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

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**

Standard PROPOSAL

Document Status: With Council
EPSRC Reference: EP/N002563/1

Organisation where the Grant would be held

Organisation	University of Bath	Research Organisation Reference:	DM4(B)T
Division or Department	Computer Science		

Project Title [up to 150 chars]

DM4(B)T: Data Management for (Build)TEDDI(NET) using Semantic Technologies

Start Date and Duration

a. Proposed start date	01 March 2015	b. Duration of the grant (months)	18

Applicants

Role	Name	Organisation	Division or Department	How many hours a week will the investigator work on the project?
Principal Investigator	Dr Julian Padget	University of Bath	Computer Science	1.77
Co-Investigator	Dr Sukumar Natarajan	University of Bath	Architecture and Civil Engineering	1.77
Co-Investigator	Dr Catherine Pink	University of Bath	Library	1.77

Objectives

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

The Data Management for (Build)TEDDI project (DM4(B)T) aims to initiate the process of building a sustainable and strong data legacy for projects funded by the `Transforming Energy Demand in Buildings through Digital Innovation' (BuildTEDDI, 2011) call. To achieve this aim, the DM4(B)T project has three high-level objectives:

1. LISTEN+ANALYSE: To engage with a range of (Build)TEDDI projects via TEDDINET to discover their problems with and solutions to data capture and data management. Based on this comprehensive review of (Build)TEDDI data, tools and practices, DM4(B)T will identify common problems to explore, as well as potential solutions to develop, in objectives 2 and 3.

2. BUILD+TEST: To develop a novel proof-of-concept demonstrator for cross-repository access to (Build)TEDDI project datasets. This will involve the development, demonstration and release of a prototype tool that will access sample datasets created by projects in the (Build)TEDDI programme. The tool will address challenges imposed by the use of differing ontologies by different project datasets and will build on semantic web technologies to establish a framework for cross-

dataset queries.

3. LEARN+DISSEMINATE: To identify continuing ethical, practical and technical responsibilities for (Build)TEDDI projects, TEDDINET, institutions and the EPSRC in order to sustain the overall (Build)TEDDI data legacy and feed into the on-going policy debate on data management. Achieved both by the projects recommendations, and through workshops on both data management planning and ethics, this will raise issues for TEDDINET partners to consider and will promote engagement and debate between projects, their institutions and the EPSRC. This objective will contribute to the current development of infrastructure to meet the RCUK and UK Government's agendas for open research data, whilst maintaining individual data legacies in line with institutional and national guidelines.

Summary

Describe the proposed research in simple terms in a way that could be publicised to a general audience [up to 4000 chars]. Note that this summary will be automatically published on EPSRC's website in the event that a grant is awarded.

CONTEXT

EPSRC funded 22 projects over two calls in 2010 and 2012 to investigate `Transforming Energy Demand through Digital Innovation' (TEDDI) as a means to find how and how people use energy in homes and what can be done reduce energy consumption. As a result a lot of data is being collected at different levels of detail in a variety of housing up and down the UK, but the mode, detail and quantity are largely defined by the needs of each individual project. At the same time, the research councils (RCUK) are defining guidelines for what happens to data generated by projects they fund, for which universities are then defining policies and finally researchers are then taking concrete actions to store, preserve and document data for future reference.

The problem at this current time is that there is relatively little awareness, limited experience and only emerging practice of how to incorporate data management into much of (physical) science research. This is in stark contrast to established procedures for data formats and sharing in the biosciences, stemming from international collaboration on the Human Genome Project, and in the social sciences, where data from national surveys, including census data, have been centrally archived for many years. Consequently, current solutions adopted by (Build)TEDDI projects may be able to meet a minimal interpretation of the requirements, but not effectively deliver the desired data legacy, such as (for example) the means to execute trans-project queries, or being able to cite the results of such queries for the sake of reproducibility.

AIMS AND OBJECTIVES

The challenges described above, which we address in DM4(B)T in the microcosm of the TEDDI projects, are tackled in three ways:

1. Raising awareness with those who are responsible for data management (principal investigators),

2. Developing a framework to guide the process of making the choices for how to go about implementing data management and

3. Demonstrating example tools that will enable researchers to bring together and re-analyse data from different projects more easily,

which together will help researchers (i) to satisfy funding and institutional guidelines for data management, (ii) begin the process of forming a data management culture in science research and (iii) create a substantial case study in science data management which can inform the three primary stakeholders (researchers, institutions and research councils) across a range of issues (see Recommendations below).

Key activities and outputs:

1. Workshops (i) to gather information about current practice, (ii) present data management problems and outline analysis and solutions and (iii) to disseminate knowledge of tools and (new) practices to support effective data management.

2. Tools and techniques: to allow researchers to harness both the variety and volume of data being collected specifically within the (Build)TEDDI projects. The tools will be made available open-source for access by other researchers to expand and adapt.

3. Recommendations: these will take the form of an online report to identify routes to facilitate a sustainable data legacy (management, curation and citation) for projects in the science and engineering domain.

APPLICATIONS AND BENEFITS

1. (Build)TEDDI projects will benefit directly from the above activities and outputs to meet institutional and research council requirements.

2. Other researchers will benefit from being able to access (Build)TEDDI data.

3. The outputs will benefit the wider research community in science and engineering through the provision of an easy-to-adopt (and adapt) data management methodology.

Academic Beneficiaries

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

Although DM4(B)T is a small project, we nevertheless envisage significant and sustainable impact across a range of beneficiaries.

Immediate beneficiaries of the project will be the (Build)TEDDI community of researchers spread across 20+ universities in the UK and several disciplines. It is noteworthy that the impact will not be limited to the investigators but will also impact the researchers working on the projects, who will be able to carry the value created DM4(B)T on to other projects as their careers progress. In addition we identify researchers in two categories who would further benefit, namely (i) built environment researchers and (ii) non built environment researchers, both in the UK and internationally.

There are a number of institutions receiving UK and EU funding for built environment research who could directly benefit from coordinated access to (Build)TEDDI data. For example, this could amongst others include researchers at (i) the Centre for Energy Epidemiology (CEE) and the CDT in Centre for Urban Sustainability and Resilience (USAR), at UCL, (ii) InSMART- Integrative Smart City Planning, Sustaining Urban Habitats - an interdisciplinary approach (Nottingham), (iii) Carbon, Control and Comfort: User-centred control systems for comfort, carbon saving and energy management (Cardiff and UCL) (iv) Conditioning Demand: Older People, Diversity and Thermal Experience (Manchester) and (v) the WHOLESEM project (UCL, Cambridge, Imperial, Surrey).

Outside the UK, as within, significant resources are being invested in research to understand domestic and commercial energy consumption behaviour and a complete list would be prohibitively long and not especially helpful. For example, Marilyne Andersen's group at EPFL and Dave Culler's group at Berkeley. By enabling the creation of an accessible data legacy for the (Build)TEDDI projects, the impact of DM4(B)T can multiply the impact of the (Build)TEDDI projects and increase the reach of its geographical academic benefits.

Many disciplines are being brought to bear on the energy behaviour problem, in particular in the context of the

(Build)TEDDI projects, which includes psychology, geography, computer science, electrical engineering, social policy (for energy policy and for fuel poverty issues), amongst others. Thus, the activities and outputs of DM4(B)T can also propagate into these disciplines.

Impact Summary

Impact Summary (please refer to the help for guidance on what to consider when completing this section) [up to 4000 chars]

Spreading outwards from the activities taking place in the project, the classes of beneficiaries and consequent impacts are:

1. The investigators and researchers in (Build)TEDDI projects. While some of these are in the investigators' immediate professional circle, the interdisciplinary nature of the portfolio means that many are not. This group are directly exposed to the problems of data management and data legacy creation and can benefit from DM4(B)T's activities to help them analyse those problems and develop appropriate solutions, while also fostering a culture in these projects (and their) researchers that begins to take account of (future) data issues. This will take place during the lifetime of the DM4(B)T project and current (Build)TEDDI projects.

