EPSRC
Polaris House, North Star Avenue, Swindon, Wiltshire,
United Kingdom SN2 1ET
Telephone +44 (0) 1793 444000
Web http://www.epsrc.ac.uk/

Engineering and Physical Sciences Research Council

COMPLIANCE WITH THE DATA PROTECTION ACT 1998

In accordance with the Data Protection Act 1998, the personal data provided on this form will be processed by EPSRC, and may be held on computerised database and/or manual files. Further details may be found in the **guidance notes**

Network PROPOSAL

Document Status: With Council
EPSRC Reference: EP/F064462/1

e-Science Networking 2007

Organisation where the Grant would be held

Organisation	STFC - RAL	Research Organisation Reference:	e-sci for o-sci
Division or Department	ISIS Pulsed Neutron & Muon Source		

Project Title [up to 150 chars]

The Open Practises E-science Network

Start Date and Duration

a. Proposed start date	01 October 2008	b. Duration of the grant (months)	36

Applicants

Role	Name	Organisation	Division or Department	How many hours a week will the investigator work on the project?
Principal Investigator	Dr Cameron Neylon	STFC - RAL	ISIS Pulsed Neutron & Muon Source	1.88

Objectives

List the main objectives of the proposed research in order of priority [up to 4000 chars]

The aims of the network are to provide a forum for identifying, discussing, and implementing solutions to the challenges noted above. This will be achieved through agreeing standard and definitions, aiding in the development and integration of specific tools, identifying and promoting examples of good practise, and promoting the benefits of the open research approach to the wider research community. The aim overall will be to widen the applicability of tools and practises as far as is possible with the resources available. The specific aims of the network will be to;

* To articulate, define, and promote the principles of properly archiving and sharing laboratory information;
* To agree a framework for identifying, and enabling the aggregation of, primary research data made available online;
* To identify practical routes towards making both raw and derived data self-describing and semantically rich;
* To publicise and promote to the research and data management communities the ethos of making research data freely available;
* To identify and implement best practise in licensing and publication of data;
* To identify and enable good practise in reconciling the twin demands of obtaining a financial return on research results and making those results freely available.

Summary

Describe the proposed research in simple terms in a way that could be publicised to a general audience [up to 4000 chars]

The web makes it possible to store and share enormous amounts of data and there are a wide range of tools availablle to help people do this, from Wikipedia to Facebook to Del.icio.us. Yet despite this the practise of scientific research remains stuck in a world where a printed page of a scientific journal that contains, at best, a summary of the details of a study, is regarded as the only important scientific document of record. This leads to an enormous waste; of the results of experiments that don't work, of experiments where another group manages to publish first, or simply results that aren't quite good enough. In fact probably the majority of research is never shared.

The web makes it possible to share all research results, as they happen. This makes it possible to re-analyse results to look for new effects, to re-check the analysis of a specific experiment, or simply to make sure that someone doesn't waste their time repeating an experiment that doesn't work. Open Research advocates beleive that this offers a better way of doing science; that it offers the potential for carrying out science in a way that is more responsive to current needs, more rapid, more effective and more efficient.

However, while the tools for sharing data are well developed in social networking sites, similar tools for scientists are still in their infancy. We propose to form a network that will examine both the needs for specific tools and identify routes towards developing them. These tools will ultimately be used by all researchers to ensure that the results of research are properly archived for the future. The network will also promote the proper archival of research results as an effective means of ensuring the most efficient long term returns on research funding.

Beneficiaries

Describe who will benefit from the research [up to 4000 chars].

The beneficiaries of this network proposal can be broady divided into four categories with the first being the most immediate and local beneficiaries and the fourth seeing the benefit of the potential increases in research efficiency over the long term:

1) The first and most immediate beneficiaries will be the international Open Research community who will benefit from having a forum and conference in which to drive forward the agenda. The opportunity for focussed discussion on the issues involved in capturing and sharing generic laboratory data has not previously existed.

2) The general research community that interacts with and adopts some of the tools and practises developed and implemented by the Open Research Community will see the benefits arising from the sharing and commenting of data including the more effective conversion of research into publications, higher citiation rates, and generally better science.

3) Research funders will benefit from a more efficient return on their investment with higher research productivity (as measured through publications) and a growing store of properly archived and annotated research data that will improve in size and quality as time progresses.

4) The general community, including commercial enterprise and technology based industry will benefit through more effective use of limited research resources, greater productivity in the research community generating more commercially relevant results, and the availability of a huge store of publically funded information that can be mined for further insights.

Summary of Resources Required for Project

Financial resources

Summary fund heading	Fund heading	Full economic Cost	EPSRC contribution	% EPSRC contribution
Directly Incurred	Staff	3888.00	3110.40	80
	Travel & Subsistence	62000.00	49600.00	80
	Equipment	0.00	0.00	80
	Other Costs	18500.00	14800.00	80
	Sub-total	**84388.00**	**67510.40**	
Directly Allocated	Investigators	9142.63	7314.10	80
	Staff	0.00	0.00	80
	Estates Costs	1066.00	852.80	80
	Other Directly Allocated	0.00	0.00	80
	Sub-total	**10208.63**	**8166.90**	
Indirect Costs	Indirect Costs	5840.00	4672.00	80
Exceptions	Staff	0.00	0.00	100
	Other Costs	0.00	0.00	100
	Sub-total	**0.00**	**0.00**	
	Total	**100436.63**	**80349.30**	

Summary of staff effort requested

	Months
Investigator	1.75
Researcher	0
Technician	0
Other	1.75
Visiting Researcher	0
Student	0
Total	3.5

Other Support

Details of support sought or received from any other source for this or other research in the same field.

