

Name of lead institution/organisation: University of Manchester	
List of project partners:	Southampton University Stanford University
Name of proposed project:	CO-ODE (Collaborative Open Ontology Development Environment)
Full contact details for primary contact	
Name:	Alan L Rector
Position:	Professor of Medical Informatics
Email:	alan.rector@man.ac.uk
Address:	Department of Computer Science University of Manchester Oxford Road Manchester M13 9PL
Tel:	0161 275 6188/6239/6248/7183
FAX:	0161 275 6204
Length of project: 2 years	
Project start date: 1 April 2003	
Total cost to JISC over life of project:	
Cost to JISC in each academic year (1 August – 31 July)	
Outline Project description	
<p>Metadata and ontologies have been widely identified as key technologies for E-Science. This project aims to integrate the most widely used tools for developing and using ontologies worldwide – OilEd from the UK and Protégé from Stanford in the US. This integrated open source toolset will be a user-oriented package designed to satisfy the needs of four use cases: a) Developing and maintaining large ontologies; b) Creating small local ontologies; c) Extending existing ontologies for specific uses; and c) Using ontologies to develop applications. It will support all three common paradigms – Frames, RDF(S) and the new Ontology Web Language OWL. A fundamental feature of the proposed architecture is its extensibility and flexibility. It will be designed specifically to be able to link to further tools being developed in other UK and EU funded projects.</p> <p>As well as supporting the UK scientific effort, the project aims to foster transatlantic cooperation on tools, content and standards. It is proposed to take advantage of recent parallel funding to the Protégé group from the National Cancer Institute and National Library of Medicine in the US</p> <p>By combining established techniques and building on tested software, the project aims at immediately achievable objectives for widespread deployment within the E-Science, Grid, and Higher Education communities. A companion proposal to the EPSRC will address longer term issues in user-oriented design, development, and management of logic based ontologies.</p>	
Names and contact details of any additional contacts:	
<p>Nigel Shadbolt, <i>University of Southampton, Department of Electronic and Computer Science</i> Highfield, Southampton, SO17 1BJ, UK, Tel 023 8059 4505 Fax 023 8059 3313, Email: nrs@ecs.soton.ac.uk</p> <p>Mark Musen, <i>Stanford Medical Informatics</i>, Stanford University Medical Center, 251 Campus Drive, MSOBX-215, Stanford, CA 94305-5479 Tel: (650) 725-3390, Fax: (650) 725-7944 Email: Musen@smi.stanford.edu</p>	

CO-ODE: Collaborative Open Ontology Development Environment Proposal to JISC under the Semantic Web Initiative

A. Introduction

Metadata and ontologies have been widely identified as key technologies for E-Science. This project aims to integrate the most widely used tools for developing and using ontologies worldwide – OilEd¹ from the UK and Protégé² from Stanford in the US. This integrated open source toolset will be a user-oriented package to satisfy the needs of four use cases: a) Developing and maintaining large ontologies; b) Creating small local ontologies; c) Extending existing ontologies for specific uses; and d) Using ontologies to develop applications. It will support all three common paradigms – Frames, RDF(S)³, and the new Ontology Web Language OWL⁴ (previously known as DAML+OIL). A fundamental feature of the proposed architecture is its extensibility and flexibility. It will be designed specifically to be able to link to further tools being developed in other UK and EU funded projects.

As well as supporting the UK scientific effort, the project aims to foster transatlantic cooperation on tools, content and standards. It takes advantage of a rare opportunity for synchronous funding in the US and UK afforded by parallel grants to the Protégé group from the National Cancer Institute and National Library of Medicine in the US⁵.

By combining established techniques and building on tested software, the project aims at immediately achievable objectives for widespread deployment within the E-Science, Grid, and Higher Education communities. A companion proposal to the EPSRC will address longer term issues in user-oriented design, development, and management of logic based ontologies.

B. Project Description

B.1 Aims, Objectives

- To provide the next steps to a user-oriented, scalable environment for domain experts to acquire, develop and use ontologies as part of the open source infrastructure for the E-Science and Semantic Web/Grid communities.
- To redevelop industrial strength services for ontology development based on the proven successes of the *de facto* editing environments for logic and frame based ontologies respectively, OilEd and Protégé.
- To integrate those services in a single platform linking the three paradigms – frames, description logics, and RDF(S) – taking advantage of existing plug-and-play facilities of Protégé to give access to a wide range of further open source tools.
- To bring together well established user communities based around the two sets of tools – OilEd and Protégé – in a single framework that can link to other developments in the UK and Europe, notably those being developed in the AKT IRC.
- To ensure that all tools are developed in cooperation with practical users.
- To facilitate transatlantic cooperation on the development of ontologies and E-Science and to take advantage of an unusual opportunity for parallel development which takes maximum advantage of the complementary skills in the two countries.

B.2 Background

B.2.1 The Problem

Expertise in ontology development is one of the UK's major strengths and a key feature of the E-Science programme. UK groups are prime movers in the new Web Ontology language OWL being developed by W3C. The OilEd editor (Bechhofer, Horrocks et al. 2001) developed at Manchester has become the *de facto* standard editing tool for DAML+OIL/OWL, despite having been designed only as a technical demonstration. However, these developments are threatened by their own success. The scale of efforts now being undertaken requires comprehensive user-oriented ontology development environments that builds on these prototypes and takes into account lessons learned from other large-scale ontology building efforts.

