JISC Final Report

Project Information			
Project Identifier	*To be completed by JISC*		
Project Title	Advanced Climate Research Infrastructure for Data (ACRID)		
Project Hashtag			
Start Date	1st August 2010	**End Date**	31st July 2011
Lead Institution	University of East Anglia		
Project Director	Dr Tim Osborn		
Project Manager	Dr Sarah Callaghan		
Contact email	Sarah.callaghan@stfc.ac.uk		
Partner Institutions	University of East Anglia (UEA), Science and Technology Facilities Council (STFC), Met Office		
Project Web URL	http://www.cru.uea.ac.uk/cru/projects/acrid/		
Programme Name	*Managing Research Data*		
Programme Manager	Simon Hodson		

Document Information			
Author(s)	Sarah Callaghan, Arif Shaon, Tim Osborn, Colin Harpham		
Project Role(s)	Project team		
Date	20th Sep 2011	**Filename**	ACRID_finalreport_20Sep2011.doc
URL	http://www.cru.uea.ac.uk/cru/projects/acrid/ACRID_D1.4_finalreport.pdf		
Access	This report is for general dissemination		

Document History		
Version	**Date**	**Comments**
0.1	9th June 2011	First draft
0.2	20th June 2011	Added contents to Section 3.1, 3.2 and 3.3
0.3	30th June 2011	Draft to be submitted to JISC
1.1	17th Aug 2011	Modifications and text added by Tim Osborn
1.2	16th Sep 2011	Minor modifications from Keith Briffa, Arif Shaon, Phil Jones, Sarah Callaghan and Tim Osborn
1.3	20th Sep 2011	Final version

Table of Contents

1 Acknowledgements

The Advanced Climate Research Infrastructure for Data (ACRID) project is funded by JISC as part of their Managing Research Data programme. It is a collaboration between UEA's Climatic Research Unit (CRU), STFC's British Atmospheric Data Centre (BADC) and the Met Office, and was part funded by STFC.

2 Project Summary

Climate research and the climate data that supports its scientific findings has recently come under increased scrutiny. It is even more important to make climate datasets available for re-use and re-examination, while at the same time capturing key information about the dataset provenance, workflows and data descriptions.

The Advanced Climate Research Infrastructure for Data (ACRID) project aimed to implement a linked-data approach for sharing some example climate datasets, and in doing so develop the necessary architecture, infrastructure and tools that might be implemented more widely within the climate science community. The ACRID project demonstrated how datasets developed by the climate science community might be published in ways that facilitate:

- the provenance of the published data to be more clearly recorded (e.g. data sources and versions, software versions, and processing options);
- the recreation of derived data from source data to be more straightforward, even a number of years after publication;
- the citation of data in a way that links more directly to the precise version of data that was used and, by using the linked-data approach, make relationships between different datasets clearly visible.

Deployment of an operational system was considered out of scope of the ACRID project; instead the project results have provided key information on approaches and techniques that other researchers might employ in the future.

This project considered four climate dataset case studies:

1. CRUTEM: This high profile dataset of monthly global gridded land temperatures is generated by processing station observation data, and is described in a series of highly-cited papers (e.g. Brohan et. al. (2006)).
2. CRU TS: The CRU TS dataset includes multiple climate variables interpolated at a relatively high spatial resolution, and updated twice per year. A complex chain of processing is performed on a large number of raw observational datasets.
3. Tree-ring chronologies: Tree-rings are a widely used proxy in paleoclimate reconstruction, but the relevant chronologies developed are highly dependent on an empirical 'standardisation' procedure. The case study used within ACRID are tree-ring data from the Yamal region of northern Russia.
4. HadCET: The Met Office Hadley Centre's 'Central England Temperature' research dataset is the longest instrumental record of temperature in the world.

These case studies formed the key elements of the project. We expect that the approaches and prototypes developed, assessed and implemented in this project will provide useful guidance and exemplars for other areas of climate science.

3 Main Body of Report

3.1 Project Outputs and Outcomes

Output / Outcome Type *(e.g. report, publication, software, knowledge built)*	**Brief Description and URLs (where applicable)**
Description of CRU's scientific workflows (Report)	In-depth analyses of the scientific workflows associated with the climate research datasets being considered within the ACRID project. From these workflows, the requirements for capturing software and dataset metadata have been identified and are also represented here. The information flow patterns have been identified in order to construct a general architecture that is applicable to other HEIs. **Report:** http://www.cru.uea.ac.uk/cru/projects/acrid/ACRID_D2.1_scientificworkflows.pdf
CRU Information Architecture (Report, Model, Infrastructure)	An information architecture that is intended to improve the current approaches to managing the CRU datasets by facilitating greater transparency and traceability of the data life-cycle. Additionally, it should also enable improved and interoperable data accessibility and sharing through adoption of suitable ISO standards and linked-data principles. The information architecture principally consists of two components: • an **Information Model** intended to accurately describe the workflows associated with the CRU datasets, thereby enabling re-enactment of the workflows to verify provenance for the CRU datasets • a **Data Management Infrastructure** mainly for capturing the metadata defined by the information model. **Report:** http://www.cru.uea.ac.uk/cru/projects/acrid/ACRID_D2.2_informationarchitecture.pdf **Information Model (RDF Ontology):** http://www.cru.uea.ac.uk/cru/projects/acrid/ontologies/cw/cru_workflow.owl **Information Model (GML):** http://www.cru.uea.ac.uk/cru/projects/acrid/schemas/cws/cru_workflows.xsd
ACRID Linked-workflow server	A linked-data server (based on the open-source GeoTOD linked-data server) for exposing the CRU workflows as linked-data that can be formally published using the DOI mechanism. **Server URL:** http://westerly.badc.rl.ac.uk:8080/alws/index.html