2. The institutions hosting (Build)TEDDI projects. The practices and culture emerging from the (Build)TEDDI projects can inform and contribute to the development of institutional data management guidelines and identification of the resources needed to support them. This can begin to occur during the lifetime of current (Build)TEDDI projects, but will continue afterwards.

3. The EPSRC that is funding the (Build)TEDDI projects. A direct benefit is improved compliance with funder data policies, of benefit to both research councils and researchers within institutions tasked with ensuring compliance. Indirect benefits are enhanced knowledge and awareness of data issues at institutional level enabling capacity for informed discussion between institutions and research councils over guidelines, compliance, resources and implementation. This too can begin during the period of current (Build)TEDDI projects, but will gain momentum subsequently as the data management agenda gathers pace and informed by the recommendations coming out of DM4(B)T's engagement with the (Build)TEDDI project portfolio.

4. Sibling projects in the same institution. Transfer of knowledge and practice to other projects, either through common investigators or local institutional enhancement activities can lead to increased skills amongst RCUK-funded researchers in data management. This would include knowledge of how to curate their data for publication, where and how to publish data, particularly if an institutional data repository is not available, and what ethical issues and licence conditions should be complied with. Expected time frame is during the lifetime of current (Build)TEDDI projects and thereafter.

5. Policy-makers (e.g. DECC/DEFRA) and non-governmental organizations (e.g. Energy Saving Trust, Carbon Trust who often report on field data) or charities (e.g. Centre for Sustainable Energy who worked on the National Household Model). DM4(B)T will initiate the process and prototype the tools to enable increased access to publicly funded data and increased value extraction from existing investment in research by support for and facilitation of publication and re-analysis of data from (Build)TEDDI projects. This in turn can increase knowledge derived from data and hence increased understanding of energy use in homes which may help better inform policy-analysis and policy-making. Impact will start through these stakeholders participating in workshops held with TEDDINET (workshops 1-3) and WHOLESEM (enabled through the CEE, workshop 4), initiating the debate and increasing awareness of what can be achieved through enhanced data management practices. Expected time frame is the duration of the above projects and thereafter.

6. The wider public. Public participation in future studies on domestic energy use will benefit from better understanding of ethical issues by researchers obliged to publish research outputs, thus ensuring privacy of and transparency for participants whilst maximising the potential for data sharing. Time frame is after the end of the current round of projects.

Summary of Resources Required for Project

Financial resources

Summary fund heading	Fund heading	Full economic Cost	EPSRC contribution	% EPSRC contribution
Directly Incurred	Staff	30201.00	24160.80	80
	Travel & Subsistence	4000.00	3200.00	80
	Other Costs	0.00	0.00	80
	Sub-total	**34201.00**	**27360.80**	
Directly Allocated	Investigators	0.00	0.00	80
	Estates Costs	3801.00	3040.80	80
	Other Directly Allocated	0.00	0.00	80
	Sub-total	**3801.00**	**3040.80**	
Indirect Costs	Indirect Costs	24050.00	19240.00	80
Exceptions	Other Costs	0.00	0.00	100
	Sub-total	**0.00**	**0.00**	
	Total	**62052.00**	**49641.60**	

Summary of staff effort requested

	Months
Investigator	0
Researcher	8
Technician	0
Other	0
Visiting Researcher	0
Student	0
Total	8

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 (£)**
EPSRC	EP/K002724/1	Energy literacy through an intelligent home energy advisor (ENLITEN)	Y	Y	01/09/2012	31/08/2016	1511972

Staff

Directly Incurred Posts

			EFFORT ON PROJECT							
Role	**Name /Post Identifier**	**Start Date**	**Period on Project (months)**	**% of Full Time**	**Scale**	**Increment Date**	**Basic Starting Salary**	**London Allowance (£)**	**Super-annuation and NI (£)**	**Total cost on grant (£)**
Researcher	PDRA to be recruited	01/03/2015	8	100	N/A	01/03/2015	36309	0	8684	30201
									Total	30201

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 Julian Padget	Y	100	0	0	68860	0
Co-Investigator	Dr Sukumar Natarajan	Y	100	0	0	57324	0
Co-Investigator	Dr Catherine Pink	Y	100	0	0	42599	0
						Total	0

Travel and Subsistence

Destination and purpose		Total £
Within UK	Visits to partner projects	1000
Within UK	Workshop attendance travel costs	1600
Within UK	Workshop hosting costs	1400
	Total £	4000

Research Council Facilities

details of any proposed usage of national facilities
Research Council Facilities are not relevant to this application.

Human Participation

Would the project involve the use of human subjects?		Yes	No✔
	If yes, would equal numbers of males and females be used?	Yes	No✔
Would the project involve the use of human tissue?		Yes	No✔
Would the project involve the use of biological samples?		Yes	No✔
Would the project involve the administration of drugs, chemical agents or vaccines to humans?		Yes	No✔
Will personal information be used?		Yes	No✔
	If yes, will the information be anonymised and unlinked?	Yes	No✔
	Or will it be anonymised and linked?	Yes	No✔
	Will the research participants be identifiable?	Yes	No✔
Please provide details of any areas of substantial or moderate severity:			

Animal Research

Would the project involve the use of vertebrate animals or other organisms covered by the Animals (Scientific Procedures) Act?		Yes	No✔
	If yes, what would be the maximum severity of the procedures?	Mild or non-recovery	
		Moderate	
		Severe	
	Please provide details of any areas which are Moderate or Severe:		

Animal Species

Does the proposed research involve the use of non-human primates?	Yes	✔No
Does the proposed research involve the use of dogs?	Yes	✔No
Does the proposed research involve the use of cats?	Yes	✔No
Does the proposed research involve the use of equidae?	Yes	✔No

Please select any other species of animals that are to be used in the proposed research.

Fish	Sheep
Rabbit	Rat
Amphibian	Poultry
Cow	Mouse
Reptile	Guinea Pig

Pig	Other Rodent
Bird	Other Animal

Genetic and Biological Risk

Would the project involve the production and/or use of genetically modified animals?	Yes	✔	No
If yes, will the genetic modification be used as an experimental tool, e.g., to study the function of a gene in a genetically modified organism?	Yes	✔	No
And will the research involve the release of genetically modified organisms?	Yes	✔	No
And will the research be aimed at the ultimate development of commercial or industrial genetically modified products or processes?	Yes	✔	No
Would the project involve the production and/or use of genetically modified plants?	Yes	✔	No
If yes, will the genetic modification be used as an experimental tool, e.g., to study the function of a gene in a genetically modified organism?	Yes	✔	No
And will the research involve the release of genetically modified organisms?	Yes	✔	No
And will the research be aimed at the ultimate development of commercial or industrial genetically modified products or processes?	Yes	✔	No
Would the project involve the production and/or use of genetically modified microbes?	Yes	✔	No
If yes, will the genetic modification be used as an experimental tool, e.g., to study the function of a gene in a genetically modified organism?	Yes	✔	No
And will the research involve the release of genetically modified organisms?	Yes	✔	No
And will the research be aimed at the ultimate development of commercial or industrial genetically modified products or processes?	Yes	✔	No

Approvals

Have the following necessary approvals been given by:			
The Regional Multicentre Research Ethics Committee (MREC) or Local Research Ethics Committee (LREC)?	Yes	No	Not required✔
The Human Fertilisation and Embryology Authority?	Yes	No	Not required✔
The Home Office (in relation to personal and project licences, and certificates of designation)?	Yes	No	Not required✔
The Gene Therapy Advisory Committee?	Yes	No	Not required✔
The UK Xenotransplantation Interim Regulatory Authority?	Yes	No	Not required✔
Administration of Radioactive Substances Advisory Committee (ARSAC)?	Yes	No	Not required✔
Other bodies as appropriate? Please specify.			

