


JISC Grant Funding 14/09

Cover Sheet for Proposals (All sections must be completed)			
Name of JISC Initiative:		Managing Research Data Programme: Strand A	
Name of Lead Institution:		UKOLN, University of Bath	
Name of Proposed Project:		SageCite: citing large-scale predictive network models of disease	
Name(s) of Project Partners(s)	Department of Computer Science, University of Manchester, British Library	Any private sector involvement in the Project YES NaturePG, PLoS, Sage	
Full Contact Details for Primary Contact:			
Name:	Dr Liz Lyon	Address:	UKOLN
Position:	Director	University of Bath	
Email:	e.lyon@ukoln.ac.uk	Claverton Down	
Tel:	+44 (0)1225 386580	Bath, BA2 7AY	
Fax:	+44 (0)1225 386838		
Length of Project:	12 months		
Project Start Date:	1 August 2010	Project End Date:	31 July 2011
Total Funding Requested from JISC: £		£120,000	
Funding requested from JISC			
August 10 – July 11		£120,000	
Total Institutional Contributions:		£40,000	
<p>Outline Project Description : SageCite will develop and test a Citation Framework (Data, Method, Publication) for complex network models of disease and associated data as Research Objects, with requirements informing a demonstrator using Taverna workflows, DataCite services as an extension to myExperiment and Sage Commons data infrastructure, as the case study. Citations of network models will be embedded in two leading publications: Nature Genetics and PLoS Computational Biology. A Benefits Evaluating mapping will be produced using the KRDS2 Benefits Taxonomy. SageCite will join up JISC RDM with the international bio-informatics initiatives: Concept Web Alliance and Bio2RDF.</p>			
I have looked at the example FOI form at Appendix A and included an FOI form	<u>YES</u>	NO	
I have read the Funding Call and associated Terms and Conditions of Grant at Appendix B (Tick Box	<u>YES</u>	NO	

SageCite: citing large-scale predictive network models of disease.

1 Proposal Description

1.1 Appropriateness and Fit to Programme Objectives

The background and rationale for SageCite is set in the context of increasing calls in the research community [1,2] to demonstrate the ability to cite data-sets and to develop new mechanisms for attribution. The exact nature of the “data” is complex and the notion of “*attribution granularity*” was explored in a recent CNI Keynote [3] where digital entities to be cited may be at the macro level (journal) and progress through article, workflow, visualisation, model, data and annotation, to a concept expressed as a number of RDF statements: micro/nano-publication [4]. This term is adopted by the [Concept Web Alliance](#), with similar nano-level approaches to attribution in human genomics work e.g. micro-attribution of annotations in the human variome project [5].

The *Open Science at Web-Scale Report* [6] described rapid growth in data volumes from gene sequencing instruments with data production on a greater scale from next generation sequencing technologies. Genome scale network biology is exemplified by a recent breakthrough publication describing family-based genome analysis where gene and clinical data-driven models were associated and informed predictions of future disease [7]. Genome Wide Association Studies underpin these emerging new approaches, where complex network models driven by distributed clinical and genetic data, are developed and integrated to examine causal relationships and associations.

[Sage Bionetworks](#) is a not-for-profit organisation which seeks to facilitate the open sharing of genomic and clinical data and associated network models via the Sage Commons infrastructure, to enable the accelerated development of large-scale predictive models of disease [8]. Sage is a collaborative effort embracing the international bio-informatics research community. An outline of requirements for citation of network models has been reported by this bid team at the recent Sage Congress and on the Sage Wiki [9].