The proposed project will combine the two most successful ontology development environments in terms of numbers of users worldwide – OilEd from Manchester and Protégé from Stanford (Grosso, Eriksson et al. 1999; Noy, Sintek et al. 2001) in a cooperative project involving both institutions. It will be developed in close cooperation with key members of the user community taking into account lessons from other widely used environments such as DAG-EDIT⁶ in the

¹ <http://oiled.man.ac.uk>

² <http://protégé.stanford.edu>

³ Resource Description Framework (Schema) – see <http://www.w3.org/TR/2000/CR-rdf-schema-20000327/>

⁴ <http://www.w3.org/2001/sw/WebOnt/>

⁵ To be announced. Please contact proposers for further details.

⁶ <http://www.godatabase.org/dev/editor.html>

Gene Ontology Consortium and KAON⁷, developed out of the EU funded OntoWeb⁸ project, and tools developed in the Advanced Knowledge Technologies (AKT) IRC in the UK. It will use Protégé's plug-and-play environment to produce an extensible environment that can be quickly adapted to the widest practical range of needs.

The approach is based on analysis of users into four groups:

- a) New users needing to develop and use small, self contained ontologies;
- b) Users needing to make modest extensions to established ontologies to suite specific applications;
- c) Users needing to author and maintain or adapt large ontologies for widespread use – *e.g.* of cell function, anatomy, star types, imaging methodologies, etc.;
- d) Users needing to use ontologies in practical applications.

Regardless of which class users fall into, they must cope with seven steps:

1. Getting started: choosing an ontology development methodology, tools, and a starting ontology.
2. Developing the ontology using familiar intuitive notions and constructs including contingent information
3. Searching and navigating the ontology
4. Understanding and debugging the ontology
5. Extending, evolving, and maintaining the ontology
6. Access to external resources indexed by or linked to the ontology
7. Using the ontology to build applications

This proposal aims primarily at steps 1), 3), 6), 7) and at providing a framework for 2) while spanning the needs of users in all four categories a) – d). It aims to provide a stable but extensible architectural platform for further development. Issues 4), 5), and the more theoretical aspects 2) are addressed in a companion proposal to the EPSRC.

The issues chosen have been selected because they are well understood and prototype open source solutions exist or are nearing completion. This proposal will provide the effort required to re-engineer and disseminate these solutions in a flexible, integrated, and well engineered toolset.

Specifically, the proposed project will

- a) Re-engineer the underpinnings of existing OilEd tools;
- b) Integrate them with Protégé's plug-and play platform in an easily maintained, loosely-coupled architecture;
- c) Ensure that the resulting tools meet the needs of established user communities;
- d) Work with users to develop a library of reusable ontologies suitable to different communities.

This proposal seeks to take advantage of a unique opportunity: the availability of nearly synchronised funding on the two sides of the Atlantic. This will allow the project to take best advantage of the two countries' complementary skills: the UK in description logics and OWL, and the US in frame systems and the Protégé architecture.

This approach will enable the UK E-Science community and OWL community to avoid duplicating the development of numerous tools already available in Protégé – *e.g.* taxonomy comparators, a constraint language, and various visualisation techniques, links to key resources such as the UMLS/PubMed, Digital Anatomist Project, and tools being developed by the National Cancer Institute Center for Bioinformatics⁹. At the same time it will allow relatively easy extensions to re-implement or include proven techniques from other projects including the EU funded GALEN project and the IRC in Advanced Knowledge Technology – AKT¹⁰.

The project will draw on ongoing activities such as the EU-Funded WonderWeb. Research issues in providing a user-oriented environment for OWL hybrid representation paradigms are addressed in the companion proposal to the EPSRC. Eventually, the proposed activity might form the seeds of ongoing mechanisms for supporting the ontology and metadata requirements of the higher education and E-Science communities. However, in the short term, we propose concrete, clearly achievable objectives.

B.2.2 The tools to be integrated

OilEd is the *de facto* standard editor for the new Web Ontology language OWL being developed by W3C. It is widely used with over 1500 downloads at last count and an active mailing list. However, while remarkably successful, it was never designed to be extensible or to be a comprehensive ontology development environment. As a demonstrator, its features mainly reflect the structure of the OWL language rather than the natural work patterns of developers. Efforts are already being made to redevelop OilEd using resources from multiple existing projects, but without a concentrated effort and dedicated team, a timely result suitable for widespread dissemination is unlikely. A major outcome of this project will be to re-engineer the foundations of OilEd as modules and services suitable to be included in any subsequent application or development environment.

⁷ <http://kaon.semanticweb.org/>

⁸ <http://ontoweb.aifb.uni-karlsruhe.de/>

⁹ <http://ncicb.nci.nih.gov/>

¹⁰ <http://www.aktors.org/>

Protégé is a frame-oriented plug-and-play knowledge acquisition environment developed at Stanford University, and probably the most widely used frame editor today. It has been used successfully to develop ontology-based applications in fields as diverse as cancer protocols and military deployment (see <http://protege.stanford.edu>). Protégé-2000 has a worldwide user community with more than five thousand registered users and at least one thousand active groups, a number in the UK where its next annual meeting is scheduled to take place.