Awarding Organisation	Awarding Organisation's Reference	Title of project	Decision Made (Y/N)	Award Made (Y/N)	Start Date	End Date	Amount Sought/Awarded (£)
Biotechnology and Biological Sciences RC	BBD00652X1	Grid compatible data management for Directed Evolution Experiments	Y	Y	01/03/2006	28/02/2008	60076

Staff

Directly Incurred Posts

						EFFORT ON PROJECT					
Role	**Name /Post Identifier**	**Basic Starting Salary**	**Scale**	**Increment Date**	**Start Date**	**Period on Project (months)**	**% of Full Time**	**London Allowance (£)**	**Super-annuation and NI (£)**	**Total Other Allowances (over period of appointment) (£)**	**Total cost on grant (£)**
Other Staff	Admin Support Staff	25919	N/A	01/10/2008	01/10/2008	36	5	0	0	0	3888
										Total	3888

Applicants

Role	Name	Post will outlast project (Y/N)	Contracted working week as a % of full time work	Total number of hours to be **charged** to the grant over the duration of the grant	Average number of hours per week **charged** to the grant	Rate of Salary pool/banding	Cost estimate
Principal Investigator	Dr Cameron Neylon	Y	100	248	1.9	60828	9143
						Total	9143

Travel and Subsistence

Destination and purpose		Total £
Within UK	Travel and subsistence for prize winners (annual meeting x 3)	9000
Within UK	Support for network members to attend annual meeting (x 3)	21000
Outside UK	Support for exchanges (two way, x2 each year)	30000
Outside UK	Travel costs for steering committee meeting	2000
	Total £	62000

Other Directly Incurred Costs

Description	Total £
Small hardware items to support specific development projects	5000
Server, hard disks, and backup storage	3000
Annual meeting costs - room hire	4500
Annual meeting costs - audiovisual support (including recording)	4500
Annual meeting costs - technical support	1500
Total £	18500

The Open Practises E-science Network:
A research network to enable data sharing in the real world

1 Track Record

The group assembled to form the core of this network are leaders in the promotion and practise of Open Research. This is a young field as well as a generally very young community. Many of the tools that make Open Practise feasible are only a few years old and developments in many groups are proceeding at a great rate. The core group of the network is made up of senior academics, more recently appointed tenured staff, PDRAs and postgraduate students, as well as developers, publishers, and hobbyists. In many cases these categories are not mutually exclusive. There is not an extensive record of publication in the specific area of Open Research because the development of these ideas is so new. The nature of the open approach generates very rapid development of ideas and research programmes and many of the references within this proposal are to Blog posts and web pages. Where this relates to original research that is not yet published in a peer reviewed journal it is not due in any way to a lack of quality in that research but simply that there has not been time for this to be published. There is good evidence that open practises lead to more highly cited papers both with respect to open access to the published research [1] and providing detailed research data [2]. The group assembled here are at the forefront of exploring and optimising this process.

1.1 Network coordinator

Cameron Neylon *(STFC Rutherford Appleton Laboratory)* is a biophysicist who has always worked in interdisciplinary areas. After undergraduate studies in metabolic biochemistry he has pursued research in molecular biology, biophysics, and high throughput methods. In 2001 he took up the position of Lecturer in Combinatorial Chemistry at the University of Southampton and in 2005 he commenced a joint appointment (80%) as Senior Scientist in Biomolecular Sciences at the ISIS Pulsed Neutron and Muon Facility. Dr Neylon is a key contributor to the 4G Basic Technology Programme (£5.1M). Through this and other projects he has gained extensive experience of the challenges of working within and managing complex multidisciplinary research programmes with recent papers in journals as diverse as *Cell, Nature Physics, Complex Systems,* and *Journal of Combinatorial Chemistry*. In 2005, in collaboration with Jeremy Frey (University of Southampton) he obtained funding to develop and optimise an electronic notebook system for biochemistry laboratories which has lead to his involvement in the Open Research movement. His group is currently moving to a fully Open Notebook approach which is being recorded and analysed in his Blog, Science in the Open. He has given three invited keynote lectures at national and international meetings since 2004. He recently gave invited talks on Open Research at Drexel University, Philadelphia and as part of the International Genetically Engineered Machines workshop at MIT, Boston. He has managed several workshops and conferences including Neutrons in Biological and Biomedical Sciences (2006) which directly lead to an increased uptake of the ISIS Neutron Scattering facility by biological scientists and Neutrons in Biology (2007) a satellite meeting of the European Biophysical Societies Association meeting with 50 attendees from Australia, Japan, and the US. He also coordinates the recently STFC funded Research Network for Biomembrane Structure and Function.

1.2 Founding Network members

Jeremy Frey is Professor and Head of Structural and Materials Chemistry in the School of Chemistry at the University of Southampton. He was principal investigator on the Comb*e*chem e-science project and Southampton e-science Platform Grant. His group has developed the Chemtools platform which includes a variety of electronic laboratory notebook approaches as well as blogging instruments, now being expanded to the blogging laboratory.

Peter Murray-Rust is Professor in the Unilever Centre for Chemical Informatics in the School of Chemistry at the University of Cambridge. He is known internationally as an advocate of Open Practise in science and Open Access publishing. His group has developed many of the key tools to enable open practise in chemistry including Chemical Markup Language, OSCAR, and Crystal-Eye.