Other Issues

Are there any other issues of which the Council should be aware? No

Provide details of what they are and how they would be addressed [up to 1000 characters]

OTHER INFORMATION

Reviewers

1	Name	Organisation	Division or Department	Email Address
Professor Chris Tweed		Cardiff University	Welsh School of Architecture (ARCHI)	tweedac@cardiff.ac.uk

Reviewers

2	Name	Organisation	Division or Department	Email Address
Professor Nigel Gilbert		University of Surrey	Sociology	n.gilbert@surrey.ac.uk

Reviewers

3	Name	Organisation	Division or Department	Email Address
Professor David De Roure		University of Oxford	Oxford e-Research Centre	david.deroure@oerc.ox.ac.uk

PATHWAYS TO IMPACT

DM4(B)T: DATA MANAGEMENT FOR (BUILD)TEDDI(NET) USING SEMANTIC TECHNOLOGIES

It is intended that DM4(B)T will have substantial impacts beyond the academic community, extending across a range of stakeholders and beneficiaries. Routes to impact are identified, each engaging different stakeholders and closely aligned to the objectives of the project.

Stakeholder	Benefit	Objectives	Route
EPSRC	Increased policy compliance. Data legacy for (Build)TEDDI projects. Additional value extracted from original investment in TEDDINET programme.	O3	Impact 5 Impact 3 Impact 7
Data Management Community	Guidance and recommendations applicable beyond TEDDINET and will inform development of institutional and national infrastructure.	O1, O3	Impact 1 Impact 2 Impact 4 Impact 6 Impact 7
TEDDINET community	Researchers will develop data management skills via a planning workshop. Challenges and issues likely to be common to many projects (ethics, data archive, access, consent) will be explored and example solutions found.	O1, O2, O3	Impact 1 Impact 2 Impact 5 Impact 3 Impact 4 Impact 6 Impact 7
Public and Energy Community	Extract additional value from original public investment in TEDDINET projects Potential to interrogate multiple datasets will enable new research questions to be asked and investigated. Open data legacy will fulfil government and funder requirements for transparency of the research process.	O1, O3	Impact 1 Impact 2 Impact 5 Impact 6

IM-1: **Research Publications:** DM4(B)T will develop a tool to access sample (Build)TEDDI datasets. As these datasets will be structured in standard formats the experience and tool will be applicable to both other datasets within (Build)TEDDI and, more broadly, to other disciplines. As well as archiving and sharing the tool software, DM4(B)T will communicate the project findings by publishing at least two articles in relevant peer-review journals such as: (a) The International Journal of Digital Curation[1] (b) Program[2] (c) The Journal of Web Semantics[3] (d) Energy and Buildings[4].

IM-2: **Project Website:** DM4(B)T will maintain a project website and blog, hosted at the University of Bath. The blog will provide an additional avenue by which the TEDDINET community will be regularly updated on project progress and the website will facilitate dissemination of outputs, including recommendations and access tools. Both the project website and blog, and the code for the access tool, will be archived at the end of the project to ensure that DM4(B)T's outputs are preserved and remain accessible beyond the end of the project.

IM-3: **Data Management Planning Workshop:** Data management plans (DMPs) are valuable tools that prompt early consideration of data management issues, enabling actions to be taken that will maximise the potential for data sharing and re-use. Principle V of the EPSRC Policy Framework on Research Data requires that "Institutional and project specific data management policies and plans

[1] International Journal of Digital Curation: http://www.ijdc.net/

[2] Program: http://emeraldgrouppublishing.com/products/journals/journals.htm?id=prog

[3] Journal of Web Semantics: http://www.journals.elsevier.com/journal-of-web-semantics/

[4] Energy and Buildings: http://www.journals.elsevier.com/energy-and-buildings/

should...exist for all data."[5] However, as institutions are still working towards full compliance with this policy it is unlikely that many projects within the TEDDINET programme will have written a data management plan. DM4(B)T will run a workshop on data management planning (**workshop 2** in workplan and JoR) for the TEDDINET community. This workshop will utilise the Research Data Scientist's (Pink) extensive experience of reviewing data management plans for a range of funders and disciplines. The workshop will be based around the recently developed EPSRC-template[6] that is available via the Digital Curation Centre's web-based planning tool DMPonline[7]. The DMP workshop will (i) provide TEDDINET researchers with the skills they will require to write data management plans, (ii) introduce TEDDINET researchers to the DMPonline tool and raise awareness of the EPSRC Policy Framework on Research Data. The data management planning workshop will support TEDDINET Network Collaboration Objective 7.

IM-4: **Ethics, Publication and Citation Workshop:** The collection of quantitative sensor data from participant homes and qualitative survey data, combined with requirements to make these data publicly accessible, raises ethical questions about consent and privacy, such as those relating to data retention and anonymisation of indirect identifiers. These ethical issues may be unfamiliar to those researchers within (Build)TEDDI who may not have extensive experience of participant-based or social science research. Research (RC-2 and RC-3) and non-technical (NTC-3) challenges explored during Tasks 1 and 4 of DM4(B)T will be used to formulate recommendations on ethical, access, consent, anonymisation, and publication issues. These recommendations will be presented to the TEDDINET community via a workshop (**workshop 3** in workplan and JoR), to be run mid-way through DM4(B)T. The ethics, publication and citation workshop will support TEDDINET Knowledge Exchange Objective 4.

IM-5: **Data Curation and Legacy Workshop:** As a key element of Tasks 3 and 4, DM4(B)T will jointly run a workshop (**workshop 4** in workplan and JoR) with both TEDDINET and the Centre for Energy Epidemiology, which will form part of the existing programme of TEDDINET events. The workshop will focus on specific data issues for the TEDDINET community and will provide an opportunity to share preliminary recommendations to sustain the Build(TEDDI) data legacy (Task 4, phase 2), as well as enabling feedback to be sought to finalise the recommendations (Task 4, phase 3). As part of this workshop DM4(B)T will provide a practical demonstration of the prototype tools developed to access sample (Build)TEDDI datasets. The data workshop will support TEDDINET Knowledge Exchange Objective 4 and TEDDINET Network Collaboration Objective 5.

IM-6: **Recommendations:** DM4(B)T will put forward criteria against which (Build)TEDDI projects can assess the suitability of the recommendations from DM4(B)T on how to archive and publish their datasets, including recommended repositories, ontologies, and supporting metadata, as well as how to cite multi-dataset analyses. In addition to directly supporting the (Build)TEDDI community, the recommenations will provide a blueprint for other multi-project programmes to follow.

IM-7: **Steering Group:** DM4(B)T will be overseen by a steering group that will include representatives from the Digital Curation Centre and the TEDDINET research network as well as the EPSRC's Research Outcomes senior manager. The steering group will have direct involvement in the practical application of RCUK policy to a discipline that is not yet fully supported by national data repositories. DM4(B)T will take place at a critical stage in the timescale for data management in the UK, spanning the run up to and months following the EPSRC's data policy compliance deadline. Members of the steering group, which will include some of the main stakeholders in mandating and supporting data management, will therefore benefit from timely information on challenges and soltuions to full data accessibility, enabling developments in national policy, guidance and infrastructure to be informed by the project findings and outputs.

[5] ESPRC Policy Framework on Research Data principle `http://www.epsrc.ac.uk/about/standards/researchdata/principles/`, retrieved 20140922.

[6] ESPRC template announcement `http://www.dcc.ac.uk/news/new-epsrc-template-dmponline`, retrieved 20140922.

[7] DMP online: `http://www.dcc.ac.uk/dmponline`, retrieved 20140922.

TRACK RECORD

DM4(B)T: DATA MANAGEMENT FOR (BUILD)TEDDI(NET) USING SEMANTIC TECHNOLOGIES

TEDDINET is a research network comprising 22 individual research projects funded under the 'Transforming Energy Demand through Digital Innovation' (TEDDI) and 'Transforming Energy Demand in Buildings through Digital Innovation' (BuildTEDDI) programmes. These 22 projects encompass 26 UK universities, 75 partners from industry and the housing sector, and over 200 researchers from engineering, informatics, design and social sciences and represent an overall investment of £21M by the UK Engineering and Physical Sciences Research Council (EPSRC). TEDDINET's primary purpose is to create added value and enhance the impact of this research by enabling interaction between these projects.