Protégé's great advantage and attraction to users is its proven flexibility and open plug-and-play framework. The active user community continues to provide third party plug-ins to support various additional techniques including standard relational databases, Prolog, constraint-based reasoning, ontology comparison and evolution, and a range of visualisation functions. Plug-ins have also been constructed providing links to key external resources including the Unified Medical Language System; it is the chosen tool for developing and delivering the Digital Anatomist Foundational Model of Anatomy; and it is currently a widely used means to access the National Cancer Institutes Ontologies and Terminologies thanks to translations provided by the Mayo Clinic.

Protégé-2000 can produce and edit RDF Schemas and knowledge bases, the foundational structures for the Semantic Web, and is in the process of being redeveloped in a web-based multi-user environment. However, although plug-ins exist to import and export in OWL syntax, there is no support for description logic based reasoning to organise the class hierarchy automatically as in OilEd. This functionality is increasingly required by users, particularly in the Semantic Web/Grid Communities. Integrating this functionality, even in the loosely coupled manner suggested here, requires significant extensions to the underlying Protégé architecture and knowledge model.

OpenGALEN – led by University of Manchester – has developed a collaborative methodology and toolset for the development of large scale description logic based ontologies (Rector, Zanstra et al. 1999; Rector, Wroe et al. 2001; Rector 2002). It treats the underlying description logic based formalism – be it OWL or some other – as an “assembly language” and allows users to work with “higher level languages” or views – usually known as an “Intermediate Representations” in deference to the knowledge acquisition community (Gaines and Boose 1988). This project will implement core support for this methodology as a plug-in for Protégé. Extensions to the view/intermediate representation mechanism are one of the topics of the companion proposal to the EPSRC.

Links to further tools – The involvement of Epistemics and Southampton University for the AKT consortium will ensure that the architecture provides the necessary links to the knowledge acquisition tools and large triple stores being developed in those environments.

B.3 Architecture and Methodology

B.3.1 Users and Requirements

There are several active communities using and developing ontologies and metadata in the UK. One is based around the E-Science projects focused on myGrid¹¹ and including AstroGrid¹², Geodise¹³, MIAS-GRID, the AKT and MIAS IRCs, and extending into international collaborations including OntoWeb and WonderWeb. There is a second group of projects centred around Health BioInformatics including: the National Cancer Research Institute (NCRI¹⁴) Cancer Therapeutics Ontology, the Prodigy Project¹⁵, the Gene Ontology Next Generation¹⁶, Multiflora, CLEF¹⁷, the NHS through both the National Electronic Library of Health¹⁸ and the emerging standard healthcare terminology SNOMED-CT, and incipient projects in large scale repositories such as BioBank. There is a third community around Digital Libraries, the Information Environment and Digital preservation and World Universities Network (WUN)¹⁹. The move towards making much more primary and secondary research data available through the Grid is likely to lead to a rapid increase in the need for special purpose ontology development and deployment. In the US the National Cancer Institute programme of ontology and terminology development is a significant keystone of its BioInformatics infrastructure.

“Vocabulary” is a central to developments in medical informatics, and the Mayo Clinic's medical informatics group, using Protégé, is committed to providing extended ‘coding and language’ resources linked to medical vocabulary. They are anxious to find a means to link it to OWL with full reasoning support as this proposal seeks to provide. Discussions with commercial collaborators in the associated projects are also in progress. These and many other groups already use either Protégé or OilEd, or both, for parts of their work. Almost all are committed in the medium term to moving towards OWL provided the tools can be made available. Through the various established user groups they provide manageable means for consultation and experimentation with prototypes and proof of concepts.

¹¹ <http://mygrid.semanticweb.org>

¹² <http://astrogrid.semanticweb.org>

¹³ <http://www.geodise.org/>

¹⁴ <http://www.ncri.org.uk/>

¹⁵ <http://www.prodigy.nhs.uk>

¹⁶ <http://gong.man.ac.uk/>

¹⁷ <http://www.clinical-escience.org/>

¹⁸ <http://nelh.nhs.uk>

¹⁹ <http://www.wun.ac.uk>

Increasingly, we expect that small-scale ontology development will start not from scratch but rather from existing libraries of resources – *i.e.* in terms of the analysis in 2.1, we expect a shift of users from a) to b). One of the great strengths of OWL is that, potentially, it makes extension and modularity much easier than with traditional ontology systems (Rector 2002), but to realise this potential requires tools designed for this purpose. OilEd brings the power of the OWL formalism. The combination of libraries of starting ontologies, plug-and-play resources from Protégé, and the simplification provided by Galen-style intermediate representations will bring the needed tools.

A further strength of the link to Protégé is its focus on application development alongside ontology development – *i.e.* to satisfy users in group (d). The proposed developments seek to make it easier to assemble bespoke ontologies in tandem with applications in order to ensure that the resulting ontologies fit the purposes for which they were intended.