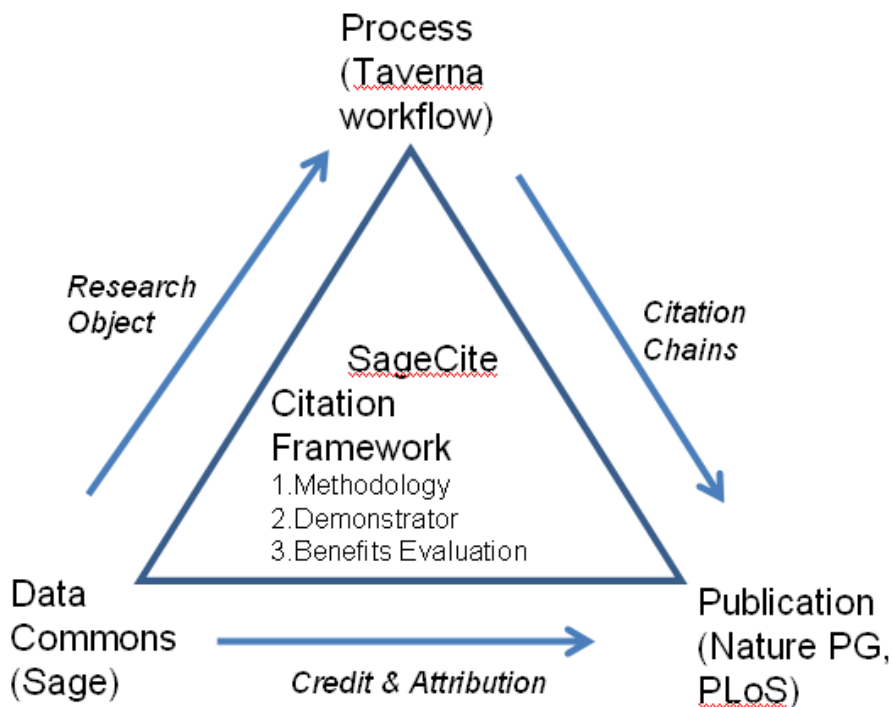
Unique to this proposal, SageCite joins up the JISC Managing Research Data Programme with established international bio-informatics initiatives (Concept Web Alliance, [Bio2RDF](#), [Chem2Bio2RDF](#)), which are progressing semantically enriched linked-data solutions for bio-medical open data. Sage has leveraged crowd-sourced community effort via the Sage Commons and has potential to radically transform scholarly communications in clinical medicine and disease biology.

Network models (bionetworks), are the outputs of an analysis (a code or workflow for example), of prior results which may be other networks or base, primary data generated through observations, instrumentation or predictions. Bionetworks are fundamentally compound Research Objects [10] that link the methods and materials used to produce them. The Sage Bionetworks case study described in SageCite, represents two important aspects of data citation reflecting this compound nature, which we dub “*black box citation*” (citing of the network as a indivisible entity partnered with its method of production but obscuring the source of its data), and “*white box citation*” (unpacking of the bionetwork to reveal its compound nature and the citations of its base components). White box citation leads to the notions of “*citation chains*”: track back the bionetwork and its source components and its mirror “*citation propagation*” of source components into the provenance records of a bionetwork. This is critical. There are many players who deserve citation: the creator(s) of the data; the creator(s) of the analysis methods; the creator(s) of the networks; the author(s) of peer-reviewed publications. We can think of this as a “*fractal citation model*”.

The white/black box citation notion poses fundamental issues that SageCite will investigate with pre-existing software and concrete use cases: What is the citation / curation boundary? How do we bottom out a citation transitive closure? When should citation DOI (digital object identifiers) be allocated to results? How do we build, instrument and report citation chains? How do we combine community standardisation initiatives on **provenance** (e.g. the [Open Provenance Model](#), [W3C Provenance Incubator](#)), **digital aggregation** models of compound objects (e.g. [Memento](#), [ORE](#)) and **citation** (e.g. [DataCite](#), [ORCID](#))? We discuss each of these community efforts in more detail below.

SageCite offers an exceptional opportunity to explore these issues through our **Citation Framework** (Figure 1) which combines three essential infrastructural research components: data, process, publication.