Jean-Claude Bradley is Associate Professor in the Department of Chemistry, Drexel University, Philadelphia. He is best known for defining and beginning the practise of Open Notebook Science [3]. His work in this area has been widely covered by the consumer and science and technology media. He is also a leader in the use and development of online tools for research and teaching including Wikis, Blogs, and Virtual Worlds and is the coordinator of the regular 'SciFoo Lives On' poster sessions held on Nature Island in Second Life [4].

Nicolas Bertrand is e-Science consultant for the Environmental informatics programme at the Centre for Ecology and Hydrology (Natural Environmental Research Council). His expertise lies with data management, bioinformatics and grid computing. He setup a project to enable scientists to capture and share their daily scientific activities using web2.0

technologies (Blogs and wikis). He has strong interests in open science as a mean to improve collaboration, data sharing, replicability of experiments and overall quality of scientific research.

Bill Hooker is currently a postdoctoral research fellow and will soon be seeking a faculty position. He has written widely on the web on Open Research and is a highly regarded commentator on open practise and its potential benefits at his Blog, Open Reading Frame [5]. He has a regular guest blog spot on the highly popular 3 Quarks Daily [6], where he has written a series of articles on Open Science [7-9].

OpenWetWare [http://openwetware.org] is an effort to promote the sharing of information, know-how, and wisdom among researchers and groups who are working in biology & biological engineering. is a collaborative web resource that provides biological researchers an online platform for storing, managing, and sharing primary and preliminary research data and know-how. By fostering the growth of online research communities, OpenWetWare has already enabled the organized capture of information and knowledge that is otherwise not stored electronically or disseminated. OWW provides a place for labs, individuals, and groups to organize their own information and collaborate with others easily and efficiently. The aim is that OWW will not only lead to greater collaboration between member groups, but also provide a useful information portal to colleagues, and ultimately the rest of the world.

Mat Todd is lecturer in Organic Chemistry at the University of Sydney and community leader for the Schistosomiasis Research Community at Synaptic Leap [10]. His research in antiparasitic drug discovery for neglected tropical diseases had lead to an interest in the development of approaches for carrying out synthetic chemistry in an open environment. He has interests in how to best share synthetic data and arrange it such that scientists, not computer scientists, can collaborate most effectively in a growing, iterative experimental project.

ChemSpider [http://www.chemspider.com] is a free access website established with the intention of "Building a Structure Centric Community for Chemists". ChemSpider was released to the public in March 2007 and is presently accessed by about 2000 individual users per day for the purpose of researching information associated with chemical structures, for the purpose of transaction-based property predictions on a chemical structure and for the purpose of searching Open Access chemistry articles. A collaboration with Jean-Claude Bradley has lead to the further development of online structure deposition system, spectral deposition and structure-activity deposition systems to support Open Notebook Science.

Pedro Beltrao has just completed his PhD at EMBL-Heidelburg and is soon to start a postdoc at UCSF. He has been using a blog [11] to share ideas about Open Science generally and his early research results in particular for over several years and aims to introduce his new lab to some of the tools and ideas he has been developing. He has directly seen the benefits of an open approach through the use and further development of bioinformatics tools which he developed and shared online.

Oliver Hofmann is a research fellow at the Harvard School of Public Health, working on the integration of genome-wide studies with experimental data for biomarker discovery. He is the project leader of the eVOC ontology system (http://evocontology.org, [PMID 12799354]) and maintains the OBO Cell Ontology (CL, [PMID 15693950]). His work includes the collaborative development of standard annotation formats in the biomedical field, including the MIACA project (http://miaca.sourceforge.net/) and a reasoner-based approach to alternate transcript annotation.

Ricardo Vidal is a graduate student at University of Algarve (Portugal), close to completing his Masters (M.eng.) in Biological Engineering. He will be performing a 6 month internship at the Massachusetts Institute of Technology (MIT) working on OpenWetWare. His focus of study will be the research, development and improvement of tools used in Open Notebook Science.

2 Introduction

The text of this research proposal was written in five days by an international collaborative group that did not exist prior to commencing the proposal. The group was self selecting, and aggregated around a Blog post [12], based on their interest in sharing, capturing, and annotating research data. The speed with which a community can be assembled (the network) and the ability to rapidly assimilate and present information (the proposal) demonstrates the potential of a different approach to research; one in which loose coalitions of scientists can rapidly develop around a specific problem, bringing different resources to bear. The key to this approach to working is that it is carried out 'in the open'. This means making raw and processed research data freely available for community comment well in advance of publication in the traditional peer-reviewed literature; in many cases making it available 'as it happens'. The technology that makes this possible has only recently emerged and is in its infancy. Different groups in different research disciplines are taking different approaches as well as developing diverse tools for data capture, data analysis and annotation, and data storage and publication. In addition much of this work is driven by junior scientists who in many cases do not, as yet, have permanent positions. This proposal is to provide the funding to support this community to meet and to provide a specific forum to identify and implement a route that will allow these systems to develop in parallel and to be integrated where appropriate.

The implications of this are not limited to the Open Research community. These systems will ultimately underpin the technology that will allow the international research community, and its funders, to move beyond a situation where the majority of funded research data is unavailable to one where data is properly curated, archived, and available for re-use. It will enable papers in the peer reviewed literature to be linked to the raw data that supports them. A curse of modern practice is the lack of availability of so called 'dark data' which is methodologically sound but never published [13]. Even where raw data is available it is a challenge to provide high quality and consistent metadata. Developing tools to enable the sharing of raw data and promoting the practise of using these tools will rectify this problem and enable the re-use of data in ways, in different disciplines, that the original researchers never considered (see [14] for historical and contemporary examples). The open research community are the simply the earliest and most public adopters and developers of these tools. Taking the opportunity to develop a consensus on best practise, definitions, and standards at an early stage will enable robust systems to be developed that will help to safeguard the future investments of research funders in the production of scientific data.