Tracking the provenance and the evolution of ontologies requires further development in both OilEd and Protégé, but the Protégé metadata mechanisms allow classes to be treated as first class objects which is not possible in OWL. Protégé therefore provides the infrastructure for tracking changes. Furthermore, ontology evolution is a major research area in the Protégé community. There are already several plug-ins for tasks such as taxonomy comparison and the visualisation of differences between ontologies. A standard Dublin Core plug-in is also available. Synergy in research on ontology evolution is expected to be a significant benefit of the collaboration, and an axis of collaboration on this issue already exists between WonderWeb and the Protégé group via Michel Klein of the Free University of Amsterdam.

Both Protégé and OilEd already support RDF(S), and Protégé can be used as an editor for RDF(S) schemas (although a more familiar editing interface for RDF(S) would be a helpful addition to the plug-in library). With respect to web integration, Protégé's facilities are currently implemented via Java's RMI using specific downloads rather than using one of the newer web enabled protocols such as SOAP. The feasibility of conversion will be investigated but not guaranteed within the resources requested.

B.3.2 Specific Objectives

Concretely, the proposed project will

- Establish a programme of formative evaluation and user requirements analysis with identified user groups including the commercial sponsors.
- Produce a well engineered re-implementation and extension of the underpinnings of OilEd, independent of the specific interface to Protégé, ensuring support for combinations of functions integrated around user tasks rather than the underlying theoretical functionality
- Integrate those features as a plug-in for Protégé including presenting and rendering the interface within the Protégé framework.
- Complete and fully specify the bi-directional mapping between OWL and the extension of the OKBC standard that underpins Protégé, and provide a loosely coupled interaction between OWL and Protégé.
- Re-implement the basic *OpenGALEN* Intermediate Representation transformation mechanisms in Java and integrate them as a further linked plug-in to Protégé to mediate between users and complex ontologies
- Provide additional required functionality to the basic Protégé framework in order to meet the needs of OWL, most notably recursive embedding of slots and fillers, a library of schemas for managing provenance and other metadata, and the tracking and linkage mechanisms required for version control and browsing.
- Ensure that any linkage through plug-ins to outside resources are available both from the native Protégé frame representation and the OWL/Intermediate representations.

Specific modifications to the current OilEd mechanisms include:

- Providing incremental rather than monolithic access to the underlying classifier. At present, an entire ontology must be classified at once – a task that can take minutes or even hours. This makes experimenting with the effects of small changes to large ontologies impractical. This is a matter of changing the interaction with the underlying classifier rather than fundamental re-engineering.
- Integration of hierarchy visualisation and selection in a single view taking advantage of work being undertaken by Sun Microsystems
- Separation of user instigated changes from changes inferred by the classifier

Specific modifications to the current Protégé mechanisms include:

- Editing mechanisms for recursively embedded slot fillers
- Modifications to Protégé's OKBC derived knowledge model to account for OWL DL style qualifiers – 'some', 'all', etc.

B.3.3 Architecture

We propose a layered modular architecture, allowing reuse of almost all components in alternative environments. The proposed design has been extensively discussed. Requirements and outline design documents already exist for most sections.

The design is shown in Figure 1. There are two primary new plug in modules – shown in dark green: the OWL manager linked to the Reasoner through the standard DL API (orange) and the Intermediate Representation (View) manager. In addition, the initial web service interface to the tools from AKT and Epistemics will be specified. All will be engineered so as to be separable and re-usable in other environments, either individually or together. By taking advantage of the newly specified standard description logic interface, alternative reasoners can be plugged in, although the initial development will use the FaCT reasoner (Horrocks 1998)²⁰.