- **Data** represents source material to be cited. The Sage Commons repository of bionetworks and links to the base data the bionetworks are based upon. One such data source is Bio2RDF: an integrated semantic web atlas of post-genomic data gathered from over 35 high value public datasets. *SageCite has prime access to Bio2RDF and Sage Commons.*
- **Process** represents the methods that are applied to data materials to generate data, combine data and produce new, insights and new citable scientific research objects. Scientific data analysis methods include workflows, such as Taverna. Taverna workflows using Sage bionetworks have been demonstrated at the public Sage Congress. Such data analyses produce provenance (history and dependency graph) that links citable results to the citable processes and citable source data they arise from. Taverna is one of the first systems to be “Open Provenance Model” compliant. This model is designed to allow provenance information to be exchanged between systems by means of a compatibility layer, to help developers to build and share tools that operate on such a model and to support a digital representation of provenance for any "thing", whether produced by computer systems or not. myExperiment, the open repository for workflows, supports workflow citation and attribution and Research Objects using the OAI-Object Reuse and Exchange standard as an aggregation model. However, these models have weaknesses, notably with versioning. Memento is a recent initiative to address version management of compound objects on a web-scale. *We will build on this prior and ongoing work using myExperiment, Taverna and Sage Commons to create an extensible testbed for SageCite. Goble’s group develop Taverna and myExperiment and participate in the OPM specification.*
- **Publication** represents the outcomes of research that will carry black and white box citations and need mechanisms for managing citation chains. This means combining provenance models with citation models, mechanisms and policies. DataCite is an international collaboration housed at TIB, Germany with the British Library (BL) acting as the UK member. The BL is the regional agency for the International DOI Foundation. The Open Researcher and Contributor ID (ORCID) is a community effort to establish an open, independent registry that is adopted and embraced as the industry’s de facto standard for name attribution. *SageCite is privileged to have BL, Public Library of Science (PLoS) and Nature Publishing Group (NPG) as partners. The BL represents DataCite; all are participating members of ORCID; all are involved in the Sage effort .*



By investigating the support for black and white citations from data to process, and from process to publication, we intend to join up the link from data to publication: *Credit & Attribution.*

SageCite addresses the 14/09 objectives (Circular para):

- Work Package 1 examines approaches, options and requirements for citing large-scale predictive network models of disease and compound research objects (32b-d).
- Work Package 2 demonstrates a **citation service** for network models and associated data in the Sage Commons through a **linked data** approach(35b-f).
- Work Package 3 explores **integration** of cited network models in peer-reviewed publications (32d).

- Work Package 4 reports on an evaluation summary, stakeholder analysis and a **benefits mapping** using the KRDS2 taxonomy (34) and international dissemination activities involving the bio-informatics domain and research and information communities more widely (45).

1.2 Value to the JISC Community

SageCite will significantly contribute to ground-up debate amongst research scientists and into publisher policy developments and will augment wider discussions on achieving career credit and attribution for data publication. The early calls for action in this context (e.g. Nature Special Issue on data-sharing), have emerged from the bio-science community. Our partnerships put us at the centre of this debate.

The establishment of sustainable mechanisms for data citation and attribution will contribute to the longer-term sustainability of the scientific record in these disciplines, providing persistence over time and additional provenance information relating to the network models and associated data.

There are benefits through enhanced discovery and access with identification and citation mechanisms for network models, leading to enhanced Return-On-Investment from the very significant amounts of public money invested in genomic research by bio-medical funding bodies, charities and trusts. There is potential societal value through making large-scale predictive network models openly accessible and computationally available: Lee Hood (Institute for Systems Biology) describes a vision of “*P4 medicine: predictive, personalised, preventive and participatory*” [11] and the ability to identify, cite and re-use network models, will greatly advance progress towards this goal.

All of the results and lessons learnt from SageCite will be shared with the wider community both at the institutional information level but also at the inter-disciplinary level. One of the strengths of the multi-skilled project team is the ability to reach out to both of these audiences and the track record of high profile pioneering projects and Reports, clearly demonstrates our effectiveness at developing data management methodologies, promoting good practice and influencing policy and strategy at UK and international level.

1.3 Quality of Proposal and Workplan

SageCite has the following Objectives:

- Collect, review and assess requirements and technical solutions for a scalable citation and attribution framework for network models, which is compliant and/or complementary to related citation approaches in this domain, and which may be applied across disciplinary boundaries.
- Develop and test a Sage Commons Citation Pilot which is embedded in bio-informatics workflows that consume and produce bionetworks, alongside other local and online data sets and publications.
- Demonstrate citation feasibility through pilot integration in the Sage Commons.
- Demonstrate attribution feasibility through integration of the pilot service in established journals.
- Evaluate benefits from SageCite and make recommendations for community practice, data publication strategies, long-term sustainability and data re-use and return-on-investment value.