TEDDINET serves to encourage and enable communication and collaboration internally between participating researchers and practitioners, as well as externally between the research projects and industry, policy-makers, civil society and wider academia. It undertakes a broad range of activities to collate, synthesise and share research findings, and develop the evidence base to inform government policies, societal debates and industrial strategies.[1]

University of Bath: The University has recently established the EPSRC Centre for Doctoral Training in the Decarbonization of the Built Environment in the Department of Architecture and Civil Engineering with an investment of £1.5M. The centre will support 20 fully funded doctoral studentships over the next 5 years with a strong focus on low carbon building design and the integration this needs from disciplines such as computer science, psychology, management and other areas of engineering. The Department of Architecture and Civil Engineering has invested heavily in this area over the last 5 years with 5 new members of academic staff including one professorial appointment.

The Project Team: comprises a computer scientist (Padget), an architectural engineer (Natarajan) a research data scientist (Pink) and a PDRA. The PI and the CI have been working together on energy and buildings since 2008 utilising a range of technologies, such as agent-based simulation [4], energy interface design [1] and semantic web tools applied to legacy data access [2, 3]. Pink is the research data scientist at the University of Bath. Locally, this project will be embedded in the ENLITEN project, funded under the BuildTEDDI programme, in which both proposers are CIs, and nationally in the TEDDINET community, giving access to a large body of multi-disciplinary research and researchers. The P/CIs are not charging their time to the project in order to maximize the available budget for the PDRA.

Team Perspective: We come to this project as (i) problem owners – we have datasets to curate (ii) with responsibilities to fulfil – as EPSRC fund-holders (Padget, Natarajan) and institutional advisers on research data policy (Pink) (iii) and a user's knowledge of the principles and tools for semantic annotation and querying of datasets. The rationale for this project follows the principle of "teach a man to fish" rather than "give a man a fish", where the (Build)TEDDI community is the recipient of the knowledge, semantic data management technologies are the fish, and teach means guided learning, through exploration of the resources available, informed by best practices and national resources (such as the Digital Curation Centre (DCC) and the Oxford eScience Research Centre (OeRC)). We believe ownership of the problem and subsequent ownership of a range of solutions, through the lens of domain experience, that meet our needs is an effective way to: (i) transfer knowledge on data management from the DCC and others and via them from solution developers and (ii) build a sustainable body of knowledge and experience in a user community.

Dr. Julian Padget [PI] is a Senior Lecturer in the Department of Computer Science at the University of Bath. The central theme of his research is formal and computational models of policy in the context of distributed intelligent systems and sensor networks, and its impact on agent architectures and collective decison-making (prediction markets using machine learning agents). He and Natarajan have been collaborating since 2008, when they first developed an agent-based simulation

[1] Extracted from `http://teddinet.org/`, retrieved 20140808.

for housing stock carbon-footprint evolution, followed by energy monitoring pilots in student accommodation on campus and in commercial offices and most recently the EPSRC-funded ENLITEN project. He has published widely on multi-agent systems and their governance as well as its application in other fields (including social policy, management and engineering), and in the technologies of sensor networks and immersive environments. Current and recent relevant projects include: ENLITEN (EPSRC, Co-I, occupant models and ABM), Formal techniques for sensor network design, management and optimization (Leverhulme, PI), Science Education in the Cloud (iNets, with Sciencescope Ltd), Evaluating smart metering interfaces (with Natarajan) and ALIVE: Coordination, Organisation and Model Driven Approaches for Dynamic, Flexible, Robust Software and Services Engineering (EU FP7, Bath PI).

Dr. Sukumar Natarajan [CI] is Lecturer in Environmental Design, Deputy Director of the research unit for Energy and the Design of Environments (EDEn) and Director of Studies for the MSc in Architectural Engineering: Environmental Design in the Department of Architecture & Civil Engineering at the University of Bath. His research investigates carbon emissions from the built environment, climate data for building energy modelling, domestic indoor thermal comfort, smart homes, smart metering interfaces and the impact of occupants on building energy use. He has been examining the problem of large scale energy use and carbon emissions modelling since 2007, as evidenced by the widely cited Domestic Energy and Carbon (DECarb) model of the UK, on which he was the lead author. More recently, he has worked on developing novel sensing solutions that can be used to better understand the sources, timing and scale of energy consumption in buildings. This has formed the basis of the collaboration with Padget and Walker, most recently on the £1.5M EPSRC ENLITEN project on which he is a Co-Investigator, leading the work on sensor design. He was Bath's PI on the EPSRC-funded COPSE project which examined the impact of changing climate on thermal comfort and the large scale energy use implications on domestic and non-domestic building stock. He has advised Manchester City Council and the London Borough of Islington on energy and comfort policy and has published widely on building modelling, energy use, retrofitting, and thermal comfort. He is an expert reviewer for a number of international journals as well as EPSRC. He is a contributor to the influential CIBSE Guide A (2006) and is a steering committee member of the EPSRC Network on Comfort and Energy Use in Buildings (NCEUB) and CIBSE Guide A (Chap. 2).

Dr. Catherine Pink is the Research Data Scientist in the Library at the University of Bath. She managed the Research360 project, part of Jisc's 2011-2013 Managing Research Data Programme, which piloted and developed infrastructure to improve institutional capacity for data management. She co-authored the University of Bath's Roadmap for EPSRC and wrote the University of Bath's Research Data Policy, both of which have been identified as exemplars and re-used by other UK universities. She has a background in industrial agrochemical research and bioinformatics research, the latter re-structuring a range of different datasets for re-analysis. Her current experience in developing data management infrastructure, particularly aimed at institutional compliance with the EPSRC Policy Framework on Research Data, will allow her to provide advice and training on research data management issues, particularly relating to objectives 2 and 3 of DM4(B)T.

CASE FOR SUPPORT

DM4(B)T: DATA MANAGEMENT FOR (BUILD)TEDDI(NET) USING SEMANTIC TECHNOLOGIES

Proposal context and format: The proposal has been developed in consultation with the EPSRC's Energy Portfolio Manager and the investigators leading the TEDDINET project. This project would take place within the context of the TEDDI/BuildTEDDI/TEDDINET projects, but is, as determined by the EPSRC's Energy theme manager, outside the TEDDINET remit. This last is the basis for the separate funding of the data management activities that are identified in this proposal. Furthermore, the scope (see below) and format are a result of discussion with EPSRC, leading to the idea of a short-term project as a way to pump-prime the activity and raise awareness of data management issues through practical experience. The P/CIs are not charging their time to the project in order to maximize the available budget for the PDRA.

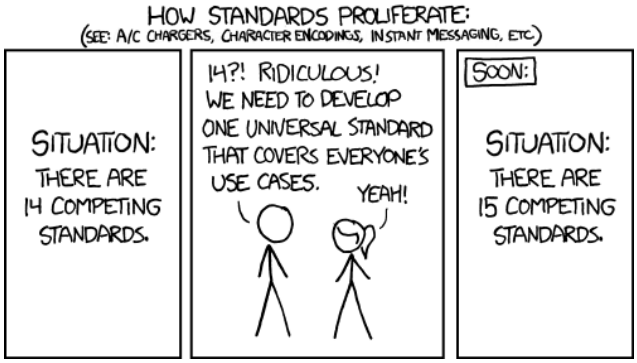
DM4(B)T Background, Aims and Objectives

The 'Transforming Energy Demand through Digital Innovation' (TEDDI, 2009) and 'Transforming Energy Demand in Buildings through Digital Innovation' (BuildTEDDI, 2011) funding calls are part of a major funding initiative by RCUK in the framework of the Digital Economy and Energy programmes, aimed at exploring the potential impact of digital innovation on energy demand reduction. In all, the (Build)TEDDI portfolio comprises 22 projects with a funding value of £22M, involving 29 universities, 70 commercial and governmental organizations, over the period 2009–2016.

The aim of DM4(B)T is to initiate the process of building a **sustainable and strong data legacy** for the (Build)TEDDI projects. This activity will be informed by the RCUK principles[2] and the EPSRC Policy Framework for Research Data[3], while supporting individual institutions' policies and practices for data management. This activity comes at a particularly early and crucial stage in the development and implementation of institutional policies.

