008); T18: project end date (31-Mar-2009)

and these domains provide an outstanding context in which to **bridge gaps** and **create the semantic layers** which will showcase JISC investments in world-leading technologies.

4.2 Value to JISC Community

36 The core value to the JISC community will come from the code developed by this project – semantic layer components and applications which make use of that infrastructure – all of it released under an OSI certified license. The more generic parts will be packaged as an easily deployed **toolkit** for use by developers working in astronomy and other fields to extend the e-Infrastructure with SKOS-type components. JISC will be able to promote our results, working within the existing Virtual Observatory infrastructure, as an **exemplar sematic layer** developed and deployed according to SOA principles.

37 We intend, as part of this project, to undertake extensive **dissemination** activities, not only of this project, but of the benefits of engaging with semantic technologies within normal tools development. We plan to run a special **workshop** on how semantic technologies might be integrated into the development of astronomical software. In addition to this, we will make presentations on our progress at AstroGrid workshops and will prepare **papers** for presentation at IVOA and ADASS meetings.

38 Complementary to these benefits to JISC will be **feedback** on the practicality of using existing semantic technologies in the field of astronomy and VO efforts. The VO has developed from the outset with a concern for SOA principles, so it will be critical to demonstrate that semantic technologies can adapt to the SOA-based architecture of a working set of services. In addition, we will make available to the JISC community via our website, all discussions and findings relating to choices of technologies.

4.3 Value to lead institution

39 By adding value to the VO, the SKUA project will benefit the University of Leicester in its support of the AstroGrid project, in which it is a lead partner.

5 Workplan

5.1 Project management

40 The project will be managed according to JISC standards and procedures. A project of this type, however, ought to benefit from agile techniques, so we will have a rapid prototyping and release cycle, ensure close user engagement at all stages of development, and revise plans and deliverables according to user feedback. This is the overall methodology used within the AstroGrid project, and both PIs are familiar with its advantages and risks. We would certainly recommend this approach and can discuss with programme management how it would fit into the JISC process.

41 Both of the project PIs are currently part of the VO community, and fully involved in consor-

tium and project strategy meetings, giving ample opportunity for formal and informal dissemination within, and feedback from, that community, both within the UK and internationally. Separately from the important ADASS conference, the IVOA organises a bi-annual set of interoperability workshops which would provide a vital forum for dissemination and feedback of the project.

42 In coordination with the JISC, we will produce a detailed project plan within the first month of the project. We have clear use-cases in mind, and vocal clients for them, so that the scoping required here can be very specific. See milestones in Table 1.

Work Package WP1 Project management

5.2 Phase 1: Architectural elaboration (T0-T12)

43 This is a long thin phase. After an initial block of activity, an initial version of the SAC network will be quickly 'delivered' to the case study developers, and effort shift to evolving it in the light of emergent requirements. This 'agile' methodology ensures that applications work can start very promptly, and that the services and API closely match emergent tool requirements.

WP2 Implementation of SAC nodes and network (time: 6sm)
Deliverable D2.1 Set up RDF version of VO registry
D2.2 Implement SAC, building on well-established triplestore implementation
D2.3 Develop thin user-facing client for simple SAC management and setup
D2.4 Produce initial API documentation for client authors

5.3 Phase 2: Case-study implementation (T3-T16)

44 The case-studies will be implemented in parallel.

WP3 Spacebook application (time: 12sm)
D3.1 Develop spacebook application, and refine interface through regular releases

WP4 Suggestion service (time: 12sm)
D4.1 Develop core service, with a plugin architecture
D4.2 Implement plugins implementing various heuristics

5.4 Phase 3: Dissemination (T8-T18)

45 The 'dissemination' activity will be more-or-less continuous, because of the continuous engagement of the PIs in the target developer community, and participation in project strategy meetings.