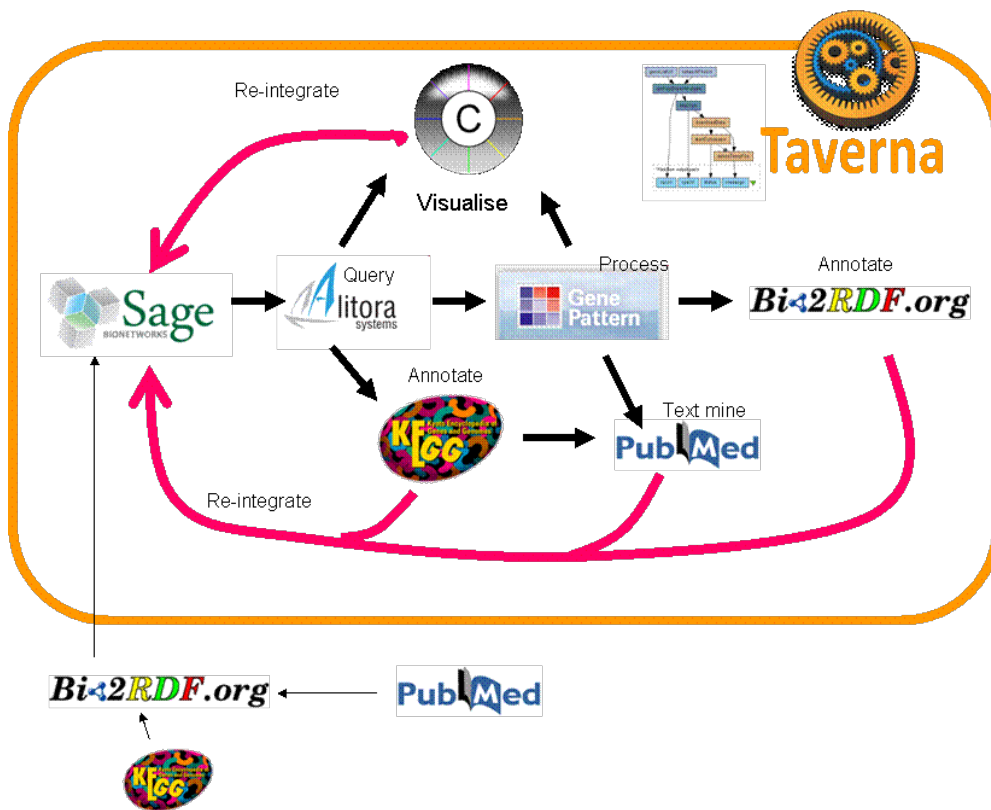
1.4 Work Package 1 Framework Foundations (UKOLN/DCC lead)

1.4.1 Task 1.1 Network models component analysis and curation lifecycle

Building on preliminary work presented at the Sage Congress (April 2010), we will carry out desk research to collect evidence of types of network models, data types, data formats, data repositories, metadata schema and standards, ontologies, identifiers, annotations, visualisation and analysis software. We will liaise with Bio2RDF, Concept Web Alliance and Sage groups. **We will also investigate other disciplines where large-scale network models have been developed to predict physical outcomes: engineering, paleo-climate, and make assertions on the wider applicability of the SageCite approach.**

1.4.2 Task 1.2 Requirements Capture

We must grasp the research practice and requirements of the biomedical community working with genomic and clinical data, in order to develop services that deliver the required functionality. The predictive computational network models of disease are derived from advanced integrative genomic analysis of genetic and clinical datasets (gene expression, clinical trait and genotype data). Network models include co-expression networks and Bayesian networks rendered as complex visualisations using the open-source Cytoscape software platform. In order to derive and analyse network models, various datasets are



processed through complex analysis pipelines. A simplified diagram (Figure 2) of a Sage Pipeline defined using Taverna is given below, using software tools such as the Alitora data service, a GenePattern Key Driver Analysis service, Gene annotation using KEGG and Bio2RDF and Text mining using PubMed services. Sage Commons is fed by Bio2RDF sources. We need to understand: at which intervention points there is a requirement to cite a data-element, a data-file, node, edge, model, visualisation or “package” of associated entities; how the workflow collects provenance and how the

workflow could generate a citation for resultant bionetworks.

Sage is forming a Federation of participating laboratories to contribute data and network models to the Commons (Ideker Lab, UCSD, Califano, Columbia Univ, Schadt, Pacific Biosciences/Sage, Friend, Sage), which will act as a data repository testbed. Key UK-based research groups at Cancer Research (Letter of Support), are also collaborating with the Sage effort. We will explore a range of methodologies to investigate citation requirements from Federation members, including iterative prototyping and mock-ups, semi-structured interviews and surveys. Our preliminary work included development of a draft set of questions for interviewing scientists and network modellers. The results of the foundational work will be a **Network Models Citation Requirements Report** (Deliverable 1) to inform the pilot/demonstrator.