The high-level objectives of this project are to:

O1: Develop proof-of-concept demonstrators for cross-repository access to (Build)TEDDI project datasets
O2: Develop a framework for (Build)TEDDI projects to follow for new and existing datasets
O3: Identify continuing ethical, practical and technical responsibilities for (Build)TEDDI projects, TEDDINET, institutions and EPSRC to sustain the (Build)TEDDI data legacy



http://xkcd.com/927/ (Licensed for re-use under Creative Commons Attribution-NonCommercial 2.5 License.)

We firmly believe that initiating the definition of a standard is the wrong approach for achieving objectives 1 and 2 above because: (i) it is impractical for completed or mature projects that have no resources to apply the standard, meaning it becomes someone else's problem (ii) the outcome of such processes can have limited utility because agreement can often only be reached on the lowest common denominator[4] (iii) a standard is typically prescriptive and requires more foresight than is currently available. In contrast, we propose to use semantic annotation and query technologies that can provide the complementary *technical* flexibility to align similar but not identical datasets. This approach at the technical level reflects the adoption of guidelines and flexible frameworks by the Research Councils and bodies such as the Digital Curation Centre, that can be adapted to meet specific needs that could not be foreseen at the time of writing (the guidelines), although these needs are necessarily limited to known unknowns. In

[2] http://www.epsrc.ac.uk/about/standards/researchdata/principles/, retrieved 20140805.
[3] http://www.epsrc.ac.uk/about/standards/researchdata/expectations/, retrieved 20140805
[4] The PI experienced this working with BSI and ISO on programming language standards in the late 1980s/early 1990s.

Relevant TEDDINET Knowledge Exchange Objective:

KEO-4. Create a strong legacy for (Build)TEDDI by promoting open data practices, influencing the international research agenda and disseminating outcomes beyond project consortia and project end-dates;

Relevant TEDDINET Network Collaboration Objectives:

NCO-5. Facilitate collaboration between TEDDI and BuildTEDDI research projects to share research experience, establish best practice, stimulate academic output, enhance project delivery, reduce duplication and ensure different projects' research is complementary;

NCO-7. Train and provide support to early stage researchers (ESR), especially post-docs and PhD students;

Figure 1: TEDDINET data management objectives (taken from TEDDINET CfS)

addition, the use of semantic technologies offers the opportunity for automation of some aspects of the documentation of datasets by formalising in both machine- and human-readable form information about their structure and content. Thus, the DM4(B)T objectives above will contribute to and complement specific TEDDINET objectives (see Figure 1), through:

1. Establishing a concrete basis and methodology for the (four) TEDDINET Knowledge Exchange Objectives, but in particular advancing **KEO-4** beyond what TEDDINET sets out to achieve.
2. Facilitating and enhancing the TEDDINET Network Collaboration Objectives, especially **NCO-5** and **NCO-7**, through example and dissemination of good practice, regarding data management.

The (Build)TEDDI Data Problem The projects under the (Build)TEDDI umbrella are generating substantial (but not huge) quantities of data (1Gb/day for ENLITEN (actual), 3.5Gb/day for IDEAL (projected), for example) in a range of formats subject to the (emerging) data management guidelines of many institutions. These data sets are typically inside (and, unless archive and ethical issues are resolved, may have to remain inside) institutional firewalls and many are live, in that they are collecting data, possibly in real-time, and may be doing so until at least 2016Q1. Each data set has been designed to meet the needs of the project it is supporting, so that any similarities in structure are more likely due to chance than by design.

Project Drivers: In addition to the immediate issues raised by data generated in (Build)TEDDI projects, the EPSRC and beyond that the Research Councils and the government as a whole have identified open data[5] as an emerging practical and policy priority[6], while the responsbility is upon individual institutions to specify guideline-compliant implementations. The EPSRC's research data policy framework is expressed in terms of *expectations*, of which item (viii) is of particular relevance:

> Research organisations will ensure that effective data curation is provided throughout the full data lifecycle, with 'data curation'[7] and 'data lifecycle'[8] being as defined by the Digital Curation Centre. The full range of responsibilities associated with data curation over the data lifecycle will be clearly allocated within the research organisation, and where research data is subject to restricted access the research organisation will implement and manage appropriate security controls; research organisations will particularly ensure that the quality assurance of their data curation processes is a specifically assigned responsibility.

[5] Francis Maude (2010) stated aim to make the UK "the most transparent and accountable government in the world" leads to `https://www.gov.uk/government/publications/open-data-white-paper-unleashing-the-potential`, retrieved 20140908

[6] Evidence and economics-based policy-making goes back to 1999's Cabinet Office report: Professional Policy Making For The Twenty First Century.

[7] defined as "Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle.".

[8] defined as an 11-step process at `http://www.dcc.ac.uk/digital-curation/what-digital-curation`, retrieved 20140805

The reality is that institutions are just getting to grips with these requirements (although institutions need to be compliant with EPSRC policy by May 2015). PIs similarly are in the early stages of working out how to achieve compliance and both PIs and institutions are engaged in establishing the balance between project and institutional effort to achieve the stated research council outcomes. This situation underscores the timeliness (at the higher policy level) of DM4(B)T in relation to the three primary stakeholders groups, namely PIs, institutions and the Research Councils, in addition to its timeliness in respect of (Build)TEDDI.

DM4(B)T Methodology This project is a 12-month pilot + 6 month dissemination activity. Our aim is to establish routes to viability first and then refine on the criteria of practicality and sustainability. Thus, the primary outcomes, in line with the objectives given at the opening, are: (i) practical demonstration of and the release of prototype tools to access sample (Build)TEDDI datasets, (ii) recommendations for (Build)TEDDI projects to follow in releasing their datasets, (iii) recommendations for (Build)TEDDI projects to follow in maintaining their data legacy in line with institutional and national guidelines, and (iv) open issues for TEDDINET partners to consider and to engage in debate with their institutions and with EPSRC. The deliverables will take the form of software and reports, all of which will be disseminated through a Bath-hosted website. These outcomes will be achieved through working with the TEDDINET and EPSRC community as a whole, initially by undertaking a survey of each (Build)TEDDI project and following up by working closely with volunteer projects to transfer knowledge and practice to them – including of course ENLITEN (Natarajan and Padget).

The (Build)TEDDI projects have some important characteristics for this project, specifically they are: (i) **data-intensive** (quantitative from sensors and qualitative from surveys), but each is different: providing different perspectives and challenges (ii) at **different stages** in their life-cycles: some are at end-of-life (requires adaptation of DM4(B)T to accommodate fixed data asset), others are mid-life (requires mutual adaptation of DM4(B)T and the project, with some fixed and some developing data assets) and some are early-life (requires active feedback with DM4(B)T) and (iii) **strongly connected** to TEDDINET (and direct responsibilities of the investigators, in the case of ENLITEN): providing risk mitigation for DM4(B)T because outcomes are mutually beneficial.

Research Challenges We describe the challenges that we consider the key (very) early-stage requirements, expressed as normative statements. Underpinning all of these is the principle that open-linked data is both desirable and in line with research council and hence government policy. Each is linked to tasks in the work programme:

RC-1: **Discoverability:** Berners-Lee's vision of the web is a collection of discoverable resources: the (Build)TEDDI datasets should be search-engine friendly (in respect of the resource as a whole, not individual records). This implies RESTful services, rather than crawler-opaque SOAP-style services. Control of discoverability should be in the hands of the dataset creators, but affected by the availability of (national) archive infrastructure. Tasks T2–T4.