WP5 Documentation and dissemination (time: 3sm)
D5.1 Papers for appropriate semantic web and astronomy conferences
D5.2 Contributions to astronomy-specific publications (in particular IVOA Standards)
D5.3 SKUA final report, including discussion of applications outside astronomy
D5.4 Organise a workshop on the project outcomes; JISC disssemination/evaluation

5.5 Risks analysis

46 *Risk: Query federation difficult to implement (medium probability/medium impact)*. Mitigation: if true federation (decomposing and rewriting queries) proves more difficult than expected, then a simpler form – where SACs simply forward queries to their federation and combine the results – will provide most of the benefits. Conduct trials at outset of project, test existing technologies, identify potential solutions and engage with originating projects.

47 *Risk: Cannot integrate semantic technology with VO (very low/high)*. This is an integration project, combining existing technologies, all of which the project proposers are familiar with. Run early trials. Build on existing VO service standards.

48 *Risk: Lack of take-up of toolkit by astro-developers (low/medium)*. Engage with existing VO projects early and often. Build components useful to developers as well as applications. Fitting in with a larger project means that that project shares the work of technology evangelism.

49 *Risk: Lack of take-up of applications by astronomers (medium/medium)*. Build community of users

to act as beta testers and promoters to other users.

50 *Risk: Technology not sustainable (low/high)*. Demonstrate to AstroGrid and other VO projects usefulness of technology. Show technology to e-Science projects in other fields. Engage with OMII for possible take-up.

51 *Risk: Users and projects do not populate SACs (medium/high)*. Demonstrate usefulness of federated queries and tagging early in project. Use text mining and existing knowledge bases to populate SACs. UI design of Spacebook and of suggestion-server clients is crucial, so ensure that recruited developer has significant UI skills.

52 *Risk: Personnel changes or major requirements change (medium/medium-low)*. Because of the close relationship between this project and the AstroGrid and VOTech projects, to the extent that VO developers will work with early versions of the SKUA deliverables, it should be possible to contract staff from those projects in the event of unanticipated changes in staff or requirements.

5.6 Community engagement

53 The project has a range of stakeholders. The **VO software developers** have a primary interest (using the SAC network, and using clients of that network); These developers are vital, as the immediate users of the network: they must be engaged, and mined for use-cases. **Astronomers** are the primary stakeholders, with the obvious interest in using the resulting tools; they are ultimately vital, but their interest is mediated by the developers. The **IVOA, and other VO projects** have a broader interest (want to see increased use of VO and emergence of functional tools), but their interests follow in turn from the astronomers.

54 In parallel to this, the **Semantic Web and JISC development communities** have interests (seeing working technology examples; learning from practice), which will be addressed through the project's dissemination activities, and by working with JISC to embed the project's educative outcomes within the JISC communities.

6 Budget and justification

Short version:
(i) Budget: £310k
(ii) Justification: because we're worth it. . . .

7 Personnel

55 Linde has over 30 years IT experience, latterly in the development of web-based and service oriented architectures. Since 2001 he has been involved in UK and European e-Science activities, as Programme Manager of the PPARC-funded AstroGrid project (`http://www.astrogrid.org`) and Project Manager of the EU-funded EuroVO VOTech project (`http://eurovotech.org`). He has been a member of the JISC JIIE committee since 2003 and chair of both the JISC User's and Innovation (U&I) workgroup and OSSWatch advisory group. With the AstroGrid and VOTech projects he introduced the use of collaborative technologies – including forums, wikis, Jabber-based IM and online conferencing tools – all extremely useful given the highly distributed nature of e-Science projects. He successfully campaigned for the VOTech project to include a work package looking into resource discovery using semantic technologies. For three years he chaired the IVOA Registry workgroup, and has been active in the IVOA Semantics workgroup. His technical skills include a range of scripting languages and working with portals.

56 Gray has been from November 2005 a senior developer on the VOTech project, working on semantic resource discovery. Before that, he was a member of the UK Starlink project in the period 1998-2005, delivering a range of astronomical data analysis software in a highly distributed team. He brings experience of a wide variety of languages and systems, astronomical software engineering and design, an early and continuing involvement in the IVOA (Data Modelling, Grids and Semantics workgroups), and experience with SW technologies.