1.5 Work Package 2 Demonstrator (Manchester lead)

1.5.1 Task 2.1 Evaluation of technical solutions

A number of technical solutions have been proposed to support the citation of datasets, the role of identifiers for data and researchers, provenance of data, aggregation of compound data research objects and the citation of methods. The demonstrator will be founded on the use of [Linked Open Data](#) (LOD) and several pre-existing resources: Bio2RDF, Sage Commons, Taverna, myExperiment and services that operate on Sage data. Linked Data is a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using [URIs](#) and [RDF](#). myExperiment and Bio2RDF are already LOD compliant; in SageCite we will make Taverna workflow provenance and Sage Bionetworks LOD compliant. For provenance and aggregation we propose to extend the OPM and ORE representations as necessary, leveraging emerging models such as [Void](#) for dataset attribution and Memento for versioning. We intend to build on the model of Research Objects as proposed in [10].

For data citation the technology choices are less clear. The DataCite service offers the assignment, registration and resolution of an identifier (DOI) to a dataset. This approach could also be adopted for methods, such as workflows. The drawback of DOIs is their cost and, like the Life Science Identifier, they do not use the standard URL resolution protocols. Citation Vectors, proposed by Axton (Nature Publishing Group), used an OpenURL, in which the parameters are used to give credit to the contributors and to locate the resource on the Web, using a citation repository as the resolver service. A TrackBack mechanism, an extension to the TrackBack protocol of notification used by the blogging community, has been demonstrated by the [CLADDIER](#) and [STORELINK](#) projects. The Concept Web Alliance proposes a model of “nano-publications” that are a fine-grained description of research statements and assertions. A Concept Wiki acts as an identity resolution service and a simple nano-publication citation mechanism. Science

Commons proposes the use of PURLs (persistent URLs) for permanent identification of database records for the Life Sciences (sharedname.org). In Europe a web-scale open service called **Entity Name System** for supporting the systematic reuse of identifiers for "things" has been developed in the [OKKAM](#) project. These approaches will be evaluated during the course of the study against the requirements of the network modellers. Selected implementations will be developed as prototypes or demonstrators in partnership with DataCite and two publishers / journals, *Nature Genetics* from NPG and *PLoS Computational Biology*; all three organisations are partners in this bid (Letters of Support).

1.5.2 Task 2.2 Citation-enabled workflow demonstrator

We will investigate development options for datasets and models made available through Sage Commons by developing the extensions needed to create a citation-enabled workflow system. Extensions include:

- Implementation of experimental identification and citation schemes for Sage Commons using TrackBack and DataCite and compare the approaches;
- Creation of a citation service to be incorporated into Taverna to enable auto-allocate blackbox citation identifiers and auto-citation resolution;
- Extension of the Taverna provenance collection mechanisms to gather citation data for whitebox citation chains;
- Extension of myExperiment to include DataCite and/or TrackBack citation mechanism for Research Objects, and to include workflow provenance in research objects;
- Development of experimental citation browsing tool to explore citation drill through for bionetworks generated by a Taverna workflow.

Using DataCite services and Sage bionetworks as the case study, we will investigate the following citation service issues: What is the scope for national vs international services related to citation? Are there national level services that the British Library could support to facilitate data citation? What are the issues around attribution and granularity in this bioscience community, what are the links with similar discussions happening within other disciplines? How can linked data be deployed to expose data and metadata in ways that enhance citation / citability? What are the implications of 'doing' data citation (and the mechanisms by which it is achieved), for the long-term stewardship of the scientific record? What is the role of open access publishers in enhancing the practices of citation of data? Deliverable D2 **Demonstrator & Briefing Notes**.

1.6 Work Package 3 Integration with STM Journal Publications (Nature PG/PLoS lead)

We will work with two journal publications to demonstrate accreditation and attribution.