RC-2: **Accessibility:**[9] Access should be established by the publication of the resource, thus making it available for search-engine indexing and consumer querying. In principle, in the case of (Build)TEDDI, the data is public because the research that captured it is publicly-funded, but there should be scope to express access restrictions through policies, reflecting the decisions of institutional ethics committees and other regulatory bodies. Tasks T3, T4

RC-3: **Citability:** Once the data is discoverable and accessible, it can be used to underpin analyses and visualizations. We contend that reproduceability should be as critical an aspect of data science as it is of physical science and so the notion of citation should be extended to: (i) the query over a set of datasets and (ii) the time when a query is executed. The former requires capture of the process leading to a result, while the latter requires capture of temporal constraints on the data processed, so that queries on dynamic datasets produce the same results whenever they are processed[10]. Task T4

[9] Accessibility here means who has access (authentication and authorization), rather than mitigation of physiological impairments.

[10] The Research Data Alliance are also working on dynamic data citation. See: `https://rd-alliance.org/sites/`

DM4(B)T will frame these questions more precisely as a result of working with actual research projects, oversight bodies (DCC and EPSRC) and (indirectly) with institutions, in order to deliver preliminary, approximate answers and develop more detailed questions for subsequent exploration.

Technical Challenges These broadly fall into the following categories, which in each case are linked to tasks in the work programme:

TC-1: **Format:** Projects will have chosen one or more of (at least) SQL, no-SQL, CSV, RDF triples, and/or JSON based on requirements, experience and tools, but data consumers should not need to be aware of of these choices. Tasks T1–T2.

TC-2: **Location:** A number of different repository platforms are either in use or being adapted for data archiving and preservation, presenting different technical environments within which data linkage tools must function. Tasks T1–T2.

TC-3: **Access:** Active datasets are and should remain protected by institutional firewalls. It remains to be seen whether finalized datasets stay there or migrate to national or institutional repositories[11]. Location aside, the datasets' collective value depends on provision of access to data consumers, requiring the addressing of issues such as consumer authentication and authorization, appropriate data anonimization (which may require additional institutional ethical committee procedures) and possibly restrictions on fields or combinations of fields in given datasets. Effective provision of cross-dataset access is a pre-requisite for the TEDDINET activity on shared data analysis, for example. Tasks T1–T3.

TC-4: **Linkage:** For DM4(B)T outputs to deployed across other TEDDI projects, mechanisms to update and publish new datasets will be required. Whether (Build)TEDDI projects manage the process via XML upload, or automated, dynamic harvesting using OIA-PMH[12], metadata normalisation is integral to functionality to enable cross-dataset querying. DM4(B)T will investigate and build on experiences and tools used by the Jisc/DCC Research Data Registry[13] and the Australian National Data Service[14] to identify solutions. Tasks T2–T3.

TC-5: **Reproducibility:** The data consumer should be able to carry out the same queries whether a dataset is dynamic or finalized and regardless of its location. For successive queries on dynamic data this raises issues of reproducibility that will, in part, require technical solutions[15]. Cross-database interogations, particularly on dynamic datasets, also raise questions of provenance, but time constraints mean that this challenge will be left to a following study. Tasks T2, T4.

Non-technical Challenges These too are linked to tasks in the work programme:

NTC-1: **Project-specific Resources:** Funding for many TEDDI projects ends in 2014 and for most BuildTEDDI projects by early 2017. Consequently, most TEDDI project deliverables are being wrapped up now, while those for BuildTEDDI will be completing from about 2016Q1. In some cases, this may mean that the PDRAs with the detailed knowledge of the datasets may only be available for a relatively short period, before the datasets are passed into institutional care. Task T1.

NTC-2: **Data Legacy:** identified as activity 7 by TEDDINET, this reiterates the research council's requirements surrounding research data, emphasising "the correct documenting and storing of data arising from measurements and models carried out within projects.". We have noted earlier how semantic technologies can potentially contribute to the documentation process. To this, DM4(B)T will add small-scale demonstrators that build on the semantic interfaces to establish a pathway to *active* data legacies taking account of the planned collaboration between the Centre for Energy Epidemiology at University College London and TEDDINET.

default/files/Workshop%20Report%20London%20July%201-2%202014.pdf

[11] c.f the UK Data Service for the social sciences, DataShare at the University of Edinburgh, and FigShare

[12] Open Archives Initiative - Protocol for Metadata Harvesting: http://www.openarchives.org/pmh/

[13] DCC Research Data Registry project: http://www.dcc.ac.uk/projects/research-data-registry-pilot

[14] The Australian National Data Service automatically harvests metadata via a number of methods: http://www.ands.org.au/support/configharvest.html

[15] See discussion of 'citability' under **Research Challenges**

Tasks T2–T4.

NTC-3: **Persistent Identification:** For policy compliance both the datasets and metadata, as well as methods of locating and identifying them, must persist for at least 10 years. Decisions will be required on the application and granularity of persistent identifiers, such as accession IDs, Digital Object Identifiers and ORCIDs. Recommendations will build on best practice employed at existing data archives. Task T4.

Work Programme The case so far has presented a summary of the problem both at the micro and macro level and a vision of the goals and challenges on the path to the publication and long-term use of open-linked data from publicly-funded research. We now describe the tasks to be done in the context of DM4(B)T, and in particular for the (Build)TEDDI projects, to demonstrate the feasibility of that vision and create the awareness that is necessary for the community to develop and sustain the habits to bring the vision about:

T1: **Gathering practice and tools:** DM4(B)T will undertake a comprehensive review of potential data types (e.g. sensor data, social survey data, photographic data), formats (e.g. CSV, SPSS, RDF triples) and the range of available tools (building on a previous review undertaken by the investigators (Padget and Natarajan [2])). DM4(B)T will also collect information from selected BuildTEDDI projects on data, tools and practices. These activities will be supported by DM4(B)T hosting a TEDDINET-sponsored project launch workshop (**workshop 1** in workplan and JoR). Timing: months 1–3.

T2: **Prototype demonstrators:** working in conjunction with selected (Build)TEDDI projects

(a) DM4(B)T will use existing semantic web technology tools to establish a framework for building wrapping services for datasets. These services will support the translation of queries into the representation adopted in the dataset and the subsequent translation of results into the terms in which the query was expressed. We are aware of the ontology alignment challenges this poses: this project aims to establish an understanding of their scope and scale in the context of (Build)TEDDI projects. In doing so, we draw on our experience of constructing a prototype [2] linking the English House Condition Surveys from 1970 to 2006 – wherein later editions have added and refined the (data) columns of earlier editions, making direct comparison infeasible because text labels do not match – as a preliminary indicator of the viability of the approach.

(b) DM4(B)T will demonstrate functionality by selecting a repository environment and development of the data access tool will be informed by repository integration with data visulisation platforms e.g. *F1000Research* with Figshare and data.gov.uk's use of CKAN.

(c) DM4(B)T will develop a web portal to enable TEDDINET to deliver direct access to all data outputs from (Build)TEDDI projects.

Timing: months 4–12

T3: **Disseminating practice:** As set out in the *Pathways to Impact*, DM4(B)T will run a series of workshops for the TEDDINET community. Workshops will be integrated within existing TEDDINET events and will cover (i) data management planning (**workshop 2** in workplan and JoR); (ii) recommendations on ethics, data publication and data citation (**workshop 3** in workplan and JoR); Timing: months 3–12. This task will contribute to and expand upon the theme of data curation and legacy within TEDDINET, working in conjunction with the Centre for Energy Epidemiology (CEE)[16]. A specific planned activity is a joint TEDDINET/CEE workshop on data access and management (**workshop 4** in workplan and JoR), which will include a demonstration of the prototype tools. DM4(B)T will facilitate access to the outputs of example (Build)TEDDI projects via the web portal developed under Task T2, and will also maintain a project website and blog. Timing: months 3–12 (core) + months 13–18 (wrap-up).

T4: **Recommendations:** DM4(B)T will prepare a phased on-line report that (i) analyses what has been learned from the project regarding practice (task 1), tools (task 2) to identify various routes and conditions for sustaining data legacies from RCUK funded research and (ii) documents and

[16]This engagement is part of existing TEDDINET arrangements and coordinated by Nigel Goddard (Edinburgh), one of the TEDDINET Co-Is.

provides early and on-going assessment of the (Build)TEDDI data legacy (task 3). Timing: months 4–5 (phase 1: practice), 9–12 (phase 2: tools + (preliminary) legacy), 17–18 (phase 3: (final) legacy).