3 Open Research: The Publication@Source Paradigm

The world wide web has enabled a step change in the way science is communicated internationally. Few people today read an issue of a journal from cover to cover and a relatively small proportion of scientists even look at tables of contents. Large archives of data are available online, particularly in the sphere of biological sciences, providing both repositories for the increasingly high data content of modern high throughput science as well as providing an enormous resource for scientists world wide. However despite the promise of electronic media the basic format of traditional science communication remains fundamentally trapped in 19th century concept of the printed page. A large proportion, perhaps the majority, of published papers, do not contain sufficient information to allow the precise replication of results (see [15] for an example). A large quantity of data are trapped in inadequate formats within the published literature which in practise is limited to the static piece of paper. Even within the data that are available there are significant misassignment and metadata errors. Vast quantities of research data languish on laboratory computers and in laboratory notebooks without seeing the light of day and are lost when equipment is retired, computers are replaced, or research workers move on. In total a large proportion, almost certainly the majority, of government and charity funded research is never made available to the scientific community, representing a monstrous waste of resources. By contrast the availability of specific data sets in the biological sciences (the PDB, Genbank, microarray data) has enabled a huge range of new science while also providing strong quality controls for the publication process.

An international community of scientists, publishers, software engineers, and library and archive managers are taking a variety of approaches to tackle these problems. The consistent theme that links these efforts is the idea that data should be made as freely available as is practicable. This ranges from the Open Access publishing championed for example by the Public Library of Science [http://www.plos.org/], through efforts to scrape data from the published literature and provide this in useful and accessible forms [e.g. CrystalEye, http://wwmm.ch.cam.ac.uk/crystaleye/index.html], through to efforts to place the laboratory itself online and ultimately to make all the raw data from the laboratory available as it is generated. In Open Notebook Science, the logical extreme of this open approach, the researcher's note books are made fully and publicly available as soon as is practicable [3]. The promise of these approaches is a more responsive and connected science community. Coalitions could rapidly form to solve specific problems based on the specific resources and expertise available, the communication of important results could be much more rapid approaching instantaneous, and resources that are currently wasted on replication of other groups unpublished data could be put to better use.

There are two major challenges in delivering this vision.

1. The supporting technology: Interoperability and communication between tools is a serious issue due to the disparate nature of the community and limited common interchange and communication formats for research data. Many tools are being developed by groups ranging from publishers to academics, and archivists to hobbyists. However communication between these tools is haphazard and agreement on standards is limited.

2. The size of community: The large potential benefits of sharing information and collaborative approaches require critical mass before they can be realised. This goes hand in hand with the need to educate the research community on the value of properly archiving and sharing data. As the need for proper data storage and annotation are more widely recognised and valued, researchers will look for tools that make this possible. Only through providing critical mass will we realise the added value that data sharing can provide.

4 The Open Practises E-science Network

We propose to form an international network of researchers, publishers, and software developers with an interest in developing tools and practises to enable open research approaches. This network will include active researchers across a wide range of disciplines, developers of tools and standards, publishers, and funding agencies. While much of the initial development work has been in the general area of molecular science the network will directly seek the involvement of other disciplines including social and biomedical sciences. It will focus on the identification of common standards and tools, the identification and promotion of best practise, and above all enabling the communication between archives, researchers, and publishers that will realise the potential of the open research approach.

4.1 Why is a network required?

Compared to the large scale end of the e-science research effort the open research community is currently disparate and distributed but already showing great promise. Different players in this community meet at a range of disparate conferences but there are very few opportunities for meetings where the focus is on driving the open research agenda itself forward. This means that the underpinning technology, standards, and philosophy is not developing at the required pace, and when parts are developed, integration of these processes is less efficient than it should or could be. The emergence of an international community is a quite recent event with many key developments over the past 18 months. Many of the members of this community have never met face to face and existing collaborations are generally recent. In summary the tools are simply not as yet developed enough to allow the development of the 'open process' to happen automatically without face-to-face meetings.

In addition many of the most energetic people driving the agenda are graduate students and PDRAs who do not have their own budget to call on to attend meetings. Where funding for meeting attendance is available these young scientists are generally forced by the needs of advancing their career prospects to attend meetings directly related to their area of research. The development, implementation, and critical analysis of tools and practises, particularly online, is not as yet recognised as a valid contribution to career development in its own right for these early career researchers. The funding of a network would support these people, who will be leaders amongst the next generation of researchers in implementing these practises, to be directly involved in determining the direction the community takes. It will also provide an external validation of the value of their work in this area, providing both an added boost to their CVs and promoting the value of this development work more generally.

Finally, the archival and sharing of data and associated metadata is near the top of the agenda for most funding agencies, yet there is little resource being directed at the central problem of making it possible to capture and share general research laboratory data in an effective manner. The proper archival of general laboratory data (as opposed to the deposition of specific types of data) is generally not valued or well resourced and this poses a serious challenge to research funders in encouraging best practise. The network will define and promote best practise, develop tools enabling data archiving and sharing, and implement these tools and practises. This will raise the profile of the proper archival and sharing of data in general within the international research community as well as driving the development and improvement of the tools that will enable best practise. In short, the network will provide an opportunity for research funders to close the gap between the aspirations of current data archiving and sharing policies and the reality of researcher practise on the ground.