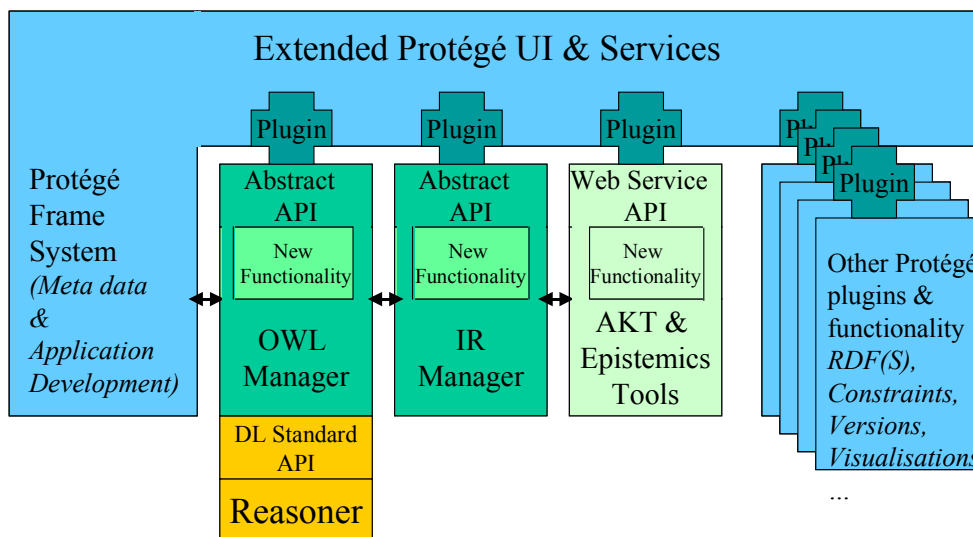


Figure 1: Simplified layered plug-and-play Architecture

The loose coupling between OWL and Protégé will be achieved by confining the functionality of the OWL reasoner to inferring the classification (subsumption lattice) for domain classes. All OWL classes will be instances of a specific metaclass in Protégé. The inferred lattice from OWL will be mapped back into the equivalent Protégé frame structure using a version of its Open Knowledge Base Connectivity (OKBC) model, but slightly extended in order to accommodate OWL's greater expressivity. Protégé's ability to treat classes as first class objects via its metaclass mechanisms will be used for metadata on the ontology itself, *i.e.* Dublin Core metadata, comments, explanations, provenance, history, etc. Protégé's mechanisms are highly tailorable and preliminary experiments suggest no problems. The overall tracking mechanism to link the three loosely coupled representations – Protégé's native OKBC, OWL, and the Intermediate Representation – will be expressed using RDF(S).

This JISC funded project will provide the OWL manager and a basic Intermediate Representation (IR) manager along with the corresponding abstract API layers (dark green in figure 1). It will provide the architecture within which other projects, including the companion proposal to the EPSRC can provide the "New Functionality" (light green inner boxes in Fig 1). Subsequent work is expected to extend the links to web service versions of functionality from AKT and Epistemics. The overall framework including modifications to Protégé itself – dark blue surround and stacked plug-ins – will be performed primarily by the Stanford team under separate funding, although a small subcontract to cover bespoke changes is requested.

B.3.4 Related projects and proposals

This proposal aims to provide a well engineered integrated piece of software for rapid deployment to the growing ontology community, who will provide experience and rapid feedback on functionality. A companion proposal to the EPSRC under the same call (HyOntUse) addresses the research topics of ontology debugging and transformation.

Other related efforts include earlier EU funded Ontoweb²¹ and current WonderWeb²² and KAON²³ architecture for RDF(S) resources and increasing support for DAML+OIL/OWL. Manchester leads Wonderweb and sees it as complementary to the integration with Protégé. Discussion is under way with several representatives of the Digital Libraries and Digital Preservation Communities who will be carefully consulted during the project.

For applications, Manchester has a central place in ontology development for numerous E-Science and related projects – see Sections D & E. Special effort will be paid to the Gene Ontology Next Generation Project and proposed developments in comparative anatomy – see companion proposal.

²⁰ <http://www.cs.man.ac.uk/~horrocks/FaCT/>

²¹ <http://ontoweb.aifb.uni-karlsruhe.de/>

²² <http://wonderweb.semanticweb.org/>

²³ <http://www.ontoweb.org/mailarchive/msg00006.html>

B.3.5 Software Engineering & Testing

A small team at University of Manchester in collaboration with the Protégé team at Stanford will develop the software. For re-engineering of the underlying software where the specification is well defined, it will use carefully managed software engineering principles – paired programming, predefined testing, exchange and review of code, etc. For the user interfaces, a more evolutionary prototype-and-test philosophy will be adapted during the first phase to establish usability, after which the focus will be on robustness and reliability. Insofar as possible the team will be integrated with the staff from related projects involved in specialised tool development - Geodise, myGrid, GOAT, WonderWeb, and the companion proposal to the EPSRC. However, whereas the members of each of these teams have specific objectives of which ontology development is only a part, the CO-ODE team will have as its focus the engineering and implementation of a robust ontology development environment for wide dissemination and re-use. Project management will be handled by the CLEF project manager who has extensive experience of managing industrial projects.

B.3.6 Deployment & Evaluation

The rapid and wide dissemination of both Protégé and OilEd provide strong evidence that the tools developed will find wide usage in many communities. The project is to build on these established user communities. Manchester and Southampton Universities' central roles in the UK E-Science/Grids programmes place them in ideal positions to foster wide dissemination.

Extensive ontology development activities occur within Manchester, both within the Computing Science Department and in projects linked to the National Genetics Research Laboratory and the Health and Safety Executive. The Department also conducts regular workshops and tutorials for other members of the academic and commercial community in ontology development and use. There is therefore wide opportunity for local evaluation and experimentation during development. There are several specific projects including AKT, and ontology work on Cancer Therapeutics, and cancer ontologies under the NCRI, who have agreed to act as early adopters and critiques of the work. Discussions are under way with industrial collaborators in CLEF and myGrid concerning requirements and deployment. The web content developer CSW who supports many of the NHS and British Medical Journal sites will take an active part in requirements analysis. Sun Microsystems is both contributing to the requirements and, it is hoped, to modules to be incorporated in the environment. Epistemics, the commercial partner of University of Southampton/AKT, who will have a subcontract from Southampton, will contribute to both requirements and evaluation. Network Inference Ltd, the spin off of Manchester for developing classification engines, will also participate in the evaluation and dissemination.

On a wider scale in the UK, both the requirements gathering and deployment will be organised in conjunction with the E-Science programme including other large E-Science projects such as AstroGrid and MIAS-Grid. Collaborations with the Digital Library community are under discussion. It is expected that workshops will be organised with the National E-Science Centre in conjunction with national metadata initiatives.