DM4(B)T will demonstrate outputs by working with TEDDINET and several (Build)TEDDI projects.

Project Management DM4(B)T will collaborate closely with the TEDDINET principals, to effect coordination of the workshop and consultation activities with (Build)TEDDI projects.

1. **Data Management plans:** For compliance with EPSRC policy and to lead by example, DM4(B)T (Pink) will develop and maintain a data management plan.
2. **Weekly Project Meetings:** The DM4(B)T team will meet weekly to review progress of tasks.
3. **TEDDINET Oversight:** DM4(B)T will have quarterly meetings with the TEDDINET management team to provide a direct connection to TEDDINET and coordinate on joint activities.
4. **Steering Group:** DM4(B)T will hold three steering group meetings (months 4, 8 and 12), where progress against research, technical and non-technical challenges will be reviewed.

National Importance RCUK as a whole has invested a significant amount in energy research, but the three main stakeholders (RCUK, institutions and grantholders) are only just beginning to make a sustainable data legacy part of research practice. At the same time, the data retention policy is coming into force, but without guidance and support the minimal approach of a downloadable zip file may be the common compliant solution. DM4(B)T will produce:

OP1: **Prototype tools:** to demonstrate access to sample (Build)TEDDI datasets.
OP2: **Publications:** to share findings and experience with relevant communities, including data curation, semantic techniques for exploiting digital information, and energy and buildings.
OP3: **Data access workshop:** to share recommendations on data sustainability and to demonstrate the prototype tool.
OP4: **Planning workshop:** to train (Build)TEDDI researchers to write data management plans (DMPonline), to promote policy compliance and to maximise potential future data sharing.
OP5: **Ethics workshop:** to train (Build)TEDDI researchers in how to manage ethical questions raised by the collection, publication and citation of (Build)TEDDI data.
OP6: **Recommendations:** an online report to identify routes to facilitate a sustainable data legacy for (Build)TEDDI projects.

These outputs will provide a blueprint for data sustainability in the engineering and physical sciences, which currently lags behind the biological, earth and social sciences in terms of data preservation and access. DM4(B)T will not only promote an open data legacy across (Build)TEDDI projects, but will inform national and international developments in this field, a focus of activity for bodies such Jisc, the Research Data Alliance, the Australian National Data Service, and the National Institute for Health. Exploring, understanding and solving the challenges will benefit both EPSRC-funded researchers and contribute to the RCUK and UK Government agendas for open (research) data.

References: 1. Teresa Chiang, Gokhan Mevlevioglu, Sukumar Natarajan, Julian Padget, and Ian Walker. Inducing [sub]conscious energy behaviour through visually displayed energy information: A case study in university accommodation. *Energy and Buildings*, 70(0):507 – 515, 2014. 2. Gokhan Mevlevioglu, S Natarajan, and J A Padget. Using semantic annotation in building databases to improve information and energy modelling: a use-case of uk domestic time-series data. 2014. 3. Gokhan Mevleviolgu. Potential benefits of building semantic database and semantic query application for research in uk domestic housing stock. Master's thesis, University of Bath, 2011. 4. Sukumar Natarajan, Julian Padget, and Liam Elliott. Modelling UK domestic energy and carbon emissions: an agent-based approach. *Energy & Buildings*, 43:2602–2612, 2011. http://dx.doi.org/10.1016/j.enbuild.2011.05.013.

JUSTIFICATION OF RESOURCES
DM4(B)T: DATA MANAGEMENT FOR (BUILD)TEDDI(NET) USING SEMANTIC TECHNOLOGIES

1 Staff costs

The project will run over 18 months and Padget (PI), Natarajan and Pink will commit 5%, 5% and 5% of their time, respectively, to allow for adequate engagement over this period. However, the charge in each case is zero because the activities of this project can be deemed part of either ENLITEN (Natarajan, Padget) or job description (Pink: see below). We divide the project into two phases: 12 months of core activity, during which a RA is employed and 6 months of wrap-up, during which conventional dissemination activities, such as (preparation for) publication occurs.

Data Scientist (Pink) The Research Data Scientists' role is to assist researchers with the development of data management plans and general data stewardship, and to develop and deliver advocacy and training to ensure that researchers are able to fully meet their responsibilities and opportunities in terms of research data management. The Research Data Scientist is also expected to seek out opportunities for collaboration and development on research data initiatives.

P/Co-Is Padget and Nataran are working on the ENLITEN project, funded under the BuildTEDDI call. Pink is the University of Bath's Research Data Scientist as described above. Thus, all three permanent staff are already contributing indirectly to the aims of this project and which is why we have given indicative, nominal percentages above and the associated costs are zero.

Investigator	Department	WP	Project input
Natarajan	Bath/Architecture	All – Co-I	user perspective, RA supervision
Padget	Bath/Computer Science	All – PI	overall management, RA supervision
Pink	Bath/Library	All – Co-I	data management policy, RA supervision

RA One PDRA (based in computer science) is needed for consistent engagement across the 3 work packages, each of which brings different perspectives to bear on the problem of data management for TEDDINET. The PDRA will therefore work with the investigators to oversee work in all WPs. The choice of a RA in computer science is appropriate as all the work packages require substantial and broad knowledge of computational techniqes. We request funding to appoint at higher point/grade than normal to give us the flexibility either to appoint an individual with experience for a shorter time or someone with less experience for a longer time. We have chosen to work within the constraint of a funding cap of £50K, in order to permit the programme manager to make a funding recommendation without taking the proposal to panel. The selection of a higher point/grade does reduce the period of the appointment that can be funded (to approx 8 months full time), but we believe that the benefits of experience would outweigh that of a longer appointment, if a suitable candidate were available. Furthermore, even in the case of a more experienced individual, the duration of the appointment could be retained at 12 months by making the contract part-time.

2 Management and travel costs

All the Bath investigators have strong working relationships and will be able to meet proactively as required, at zero cost. We will also hold regular teleconferences with the TEDDINET management team. The RA will primarily use teleconferences to communicate with sample (Build)TEDDI projects, again at zero cost. However, we have costed in a small budget of £1,000 for the RA to travel between Bath and specific (Build)TEDDI projects (5 person trips @ £200/trip = £1,000), to faciliate this work where required.

Total management + travel costs = **£1,000**

3 Dissemination

The essence of the project is dissemination and along with the development of prototype components forms the majority of the activities to be undertaken.

Event locations are yet to be determined, but Bath, Edinburgh, Loughborough and London are the most obvious venues. Given the 3-person project team, we have costed for 4×2 person trips over the project lifetime at £200 each (to allow for overnight in some cases or the high cost of day returns to London), totalling 8 × 200 = £1,600, noting that in several cases, attendance at the workshops for at least one additional member of the project team may be covered by TEDDINET or by ENLITEN.

- The RA will attend the various national events that will be organized as part of the project along with some or all of the project team, as outlined and costed above.
- We expect to write up the results of this work for two international journals. Because of the short time frame of activity in the project, we consider it unlikely that submission can occur during the 12 months of core project activity, not least because the main impact of the work will take place towards the end of that period. This provides one of the motivations for the proposed 18 month duration.

Total dissemination costs = **£1,600**

4 Impact

A number of dissemination activities are detailed in the Pathways to Impact and are costed below:

- Two workshops (project launch (**workshop 1**)) and data legacy (**workshop 4**)) to be run as part of the existing TEDDINET programme and so will not incur additional costs (£0). The project team's attendance at these events will be covered by the travel costs described under dissemination.
- Two workshops (2 × £700 = £1,400) on data management plans (**workshop 2**) and ethics, publication and citation (**workshop 3**). These will be coordinated with TEDDINET, but will at the same time be open to the UK research community, in order to engage with a range of stakeholders.
- Recommendations report at Month 18 published online by the University of Bath (£0)

Total Impact costs = **£1,400**

5 Other

Hosting, room booking costs, or video and teleconferencing costs will be covered by Bath.