4.2 Aims of the network

The aims of the network are to provide a forum for identifying, discussing, and implementing solutions to the challenges noted above. This will be achieved through agreeing on standards and definitions, aiding in the development and integration of specific tools, identifying and promoting examples of good practise, and promoting the benefits of the open research approach to the wider research community. The aim overall will be to widen the applicability of tools and practises as far as is possible with the resources available. The specific aims of the network will be:

- To articulate, define, and promote the principles of properly archiving and sharing laboratory information;
- To agree a framework for identifying, and enabling the aggregation of, primary research data made available online;
- To identify practical routes towards making both raw and derived data self-describing and semantically rich;
- To investigate different approaches to data sharing in distributed collaborative projects;
- To publicise and promote to the research and data management communities the ethos of making research data freely available;
- To identify and implement best practise in licensing and publication of data;
- To identify and enable good practise in reconciling the twin demands of obtaining a financial return on research results and making those results freely available.

4.3 Implementation

The key to the success of the network is enabling and encouraging face to face communication between the founder members and seeking new members to expand the scope and reach of the network. The network will be run by a steering committee to be drawn from the founder members which will meet twice a year (once at the general meeting and once via video conference/online). The network will organise one annual meeting in each of the three years of the grant. The first meeting will be based in the UK and will be held in the second half of 2008. The network will directly fund some meeting attendees with priority given to early career scientists and those without independent travel budgets. The meetings will be run as a mixed programme of organised sessions on specific topics that relate to the aims of the network and an open 'Unconference' approach [16] to allow open discussion and presentation of issues of general interest to the community. The location of the meeting will be chosen to enable material (audio, video, photos, presentations) from the meeting to be freely and immediately available online. Support is requested for recording and streaming presentations, workshops, and meetings. The meetings will therefore provide a demonstration of the ability to effectively record, annotate, discuss and showcase scientific research projects. At the annual meeting two prizes will be awarded, one to the tenured academic or fully employed person or organisation that has made the most significant contribution to the promotion and/or implementation of open research in the previous twelve months. A prize will also be awarded to a non-tenured early career researcher who has made a significant contribution through research, promotion, or tool development in the previous twelve months. The prizes will be decided by an independent committee appointed by the steering committee. The recipients will be invited to present a plenary lecture at the meeting. The prize will not include a cash component derived from the grant (beyond the costs of attending the meeting).

Funds will be made available for extended exchange visits. Our experience is that this is the most effective way of understanding differences in local practise as well as raising new issues that are not obvious in isolation. Funding for exchange visits (ideally one visit in both directions) or small gatherings (3-5 people) to resolve specific issues will be distributed by the Steering Committee. Exchange visits will be for two weeks to one month and small gatherings for a day to a week.

The OPEN group will be represented by a web presence which will aggregate material from meetings, relevant blog posts, pointers to peer reviewed papers and data, and other related material. One problem identified by the open research community has been the limitations of currently available online publication tools including both Blogs and Wikis as forums for recording, presenting, and discussing primary research. The web presence will therefore provide a range of functionality building on the technical developments at OpenWetWare, Postgenomic, Chemical Blogspace, and other online projects.

5 Specific development aims

In addition to the general aims and actions detailed above, the network will work in a number of specific areas, to define and agree a statement of relevant standards, to actively identify and promote examples of linking primary research data online into databases and published papers, and to develop examples and case studies that help to resolve the potential conflicts between making data freely available and the obligations of research funders, research organisations, and researchers to exploit the results of research and safeguard the rights of research subjects.

5.1 Definitions, standards, and promotion

A key part of promoting any approach or set of tools is providing clarity on what precisely is proposed. For instance, there has been significant confusion and disagreement over the use of the term 'Open' with respect to Open Access journals (e.g. [17]). There have been several recent exchanges on whether specific projects constitute 'Open Notebook Science' and the requirements for 'Open Data'. In some cases these terms have been defined [3], but usage has drifted [18,19], and in others

there is genuine discussion as to how they should be defined so as to encourage good practise but also to provide realistic goals. At one level we are developing an 'etiquette' of good practise in this area and the discussion of standards of behaviour and how people react to specific events is just as important as agreeing on description standards, ontologies, or definitions. The network will provide a forum for discussing both definitions and standards as well as identifying and promoting good examples of adherence to these definitions. The network will also enable a wide ranging discussion of the practical consequences of these definitions. Several bodies with the aim of supporting or defining standards already exist including the Open Knowledge Foundation, Creative Commons, and Science Commons. The network will engage with these bodies to explore the practical consequences of their definitions and standards. The network will, where appropriate and where fully agreed by consensus, adopt specific standards and make specific declarations on standards and definitions.

A key activity of the network will be to identify specific examples of good practise and to publicise these through the consumer and science and technology media, the peer reviewed literature, and comment in the primary literature, as well as online through Blogs, the network web presence, and any other appropriate means, including presentations in virtual worlds. The network will work closely with ISIS and STFC Communications groups to deliver clear and targeted messages, aimed primarily at the research community. We will directly seek media coverage of meetings and specific speakers as a means of promoting the value of data archival, annotation, and storage. The consumer and science and technology media have shown an interest in this general area and this is an effective means of reaching a wide audience of researchers.

5.2 Primary data online

The central aim of the network is to make it possible to share, re-use, aggregate and publish raw data online and to promote the benefits of this approach to research. This will build on existing activities in Open Notebook Science and Open Data by identifying routes that enable data to be freely passed from one place to another. There are several aspects to this: providing a link between peer-reviewed literature and primary research data online; identifying primary research data online; aggregating data to specific portals; and enabling the automated identification and re-purposing of data. Providing technical solutions to all of these problems is beyond the scope of the resources available for the network so the main focus will lie in identifying routes towards solutions. Where feasible solutions will be tested or implemented in a variety of settings.