Internationally, the World Universities Network has expressed support and will also provide further links to the worldwide digital libraries community. A major byproduct of the proposed project is to foster collaboration with US groups including the National Library of Medicine and the National Cancer Institute Centre for BioInformatics (NCICB) who are funding the US part of the research and numerous groups in the medical community centred on the standards body HL7²⁴ and the Mayo Clinic. Major collaborative efforts on bioinformatics ontologies on comparative mammalian anatomy and physiology appear likely to emerge from the recent Standards and Ontologies in Functional Genetics meeting sponsored by the Wellcome Trust.

B.3.7 IP and Open Source Mechanism

The products of the project will be made available under open source license adapted from the Apache/BSD family on advice of NES/C/JISC through broader E-Science initiative. Actual ownership of intellectual property will be negotiated with the participants and JISC.

B.4 Benefits and Risks

B.4.1 Benefits

- Widespread availability of common tools for creating and disseminating ontologies. OilEd and Protégé have each already proved their worth as *de facto* standards for disseminating particular ontologies – e.g. Digital Anatomist, cancer ontologies from the NCI, and new work on the Gene Ontology.
- Immediate access to tools already implemented in the Protégé environment for ontology comparison and evolution, visualisation, links to the Unified Medical Language System and PubMed, Digital Anatomist, and National Cancer Institute resources, creation and editing of non-OWL RDF(S), and Dublin Core metadata, and to the rapidly growing body of plug-ins being provided by the world-wide Protégé community.

²⁴ www.hl7.org

- Increased input from open source developers around the world. Protégé already benefits from extensive input of plug-ins from outside Stanford. By providing a standard mechanism for extensions, developers in the open source community may more easily contribute while minimising the problems of supporting and integrating their work.
- Developing the community for OWL and logic based ontologies – already one of the UK’s major assets in the Semantic Web Community – and increasing the influence of UK in key standards bodies in Web and Grid Development
- Improved collaboration with key US partners

B.4.2 Risks

- Lack of uptake/poor focus on users – the project is designed to involve users and build on existing user bases.
- Staffing and recruitment – perceived as a modest risk on recent experience although recruitment can always cause delays. Some of the required staff are expected to be made available through redeployment.
- Technical incompatibilities between Protégé and OWL/OilEd – the relationship between the OKBC model used in Protégé and OWL has been looked at extensively in preliminary workshops between Manchester and Stanford and we are confident that the loosely coupled solution proposed is practical.
- Scaling difficulties of the resulting ontologies in OWL – the emphasis on the Intermediate Representation means that commitment to a particular representation in OWL for any given project can be deferred until the scaling properties of the chosen representation have been proven. The common DL interface means that alternative reasoning engines can be used if required.
- Being overtaken by other developments – the project is being conducted in close cooperation and discussion with members of the relevant W3C working groups and active participants in Grid and E-Science developments world wide. The proposers believe that the OilEd + Protégé approach provides the best approach to rapid deployment of the needed tools within the time scales of “Web development”.

B.5 Project plan

B.5.1 Work packages

- User Requirements and Evaluation: User requirements confirmation including survey of existing tools, consultation with commercial collaborators, and workshop/tutorial. (m 3) This will build on existing requirements analyses and design outlines requests for both OilEd, Protégé, and *OpenGALEN* as well as close links with current UK projects as described above. Checkpoint for user requirements against initial prototypes and design (m9), and beta testing (m 15), and dissemination workshops (m 21). This will also include presentations to the Protégé user group meeting in the UK
- Coordination with Stanford: Finalisation of partitioning of tasks, APIs, and mappings between Protégé/OKBC knowledge model and OilEd/OWL and the Stanford and Manchester teams. Joint meetings with Stanford Protégé team (m 2, m6, m12, m18) will be supplemented by access grid and phone conferences to maintain coordination. Additional issues (besides the introduction of OWL and support for a classification reasoner) will be harmonisation of handling of name spaces, metadata on ontologies, hooks for links to external sources, and mechanisms for handling non-OWL.
- Re-engineering of OilEd infrastructure with provision for a) incremental classification, (m6) b) change tracking (m9) c) hooks for modular inclusion of debugging aids and views, including mapping manager: to manage mappings between Protégé classes, OWL classes and Intermediate Representation. This will be built as part of the re-engineered infrastructure based on combining the OWL-OKBC mappings and the *OpenGALEN* mechanisms for handling the Intermediate Representation and will draw heavily on the Requirements work to establish the most convenient interface structures to enhance usability
- Development of the user interface and ‘tabs’ (plug-ins) for Protégé. The Protégé-specific user interface will be a thin rendering layer over the basic re-usable infrastructure. Key developments include modification of the Protégé slot widgets to allow recursive embedding (to be done by the Stanford team), a new hierarchy browser (to be adapted from open source work being developed at Sun), improved drag and drop mechanisms for ontology editing (Manchester) and implementation of the Dublin Core and other provenance mechanisms within the Protégé framework (Stanford).