Total other costs = **£0**

Dr Dan van der Horst
Director of TEDDINET
School of Geosciences
Drummond Street,
Edinburgh EH8 9XP

Julian Padget,
Department of Computer Science,
University of Bath,
Bath, BA2 7AY

4th December 2014

To whom it may concern,

EPSRC Proposal: Data Management for (Build)TEDDI projects

On behalf of the TEDDINET team I am happy to confirm that we strongly support the above proposal. Having held constructive discussions with Julian and his colleagues as the proposal was drafted, we are confident that this project will be of direct relevance and interest to TEDDI and (Build)TEDDI partner projects. More specifically we look forward to:

1. Benefits arising from the dissemination and sharing of data management principles and practice.
2. Contributing to the development of recommendations concerning the policy for and practice of data management for researchers, institutions and the research councils.
3. Participation by (Build)TEDDI projects in focussed workshops on the various aspects of data management identified in the proposal
4. Enabling access to data sets collected by (Build)TEDDI projects, subject to ethics conditions.
5. Contributing to the definition of the requirements for a data management framework for (Build)TEDDI projects and assisting in its evaluation.
6. Participating in the project steering group that will meet three times during the course of the project.

We fully are committed to support this project in the following ways: (a) intermediating between DM4(B)T and TEDDINET members (b) supporting the organization of the workshops in the context of TEDDINET and (c) contributing, in line with TEDDINET's aims and objectives, to participation costs for TEDDINET members at those workshops, and (d) travel/teleconference to Bath to participate in three steering group meetings.

Yours faithfully,

Dan Van der Horst

Subject: Support for EPSRC Proposal: Data Management for (Build)TEDDI projects
From: "Ben Ryan (EPSRC, CIS)" <Ben.Ryan@epsrc.ac.uk>
Date: 10/12/14 16:04
To: "jap@cs.bath.ac.uk" <jap@cs.bath.ac.uk>
CC: "Ben Ryan (EPSRC, CIS)" <Ben.Ryan@epsrc.ac.uk>, "C.J.Pink@bath.ac.uk" <C.J.Pink@bath.ac.uk>

Dear Julian

As discussed with Cathy Pink, I anticipate that if the above proposal is funded, EPSRC would look forward to:

1. Researchers and other research data users benefitting from a sustainable data legacy beyond the scope of TEDDINET, extracting additional value from the original £22m investment.
2. Enhanced data management skills amongst EPSRC-funded researchers, supporting compliance with the EPSRC Policy Framework on Research Data and informing improved data infrastructure for future EPSRC-funded projects.
3. Participating in a project steering group that will meet three times during the course of the project.

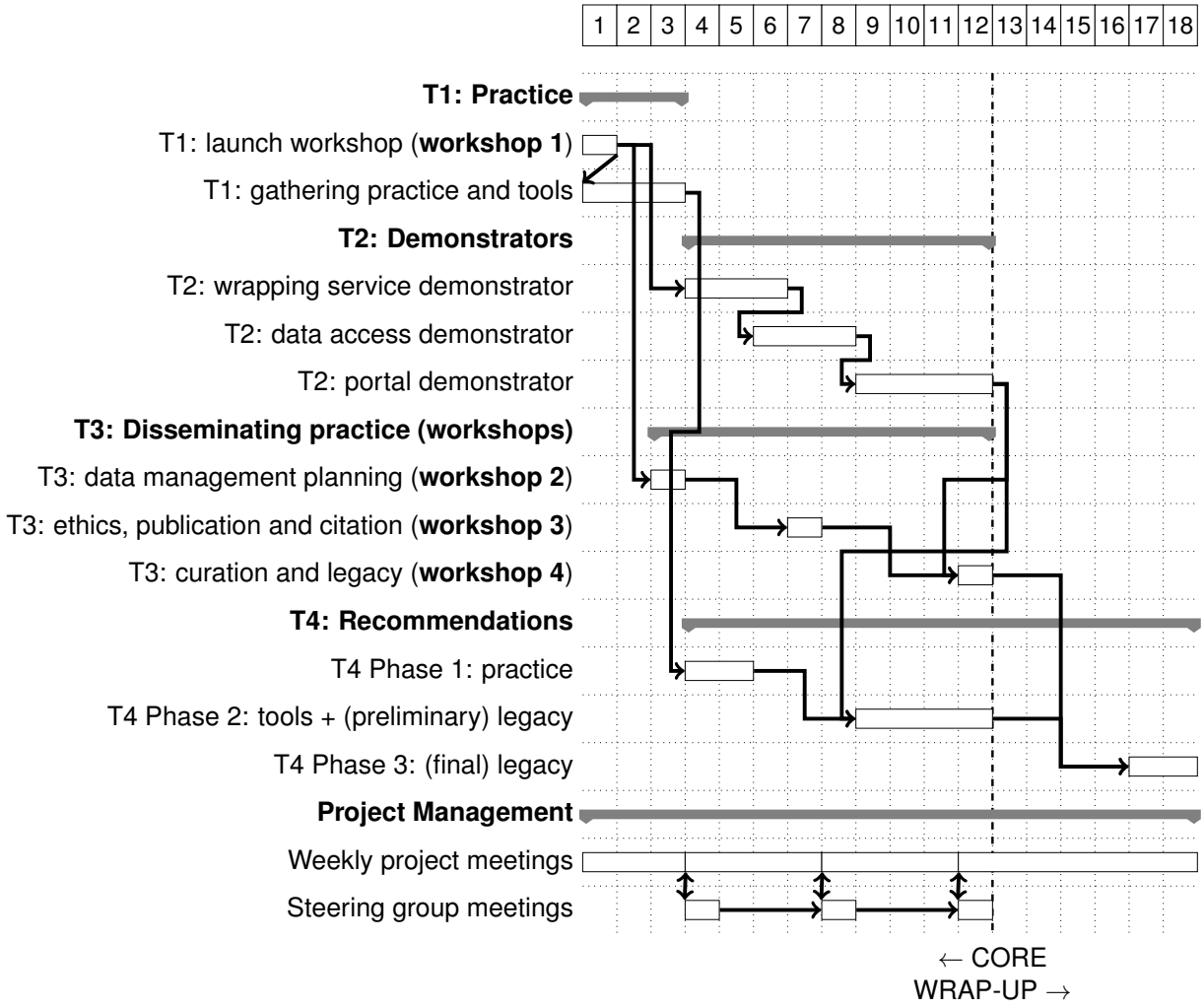
I estimate that the cost to EPSRC in time and resources for the above interactions would involve only travel to Bath to attend three steering group meetings.

Yours faithfully

Ben Ryan

WORK PLAN

PROJECTNAME: DATA MANAGEMENT FOR (BUILD)TEDDI(NET) USING SEMANTIC TECHNOLOGIES



T1: Practice begins with a launch workshop (**workshop 1**), hosted by TEDDINET and co-organized by TEDDINET and DM4(B)T. The purpose is to collect practice, policy, tools and volunteers from TEDDINET and the wider EPSRC community for on-going engagement.

T2: Demonstrators develops three functional components:

(i) to illustrate how to make a dataset into a web resource,

(ii) to show how the metadata about the dataset can be used to query the resource using semantic rather than structural information, and

(iii) to publish the resource through a web portal to demonstrate discovery, access and citation of dataset queries[1]

T3: Disseminating practice comprises three workshops (**workshops 2, 3, 4**), participation in each of which will be supported by TEDDINET. The first two are DM4(B)T specific, on data management planning and ethics, and will be run by DM4(B)T. The last on data legacy will be hosted by TEDDINET and CEE and co-organized by TEDDINET, CEE and DM4(B)T, featuring demonstrations of the tools developed in **T2**.

T4: Recommendations takes what has been found in **T1**, developed in **T2** and learned the workshops in **T3** to deliver two intermediate and one final report.

All of the above will inform, and be informed by, the weekly project meetings and the quarterly steering group meetings.

[1]The portal is only intended as a prototype to show how TEDDINET (and other) datasets could be published by host institutions and then appear to be merged as a collection of resources. It has no pretensions as a production service.