5.2.1 Linking peer-reviewed literature to primary data online

Despite the obvious potential advantages in requiring the deposition of raw supporting data for peer reviewed papers, it is very rare for such data to be provided. Partly this is a hangover from the days of paper journals where space was limited and partly due to the limited support provided by journals for providing supplementary data online. Many journals limit supplementary online information to a single pdf file making it effectively impossible to provide the raw data. While it may be possible in the future to package the raw data associated with a paper in an appropriate form it is therefore likely that journal website are not an appropriate repository. A simple solution is therefore to provide a pdf file that acts as an index, pointing out to the location of data stored elsewhere. However journal editors are often, reasonably, concerned about the stability and reliability of such data. There are therefore two central issues to be resolved: the stability of raw data online and the technical means of pointing to it (and back from it).

The first problem is a general one for research data and a number of potential solutions can be envisaged including the use of institutional, or other, repositories or third party web archive solutions such as WebCite [http://www.webcitation.org/]. These approaches are developing and significant advances can be expected over the lifetime of the network. It will therefore maintain a watching brief and seek to influence the development of these systems so as to enable long term laboratory data storage. The second problem is more immediately tractable, at least for specific cases. The Public Library of Science journal PLoS ONE [http://www.plosone.org/] has implemented a number of features with the aim of enabling a 'conversation' to take place around a published paper. These tools can be used both to 'point out', through comments and annotations on the paper, and to 'point in' via trackbacks from blog entries. It is therefore the natural place to investigate linking out through annotations of the paper to the raw data as well as linking in from blog based notebooks. The network will work with PLoS ONE to investigate this approach and its potential on specific papers.

5.2.2 Identifying and aggregating primary data online

A second issue for primary data is finding and recognising them. This is actually very simply achieved by exposing within the page html, specific tags, RDF, or microformat declarations that identify the page as containing primary research data. Specifically tagged RSS or ATOM feeds can also provide similar functionality and can easily be implemented using tools such as Yahoo! Pipes [http://pipes.yahoo.com]. This links into the issue of enabling data interchange through self describing formats. This is challenging but the first step of simply describing the page as containing research data is relatively simple through agreeing an appropriate tagging approach. The Open Notebook Science community is currently small and by taking the opportunity to agree standards now it may be possible to spread this practise effectively as the community grows. The

use of InChi and Smiles codes as identifiers of chemical compounds had enabled both the efficient searching of specific chemical compounds as well as the aggregation of material dealing with specific molecules at Chemical Blogspace. Similar approaches for other elements using specified mark up languages or other simple tagging systems (E.C. numbers for enzymes, Houben-Weyl system for chemical reactions, Genbank accession etc.) will be explored. Identifying the primary data makes it possible to aggregate it to specific portals. Several approaches are available for aggregation including the use of RSS/ATOM feeds from active research sites, spidering the web to identify specifically tagged material, or obtaining material through a customised web search. These approaches will all have advantages and disadvantages and investigating how they work in practise will be a valuable exercise. The results of practical experiments with these approaches will be feed back into discussion at the annual meeting as part of the process of agreeing standards where possible.

5.2.3 Self-describing data - enabling automated re-purposing

This is a serious and general technical problem and its solution is well beyond the scope of the current proposal. However it is possible to identify possible routes forward and to position the network to make grant applications. It is also possible to identify specific cases where information has or could be passed between systems and to encourage the development of these processes. Some examples of this can already be identified. The collaboration of Jean-Claude Bradley's group with ChemSpider has provided new means of depositing specific forms of chemical characterisation data from the Open Notebook laboratory into a publically accessible database. In many cases a large amount of data can be extracted by text-mining software and aggregated into repositories based on RDF and XML. These are then excellent resources for datamining and ontological explorations, especially involving mashups between different knowledgebases. This is extremely effectively for published crystallography, resulting for instance in the CrystalEye system with over 100,000 entries. Recently a similar project was launched - OpenNMR ("NMREye") which invited the community to comment on specific examples from a large dataset where calculated chemical shifts did not correlate with those reported [20]. Over the course of a few weeks a large database of NMR chemical shifts and associated structures was refined and 'wrong' structures removed. This provides a good test of how well the values of reported NMR chemical shifts can be calculated (answer: surprisingly well).

These examples show that much can be currently achieved. However the network will also look forward to scope out the route towards self-describing data. There is currently an active and ongoing discussion on the value of different formats for online laboratory notebooks and this will continue. However there is general agreement that there is a need for presentation tools that provide the functionality of both Blogs and Wikis, and more sophisticated tools for automating data capture. The experience at Southampton is that the much of the critical metadata that describes an experiment can be automatically captured. Linking this directly to RDF or XML documents has clear potential as a route forward and the network will investigate possible routes as well as position itself for grant applications to resource those developments. The use of XML, RDF, and the growing set of tools for the semantic web clearly have an important role to play here as does identifying routes towards the embedding of existing ontologies for the description of experimental procedures, samples, and data types. The interests of the W3C Semantic Web for the Life Sciences group will overlap here and the network will aim to initiate dialogues with the relevant development projects in this area. The OPEN group will have a valuable role to play here as a provider of use cases as well as providing a practical users perspective.