B.5.2 Milestones and Deliverables

- Requirements analysis and ongoing monitoring mechanisms (m 6)
- Design and documentation (m 6)
- Beta version of integrated software (m 15)
- Final version of integrated software with manuals, tutorials, and support material (m24)
- Dissemination workshops and tutorials (m 21-m24)

C. Budget & Justification of Resources

Two software engineers/research associates are required for the two-year period. In addition, specific work required to meet UK requirements concerning OWL and the intermediate representation, plus support and training, will be commissioned in a subcontract with Stanford University. Experience has shown that collaboration only works well when there is specific support for the additional work at the collaborating centre, although the scale of this support has been reduced from original estimates because US funding of related efforts is now assured. For user requirements analysis and evaluation, and for ensuring that the architecture will support links to the software from the AKT consortium and Epistemics Ltd, a subcontract with Southampton University is requested, some of which will be further subcontracted to Epistemics Ltd.

Four joint workshops with the Stanford team are proposed over the life of the project. The intention is that these take place half in Manchester and half in California, but final decision on locations will depend on constraints and opportunities for the team. In addition, it is anticipated that one member of each team will spend an extended (4-6 week) visit with the other. To reduce travel costs, the initial meeting with Stanford will be combined with a visit to GSK in the US to check requirements. The agreement with Stanford for providing support for the project at a greatly reduced rate is that all funding for travel and workshops for both groups be funded from this project. The total dedicated contribution of Stanford staff to the project, over and above that already funded in the US, is estimated to be on the order of one staff year.

The requirements analysis and local travel within the UK are planned as a series of bilateral consultations and informal workshops at the beginning of the project and another more formal workshop, probably at NESC, towards the end of the project.

Manchester University Computer Science Department runs a dedicated Research Office, providing substantial project management and secretarial support to relieve the investigators of routine, non-technical project management tasks. The level of support provided goes far beyond what is normally covered by indirect overheads. The secretarial and administrative charges support this facility at a level commensurate with the size of the research project in question. The system management and technical staff are associated with providing computing support to the project. All networking and central computing resources are provided as part of the normal university support for research. However, PCs for research staff, and a laptop plus projector for dissemination workshops and requirements gathering are required.

D. Capabilities

D.1 Manchester

Manchester's Department of Computer Science is one of the largest in the UK with a 5* RAE rating. It is a leading player in the E-Science programme and a major centre for the development of ontologies and tools in the UK. The department provides an unusual mix of theoretical expertise, experience in developing software tools, and practical hands on experience in using tools to capture knowledge. FaCT, the first practical reasoning engine for expressive description logics and the key breakthrough that made modern ontology languages possible, was developed by Ian Horrocks in the Department. The Department has been instrumental in the development of the OWL language and has developed and implemented OilEd, the *de facto* standard OWL editing environment. Ian Horrocks sits on the W3C WebOnt committee responsible for OWL. Alan Rector led the GALEN projects responsible for the development of tools and techniques for medical ontology/terminology development using a previous generation of the technology. Members of the Department have run major workshops on ontologies and ontology development at the Global Grid Forum, the Intelligent Systems for Molecular Biology conference, the International Semantic Web Conference, the American Medical Informatics Association Fall Symposium, and numerous local and private tutorials including for the Mayo Clinic.

In addition to the links listed under key personnel (E) below, Manchester hosts MIAS IRC – from Medical Images to Clinical Knowledge and has close links with Manchester Computing which is an active participant in Grid activities and a major player in MIMAS via Julia Chruszcz.

The Heath & Bio Informatics forum has long established relations with major ontology and terminology groups in the US including the Mayo Clinic, the major standards body HL7, and the National Cancer Institute Center for BioInformatics, which is funding much of the Stanford part of the project. Members of the Department are involved in metadata and ontology development for fields as diverse as fluid dynamics for aircraft design (GeoDise) and art history (STARCH). The CoHSE architecture for conceptual hypermedia pioneered the use of advanced ontologies in Web design.

D.2 Stanford

Stanford Medical Informatics is one of the world's leading research centres in knowledge acquisition and representation, with a history of landmark projects dating back to the original Mycin and now generalised in a family of projects around the Protégé plug-and-play environment, a key piece of infrastructure for many projects developing and using ontologies and with an established worldwide user group of several thousand.

D.3 Southampton Department of Electronics and Computer Science/Epistemics Limited

Southampton leads the Advance Knowledge Technologies IRC and has a long track record of developments in knowledge acquisition and web technology. The spin-off company, Epistemics²⁵, which will have a subcontract from Southampton for requirements gathering and testing is one of the UK's leading providers of knowledge management services. Both are members of the *my*Grid and Geodise consortiums. Both use both OWL and Protégé.

E. Key Personnel

E.1 Alan Rector

Alan Rector has pioneered the use of ontologies in practical applications in biomedicine in the PEN&PAD, *OpenGALEN*, and UK Drug Ontology sponsored by the MRC, EC, and DoH respectively. He now leads the Clinical E-Science Framework (CLEF) project sponsored by the MRC. He is a member of the CEN and ISO Technical Committees on medical terminology, and has consulted for the NHS Information Authority, Mayo Clinic, Network Inference, and Hewlett Packard. He is Professor of Medical Informatics in the University of Manchester Department of Computer Science and leads the Health and Bio Informatics Forum there. He is a member of the JCSR and the NHS/HE Forum and of the bioinformatics board of the NCRI.