1.6.1 Task 3.1 Nature Genetics

Nature Genetics currently mandates data deposition and citation of NCBI Genbank IDs for new nucleic acid sequences and of MIAME-compliant IDs for RNA expression microarrays. The journal also mandates a range of standard nomenclatures including human gene nomenclature and it is journal policy to encourage data deposition and citation of unique accession IDs within the main article. *Nature Genetics* will mandate - as a necessary condition of peer review - citation of any SAGE Commons accession codes adopted for external reference to data, individuals, consortia and models within the Commons. In addition, in accordance with international agreements on prepublication data sharing, e.g. Fort Lauderdale Convention (2003) for data producers and data users, and the Toronto International Data Release Workshop Authors (2009), *Nature Genetics* will increase its efforts to ensure data release in accordance with a citable funder-mandated project summary (data management plan). The journal will require users of publicly funded resource projects to cite as a condition of peer review: i) data accession codes, ii) DOI of published or preprint project summary and iii) email communication between data producer and data user. *Nature Genetics* will also work with selected funders, to obtain sets of project summaries for the complete set of grants funded under a resource generation program. For each program we have created a "Collection" in the preprint archive *Nature Precedings*; the summary should contain the aims of the data producers, the way in which data should be cited and any use restrictions on competitive publications. *Nature Precedings* preprints may be updated (versioned) and are not considered competitive publications by *Nature* journals. Furthermore, *Nature Genetics* will periodically count and display quantitative citations to data accessions and other unique citable items (micro-citations), for the purpose of tracking and promoting the utility and resource allocation of those data types. When author and contributor IDs become available, the journal will integrate these into its data citation metrics.

Task 3.2 PLoS Computational Biology

PLoS Computational Biology has a data publishing policy based on the National Academies Press principles: "Publication is contingent on making data integral to a manuscript freely available without restriction, provided that appropriate attribution is given and that suitable mechanisms exist for sharing the data used in a manuscript. Data for which public repositories have been established that are in general use should be deposited before publication, and the appropriate accession numbers or digital object identifiers published with the paper".

We will work with PLoS to test the SageCite Citation Framework and will explore implications on the proposed "*Scholar Factor*"[12], presented as a new metric which includes published data. We will use the citation of Sage network models and data, as demonstrator implementations of the policy guidelines. Deliverables will be **Demonstrators and Briefing Notes** describing technical and policy issues (D3).

1.7 Work Package 4 Benefits Evaluation and Dissemination (UKOLN/DCC lead)

1.7.1 Task 4.1 Benefits Evaluation

A self-evaluation of SageCite will be carried out. A core component is a mapping of SageCite outcomes using the *Benefits Taxonomy* from the *Keeping Research Data Safe2 Report*. We will focus on qualitative benefits in the Sage domain of clinical science and disease biology, and will assess the direct and indirect benefits, near and long-term benefits and private and public benefits from the citation demonstrator implementations for bionetwork models. The work will cover an analysis of SageCite stakeholders. Key elements will include consideration of the continued scalable development of the infrastructure, adoption and embedding of its associated processes, tools and workflows, by the bio-informatics and clinical research community and by STM publishers, implications for scholarly metrics and research assessment exercises, health policy issues and societal benefits for the treatment of disease. We will make Recommendations to JISC, the research community, HE institutions, funders and policy makers, learned societies and commercial publishers, to inform community practice, data publication strategies, long-term sustainability and return-on-investment value in the sector. Deliverable D4.1 **Benefits Evaluation Report**.

1.7.2 Task 4.2 Outreach and Community Engagement

Whilst the network biology, systems biology and bio-informatics communities will engage with the project through the requirements analysis tasks in WP 1, and through the exposure and testing of the pilots, two specific communities have been identified as Outreach targets to communicate the findings and further embed the outcomes from this study. Firstly the research community (scientists from disease biology backgrounds) will be addressed through presentations on SageCite at a range of high-profile disciplinary seminars, workshops and conferences during 2011 including Bio-IT World and the 2nd Sage Congress.

Secondly, the data management and Library and Information community will be addressed through the DCC Research Data Management Forum (RDMF): a focus for practitioners involved in this field. We will offer to showcase this project at a future RDMF meeting.

Furthermore, the co-PI's (Lyon & Goble) regularly author articles and present lectures and keynotes at international conferences and workshops in the areas of Open Science, Data Management, Digital Curation and eResearch. We will target presentations at the following meetings: IDCC Conference, UK eScience AHM, IEEE eScience, CNI Taskforce, and JISC Programme events. A project wikispace will be created and update postings made to the Research Data Managers blog. Deliverable: D4.2 **Outreach Programme**.