5.3 Licensing, legal, and ethical issues

A common assumption is that providing raw data online, or otherwise commenting on the progress of research, precludes the protection and commercialisation of the products of that research. Disclosure of a patentable invention does preclude patenting but the management of any patent process can be linked with the management of disclosure. There is a strong case to be made, particularly for the exploitation of academic research, that the traditional approach of maintaining complete confidentiality can *hamper* the process of exploitation by reducing the ability of the researcher to publicise their invention. A well informed approach to both the protection and release of research results therefore has the potential to maximise returns for both the potential commercial and scientific value of the research as well as returns for the researcher in terms of peer reviewed papers and career advancement. The network will therefore engage directly with Technology Transfer groups (e.g. CLIK at the Science and Technology Facility Council) in identifying the most appropriate approaches for specific cases. Specific example cases will be explored to provide detailed guidance on how best to pursue both aims.

Additional arise where the desire to make research results freely available has the potential to conflict with privacy, ethical, or safety requirements. In some cases, particularly in the social sciences, the act of making the results available may in fact influence the outcome of the study. The network will aim to identify and document example cases and engage with the social sciences community to identify approaches to resolve these conflicting requirements where appropriate as well as identifying the characteristics of those research studies where data should not be made freely available.

Finally the simple matter of which license to use for both online comment and raw data is often an area of some confusion. The Open Research community has largely selected the Creative Commons By Attribution license as the preferred standard although in specific cases various other licenses are preferred. A range of organisations are working in this area (Creative Commons, Open Knowledge Foundation, Science Commons) and the role of the OPEN group will be again to engage with these efforts by providing a practical view of the consequences of different choicices

6 Fit to assessment criteria for this call

This network proposal is to support the meeting of a currently loosely associated international grouping. This is not a continuation of a pre-existing network. The focus of the network is on enabling the collection, annotation, and publication of data directly from the laboratory general research laboratory. The network will therefore play a direct role in enabling, encouraging, and promoting the use of E-science to the widest possible research community in the UK and internationally. The network will explicitly aim to embed more tools developed by the E-science community into everyday research practise therefore strengthening the E-science community in its role as a resource for the general research community. The potential positive impact of changing research practise is enormous. The network will play an important role in promoting the potential of this way of working. This will lead to improvements in the way in which data are stored and published with significant positive effects for the research community. Additionally the network will identify those areas of concern and potential negative impact and develop specific guidance for addressing these issues.

The network is strongly linked into existing e-science programme activities across the UK and internationally but also engages in a new way with a range of current and future potential users. Connections are in place with e-science groups at Southampton, Rutherford Appleton Laboratory, and Cambridge. There are also connections with international efforts at OpenWetWare (MIT), ChemSpider, the University of Sydney (Synaptic Leap), and the Open Notebook Science group at Drexel (see letters of support). There is also interest from a wide range of researchers from PhD students through to tenured academics. This is a diverse mix of developers, data specialists and laboratory end users. The aims of the network involve linking all these groups together in a conversation about how research data is houses and presented.

The proposal is centred on dissemination routes for primary and annotated data. Where appropriate papers will be published in the peer reviewed literature but the majority of dissemination will be through online media including blogs, the development of specific tools and databases, and the aggregation and annotation of existing data. General media coverage will be explicitly sought. Making research data more available is an area which the mainstream media are interested in. Positive coverage of the networks efforts in the mainstream and general science media will be a significant component of promoting the open research agenda.

7 References

1. G Eysenbach (2006) Citation advantage of open access articles. PLoS Biol 4: e157.
2. HA Piwowar, Day RS, Fridsma DB (2007) Sharing detailed research data is associated with increased citation rate. PLoS ONE 2: e308.
3. J-C Bradley (2006) Open Notebook Science, [Internet], http://drexel-coas-elearning.blogspot.com/2006/09/open-notebook-science.html
4. J-C Bradley et al (2007) SciFoo Lives On, [Internet], http://scifooliveson.wikispaces.com/
5. B Hooker (2007) Open Reading Frames, [Internet], http://www.sennoma.net/
6. 3QD Editors (2004-2007) 3 Quarks Daily, [Internet], http://www.3quarksdaily.com/http://www.3quarksdaily.com/
7. B Hooker (2006) The Future of Science is Open, Part 1: Open Access, [Internet], http://3quarksdaily.blogs.com/3quarksdaily/2006/10/the_future_of_s_1.html
8. B Hooker (2006) The Future of Science is Open, Part 2: Open Science, [Internet], http://3quarksdaily.blogs.com/3quarksdaily/2006/11/the_future_of_s.html
9. B Hooker (2006) The Future of Science is Open, Part 3: An Open Science World, [Internet], http://3quarksdaily.blogs.com/3quarksdaily/2007/01/the_future_of_s.html
10. M Todd et al (2006 - 2007) Schistosomiasis Research Community, [Internet], http://www.thesynapticleap.org/schisto/community
11. P Beltrao (2004 - 2007) Public Ramblings, [Internet], http://pbeltrao.blogspot.com/
12. C Neylon (2007) e-science for open science - an EPSRC research network proposal, [Internet], http://blog.openwetware.org/scienceintheopen/2007/11/22/e-science-for-open-science-an-epsrc-research-network-proposal/
13. T Goetz (2007) Freeing the Dark Data of Failed Scientific Experiments, Wired, Issue 15.10
14. P Murray-Rust (2007) Data-driven science - a scientist's view, [Internet], http://www.sis.pitt.edu/~repwkshop/papers/murray.html
15. J-C Bradley (2007) Experimental Uncertainty Principle, [Internet], http://usefulchem.blogspot.com/2007/11/experimental-uncertainty-principle.html
16. Wikipedia Editors (2007) Unconference, [Internet], http://en.wikipedia.org/wiki/Unconference,
17. Open Medicine Editors (2007) OM blog - CMAJ Endorsement, and We Respond, [Internet], http://blog.openmedicine.ca/node/42
18. J-C Bradley (2007) Science is About Mistrust, [Internet], http://usefulchem.blogspot.com/2007/10/science-is-about-mistrust.html
19. P Murray-Rust (2007) Open Notebook : more ideas, [Internet], http://wwmm.ch.cam.ac.uk/blogs/murrayrust/?p=743
20. P Murray-Rust (2007) Archive for category 'NMR', [Internet], http://wwmm.ch.cam.ac.uk/blogs/murrayrust/?cat=22