E.2 Carole Goble

Carole Goble is Professor of Information Management in the Department of Computer Science, University of Manchester. Recently she has been heavily involved in the UK e-Science Grid initiative and is director of one of the largest EPSRC pilots (£4million over three years), *my*Grid, which aims to build a personalisable platform for *in silico* experiments for biologists. *my*Grid applies Semantic Web technologies to Grid services. Professor Goble is the co-Director of the e-Science North West regional centre and sits on the UK Office of Science and Technology eScience Programme Steering Committee. Other eScience grants she holds where she is applying Semantic Web technologies include Geodise, for optimising engineering designs. She is one of the prime movers behind DAML+OIL/OWL, and is the co-chair of the Global Grid Forum Research Group on Semantic Grid, dedicated to bringing Semantic Web technologies to the Grid community. She is an Editor in Chief of a new Elsevier Journal of Semantic Web and chaired the first Semantic Web track at the WWW2002 conference

E.3 Ian Horrocks

Ian Horrocks is a Reader in Computer Science at the University of Manchester, UK. His FaCT system revolutionised the design of Description Logic (DL) systems. He has been involved in numerous national and international research projects including Camelot, DWQ, OntoWeb and DAML, and is the coordinator of the EU IST WonderWeb project. He has published widely in leading journals and conferences, winning the best paper prize at KR'98. He was the program chair of the 2002 International Semantic Web Conference and is the Semantic Web track chair for the 2003 World Wide Web Conference. He is a member of the Joint EU/US Committee on Agent Markup Languages, the W3C Web Ontology Language working group, and is a prime mover in and editor of the OIL, DAML+OIL and OWL language standards

E.4 Nigel Shadbolt

Nigel Shadbolt is professor in the Department of Electronics and Computer Science at Southampton University and leader of the Advanced Knowledge Technologies (AKT) IRC. His research concentrates on two ends of the spectrum of AI - namely, Knowledge Technologies and Biorobotics. He is Editor in Chief of IEEE Intelligent Systems, an Associate Editor of the International Journal for Human Computer Systems and on the editorial board of the Knowledge Engineering Review.

E.5 Mark Musen

Mark A. Musen is Professor of Medicine and Computer Science at Stanford University and head of the Stanford Medical Informatics laboratory. He has directed the Protégé project since its inception in the 1980s, and was the 1989 recipient of the Young Investigator Award for Research in Medical Knowledge Systems from the American Association for Medical Systems and Informatics. He received a National Science Foundation Young Investigator Award in 1992 for work on Protégé. He has been elected to the American College of Medical Informatics and to the American Society for Clinical Investigation. He has served on the Biomedical Library Review Committee of the United States National Library of Medicine as well as on numerous editorial boards and program committees.

E.6 Key Research Staff

- Dr. Sean Bechhofer – developer of OilEd
- Drs. Chris Wroe & Jeremy Rogers – knowledge engineers on the GALEN, UK Drug Ontology, and Gene Ontology projects

²⁵ <http://www.epistemics.co.uk/>

F. References

- Bechhofer, S., I. Horrocks, C. Goble and R. Stevens (2001). "OilEd: a Reason-able Ontology Editor for the Semantic Web". *KI2001, Joint German/Austrian conference on Artificial Intelligence*, Vienna, Springer-Verlag: 396--408.
- Gaines, B. and J. Boose (1988). *Knowledge Acquisition for Knowledge-Based Systems*. New York, Academic Press.
- Grosso, W. E., H. Eriksson, et al. (1999). "Knowledge modelling at the millenium (The design and evolution of Protege-2000)". *Knowledge acquisition workshop (KAW-99)*, Banf, Canada.
- Horrocks, I. (1998). "Using an expressive description logic: FaCT or Fiction". *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixth International Conference on Knowledge Representation (KR 98)*, San Francisco, CA, Morgan Kaufmann: 634-647.
- Noy, N. F., M. Sintek, et al. (2001). "Creating semantic web contents with Protege-2000." *IEEE Intelligent Systems* **16**(2): 60-71.
- Patel-Schneider, P. F. (2002). "Two Proposals for a Semantic Web Ontology Language." *2002 International Description Logic Workshop.*, Toulouse, France, April 2002.
- Rector, A. (2002). "Normalisation of ontology implementations: Towards modularity, re-use, and maintainability". *Workshop on Ontologies for Multiagent Systems (OMAS) in conjunction with European Knowledge Acquisition Workshop*, Siguenza, Spain.
- Rector, A., C. Wroe, J. Rogers and A. Roberts (2001). "Untangling taxonomies and relationships: Personal and practical problems in loosely coupled development of large ontologies". *Proceedings of the First International Conference on Knowledge Capture (K-CAP 2001)*, Victoria, BC, Canada, ACM: 139-146.
- Rector, A. L., P. E. Zanstra, et al. (1999). "Reconciling Users' Needs and Formal Requirements: Issues in developing a Re-Usable Ontology for Medicine." *IEEE Transactions on Information Technology in BioMedicine* **2**(4): 229-242.