1.8 Work Package 5 Project Management (UKOLN/DCC lead)

1.8.1 Task 5 Project Management

UKOLN/DCC will provide project management capability and day-to-day operational oversight of the work. Project start-up will be informed by an initial face to face meeting of project personnel at month 1 to establish the Project Plan, with further f2f meetings at 6 & 10 months. This will facilitate oversight of the requirements; a mid point health check; review of the demonstrator implementation and preparation for report writing. Monthly project telephone conferences together with bilateral partner meetings will provide practical management between f2f meetings. The project team will work pro-actively with JISC (Simon Hodson) and reports will be provided as required. Deliverables: **D5.1 Project Plan; D5.2 Final Report**.

1.9 IPR and Accessibility Issues

The project will comply with the JISC Funding Agreement. It is expected that whilst most (if not all), outputs will be openly available (with Creative Commons licenses where appropriate), on the SageCite Web site, any intellectual property from resulting research outputs, will be subject to the Copyright, Designs and Patents Act 1988. The accessibility of Web-based systems and software will be addressed. We are aware of IPR issues which may arise from cross-sectoral collaborations and will seek expert advice from institutional and DCC legal experts.

2 Project Deliverables Summary and Timetable

WP	Description	Deliverables	Month	Lead + partners
1	Framework Foundations	Requirements Report	1-4	UKOLN DCC
2	Citation-enabled workflow demonstrator	Demonstration of workflow-based citation gathering, generation and tracking	3-11	Manchester CS
3	Integration with STM journal publications	Implementations at Nature Genetics, PLoS Computational Biology	6-11	Manchester CS + Publishers
4	Benefits Evaluation and Dissemination	Evaluation & Benefits Report, Outreach Programme	1-12	UKOLN DCC + All
5	Project Management	Project Plan, Final Report	1-12	UKOLN DCC+ All

3 Risk Assessment

Risk	Probability	Severity	Score	Action/Mitigation
Difficulties recruiting or retaining staff	2	4	8	Key members of staff already in post at UKOLN DCC and Manchester
Failure to meet project deadlines	2	4	8	Clear project plan with relevant tasks outlined, continuous review and rescheduling of work.
Failure to disseminate	2	2	4	UKOLN DCC and Manchester have very effective and proven dissemination channels.
Project is over-ambitious	2	2	4	The project plan will ensure the project does not divert from agreed goals.
Project team is working in isolation	2	2	4	UKOLN DCC and Manchester have strong existing links with international initiatives.
Project partners fail to work effectively	1	3	3	UKOLN has good links with the University of Manchester through previous joint projects.

4 Engagement with the Community

4.1 Practitioners and Stakeholders

WP1 Through systematic requirements capture, interviews, analysis, and reporting, working with a group of network modellers and biologists in the Sage Federation and at Cancer Research UK.

WP 2 Through Prototyping/Pilot/Demonstrator implementations working with network modellers / biologists DataCite at the British Library and STM publishers Nature PG and PLoS Computational Biology in WP3.

4.2 Benefits Evaluation and Dissemination

WP4 Evaluation of the SageCite approach for network models, highlight the potential intellectual, economic and societal benefits and value of the citation infrastructure and likely impact on scholarly communications.

WP4 Outreach Programme which includes a range of workshops and conferences and the Research Data Management Forum. Project PI's are regularly invited to speak at / host major events and have targeted high-profile international conferences within the bio-domain and in the LIS community.

5 Budget

5.1 Funding Summary

5.2 Value for Money

SageCite provides outstanding value-for money for three key reasons:

- SageCite will build on significant prior investment by the JISC in myExperiment. We will leverage prior investment in Sage Commons data infrastructure, which will act as the SageCite repository test bed.
- The project includes "in kind" contributions from two leading publishers Nature Publishing Group and PLoS, who have agreed to work with SageCite partners.
- Finally, we have an international team with established links into major global linked-data efforts and who will enable the JISC RDM Programme to join-up with leading bio-informatics initiatives such as Bio2RDF and the Concept Web Alliance.

6 Previous Experience of the Project Team

The members represent a partnership of very high calibre teams with demonstrable track records for delivering high-impact, high-profile outputs including articles/references to work in *Nature*, *Science*, *Times Higher*, eScience AHM and JISC/CNI keynotes, input to US, ANDS and Canada national data strategy work. The two STM publishers are leaders in their field.