Justification of resources

The primary support requested is to enable direct face to face meetings in two main forms, the full network annual meeting to be held in each year of the grant, and exchange visits between participating organisations. The overall cost is modest with a total cost to the research councils of £86k.

Staff and overheads costs

Costs are requested for the time of the principal investigator and administrative support. The PI will act as Chair of the steering committee and spend on average approximately two hours per week administering the grant and the network. This time will be concentrated around annual meeting and steering committee meetings and is based on the PI's experience of running similar workshops, meetings, and networks. Administrative and technical support at a similar level is requested to assist in the planning of logistics of meetings, particularly in organisation of travel arrangements as well as a contribution to maintaining the group web presence. Again this will be concentrated around the period of meetings.

Consumables

Support is requested for the running of meetings. As noted in the proposal a key aspect of this network is that many of its members will not have access to their own travel budgets. While some will be supported directly to attend meetings we also wish as far as possible to reduce cost barriers for those who cannot be fully supported. We therefore request funding to support room hire (£4,500) and facilities, including wireless networking, video, audio, and screencast recording of talks and workshops (£4,500) and technical support for the transfer of these to suitable formats and online (£1,500). The costs for these facilities are estimated based on those for the Neutrons in Biology meeting run at RAL in July 2007 running for three days with 50 attendees. We expect fewer attendees (20-30), at least for the initial meetings, but with a higher degree of technical support and facilities required.

A small quantity of resource (£5,000) is requested for small equipment purchases that will enable specific development projects to be taken forward. This will be used for small pieces of hardware, e.g. hard disks, tablet PCs, PDAs, barcode readers and printers that will make a significant difference in the development of a specific tool or use case and can not be funded through other sources. The funding will be available at the discretion of the Steering Committee and applications will be called for at each annual meeting. A further £3000 is requested for the provision of server hardware to support the group web presence.

Travel

Funding is requested to support the travel costs of members of the steering committee to attend the first steering committee meeting (£2000, subsequent SC meetings will be online or via video conference). In addition funds are requested to support network members (£21k for network members, 7-10 supported each year, and £9k for prize recipients, two per year) to attend the annual meeting. Preference will be given to Prize recipients (selected by an independent prize committee) and those attendees with no or limited access to independent travel budgets. The network will aim to develop to the point over three years that meetings can be funded directly through registration fees and sponsorship.

Specific funding is requested for exchange visits and small meetings both to tackle specific technical issues, transfer best practise, and to understand differences between practises and tools in use in different places. This will enable both short concentrated visits as well as extended exchanges. Exchanges involving reciprocal visits between two sites will be encouraged. This is budgeted as £5000 per visit (for a two way exchange) on average including travel costs and subsistence as these are likely to involve overseas travel. Preference will be given to those network members without an independent travel budget and/or those that can provide an additional contribution to the costs of the visit. We aim to enable six such visits over the course of the grant (two per year). Funding recipients will be expected to report to the annual meeting on the result of the exchange.

Month	Steering committee meeting	Annual meeting	Meeting themes and issues
0			1. Selection of prize committee and encouraging nomination of candidates for prizes 2. Identification of invited speakers for first annual meeting 3. Identification of existing and desired interactions with other groups
6			1. Identification of specific definitions and standards for development 2. Identification of specific challenge proposals for network to take forward 3. Formation of working groups to take specific issues forward 4. Defining expectations, resources, and future possible funding opportunities 5. Selection of first round of exchange visits
12			1. Tracking progress of working groups and exchange visits 2. Identifying possible declarations/standards for discussion at second annual meeting 3. Identifying future potential sources of funding to maintain network activities
18			1. Standards proposals and discussion, results of practical investigation 2. First reports of exchange visits and working groups 3. Working groups established for follow on funding applications 4. Working groups established for funding applications for collaborative development projects
24			1. Ongoing monitoring and identification of new and developing opportunities and issues 2. Finalisation and submission of follow on funding applications 3. Finalisation and submission of development funding applications 4. Identifying possible declarations/standards for discussion at second annual meeting
30			1. Final annual meeting funded from current proposal 2. Report on progress over the course of the grant 3. Reports on progress on funding application 4. Reports on exchanges and working groups 5. Standards proposals and discussion, results of practical investigation
36			1. Finalise accounts 2. Prepare final grant report 3. Report to community on continuation of activities

The OPEN group will be active in a very rapidly moving area with many players. It is therefore not straightforward to predict the details of the issues that will be identified. In addition the collaborative and open-ended nature of the planned meetings deliberately allows for significant flexibility. Nonetheless specific themes are identified above that may predominate at specific meeting. Tracking of progress will be required throughout the period of the grant and identifying opportunities for further funding will need to be done at the appropriate time. The schedule of meetings is given and two exchange visits are planned per year.