Liz Lyon is Director of UKOLN and Associate Director of the UK Digital Curation Centre (DCC), University of Bath. She authored the *Dealing with Data* and *Open Science at Web-Scale* Reports and has a doctorate in cellular biochemistry. **Monica Duke** has developed pilots and services within JISC initiatives and projects over the last 10 years. She has expertise in identifiers as well as several standards (such as XML, linked data and OAI-ORE). She contributed to the preliminary Sage Commons citation work.

Carole Goble is Director of the myGrid consortium, University of Manchester, which develops the Taverna Workflow Management System and a number of e-Laboratories that promote the sharing of scientific assets projects, including support for citation and attribution. Examples: myExperiment for workflows, MethodBox for statistical methods and SysMO-SEEK for systems biology data and models. Carole developed data analysis pipelines for the Sage Bionetworks Congress and is on the Sage Advisory Board.

Adam Farquhar is Head of Digital Library Technology at the British Library and was a lead architect on the BL Digital Library System, co-founded its Digital Preservation Team, and initiated the BL Dataset Programme. He is Co-ordinator and Scientific Director of the EU co-funded Planets Project and founder of the Open Planets Foundation. He is President of DataCite and serves on the board of the Digital Preservation Coalition. **Max Wilkinson** is the Programme Manager for the BL dataset programme.

Myles Axton is chief editor of *Nature Genetics*. Following a doctorate at Imperial College (1990) and postdoctoral research at MIT's Whitehead Institute, his interests broadened into human genetics, genomics and systems biology. He helped establish Oxford's innovative research MSc. in Integrative Biosciences, which emphasised the importance of an integrative overview of biomedical research.

Philip E. Bourne PhD is a Professor in the [Department of Pharmacology](#) and [Skaggs School of Pharmacy and Pharmaceutical Sciences](#) at the [University of California San Diego](#) and Associate Director of the RCSB [Protein Data Bank](#). He is Founding Editor-in-Chief of the OA journal *PLoS Computational Biology*.

Stephen Friend is President, CEO and Co-Founder of Sage Bionetworks.

7 Supporting Letters

Letters of Support have been obtained from all institutional partners and associates.

8 References

- [1] Credit where credit is overdue, Editorial (2009) Nature Biotechnology
<http://www.nature.com/nbt/journal/v27/n7/full/nbt0709-579.html>
- [2] Thorisson G. A. Accreditation and attribution in data-sharing (2009). Nature Biotechnology
<http://www.nature.com/nbt/journal/v27/n11/full/nbt1109-984b.html>
- [3] Liz Lyon, Codes, Clouds and Constellations: Open Science in the Data Decade, Keynote, CNI Spring meeting, 2010. <http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/presentations.html#2010-04-13-cni-baltimore>
- [4] Mons B. & Velterop J. Nano-publication in the eScience era (2009). http://www.nbic.nl/uploads/media/Nano-Publication_BarendMons-JanVelterop.pdf
- [5] Human Variome Microattribution Reviews Editorial (2008) Nature Genetics
<http://www.nature.com/ng/journal/v40/n1/full/ng0108-1.html>
- [6] Liz Lyon, Open Science at Web-Scale Report (2009)
<http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/publications.html#november-2009>
- [7] Roach J.C et al, Analysis of genetic Inheritance in a Family Quartet by Whole-Genome Sequencing (2010), Science, <http://www.sciencemag.org/cgi/content/abstract/science.1186802>
- [8] My Data, Your Data, Our Data, (2010) Wall Street Journal
<http://online.wsj.com/article/SB10001424052748703625304575116512173339800.html>
- [9] Sage Citation Workstream on Sage Wiki <http://sagecongress.org/WP/workstreams/Citation>
- [10] Bechhofer, S., De Roure, D., Gamble, M., Goble, C. and Buchan, I. (2010) Research Objects: Towards Exchange and Reuse of Digital Knowledge. In: The Future of the Web for Collaborative Science (FWCS 2010), April 2010, Raleigh, NC, USA. (In Press) <http://eprints.ecs.soton.ac.uk/18555/>
- [11] Leroy Hood, A Doctor's Vision of the Future of Medicine (2009) Newsweek,
<http://www.newsweek.com/id/204227>
- [12] Bourne P. & Fink, L. (2009) I Am Not a Scientist, I Am a Number. PLoS Computational Biology
<http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1000247>