3.2 How did you go about achieving your outputs / outcomes?

3.2.1 Project Objectives

ACRID aimed to develop an efficient customisable linked-data approach to publishing the workflows (a trail of provenance, e.g. processes applied, interim data generated) associated with the scientific datasets held by the Climatic Research Unit (CRU)[1] at the University of East Anglia. This is intended to facilitate the provision of greater transparency and traceability of data life-cycle, and to enable improved and interoperable data accessibility and sharing.

[1] http://www.cru.uea.ac.uk/

3.2.2 Motivation

Traditionally, the formal scientific output in most fields of natural science has been limited to peer-reviewed academic journal publications. Datasets have been and continue to be archived, but the scientific focus remains on the final output, with less attention often paid to the chain of intermediate data results and their associated metadata, including provenance. In effect, this has constrained the representation and verification of the data provenance to the confines of the related publications. This culture, however, has started to change, owing to initiatives such as the OJIMS[2] and CLADDIER[3] projects, which have developed mechanisms for formally publishing scientific datasets as scientific resources in their own right, rather than merely as an adjunct to the publication of scientific articles.

Publishing a dataset by itself, however, will not provide a complete account of its provenance. In the typical production of a dataset, there is a series of processes and operations applied, analyses conducted, and interim data results generated, i.e. a complex **scientific workflow** enacted before a scientific experiment or observation yields its final data output. These processes and interim data outputs, along with other related metadata, form a dataset's lineage. This is increasingly important for open-access data to determine their authenticity and quality, especially considering the growing volumes of datasets appearing in the public domain. A detailed history of the data will also help the users determine if the data is fit for its intended purpose(s).

The need for the publication of data provenance was also highlighted in the 2009 UK House of Commons Science and Technology Committee report into the release of private emails at the Climatic Research Unit (CRU) of the University of East Anglia which noted that although CRU's "(data sharing) actions were in line with common practice in the climate science community", went on to suggest "...that climate scientists should take steps to make available all the data that support their work (including raw data) and full methodological workings (including the computer codes)". The report also noted that "it is not standard practice in climate science to publish the raw data and the computer code in academic papers". The ACRID project was motivated by the recommendations arising from this report.

3.2.3 Methodology

3.2.3.1 Analysis of the CRU Datasets

We conducted in-depth analyses of the scientific workflows associated with the climate research datasets being considered within the ACRID project in order to identify and understand the various types of metadata associated with these workflows. The information flow patterns in the workflows were also identified in order to construct a general architecture that may be applicable to other HEIs.

In general, the metadata associated with the workflows assessed can be categorised as follows:

3.2.3.1.1 Observation

The act of measuring or calculating a particular property (e.g. temperature) associated with a certain feature of interest (e.g. air) over a discrete period of time is referred to as an Observation within the geospatial community. The CRU datasets are essentially the outcomes of such observations that primarily fall under two categories: **raw or source observations** undertaken at various land-based climate monitoring stations or sites around the world, and **computed or constructed observations** (e.g. CRU TS dataset[4]) that are derived from the source observations and typically published and/or used as the basis for publications. Also of note here is that the general structure of the CRU datasets are typically time-series[5] with varying structures.

2 http://proj.badc.rl.ac.uk/ojims

3 http://claddier.badc.ac.uk/trac

4 http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__dataent_1256223773328276

5 A series of values measured at different points of time as the result of an observation.

3.2.3.1.2 Process

A process is essentially an action or a set of actions performed to produce the result (i.e. dataset) of an observation. In practice, a process may be an algorithm, a computation, a manual procedure, or calculation that may also consist of a sequence of steps, where the outputs of one step may be used as the inputs of another succeeding step.

3.2.3.1.3 Processor

This is an entity or a set of entities that performs and/or controls a process in order to produce the result of an observation. In practice, a processor may be a human, computer software or any type of hardware, such as weather observation instrument.

3.2.3.2 The ACRID Workflow Information Model

Following the workflow analyses, we reviewed a number of existing information models with a view to identifying a suitable model for describing the CRU workflows. Of particular note among these models are: ***Open Provenance Model (OPM)*** [2], ***ISO 19156 Observations and Measurements (O&M)*** [1] ***model*** and ***Climate Science Modelling Language (CSML)***[6]. The review was conducted in consultation with a number of domain experts from the British Atmospheric Data Centre (BADC)[7] and the UK Met Office to ensure accurate interpretation of the information and concepts assessed.

The review indicated that both the ISO O&M Model and CSML could be directly applicable to the CRU observational datasets as they are specifically designed for describing environmental observations, such as the ones represented by the CRU datasets, and are commonly used in the geospatial community. CSML is effectively an application schema of the ISO O&M model specialised for representing time-series datasets (such as the CRU datasets), and also has a growing user community led by the BADC in terms of developing and providing tools and software support for understanding and manipulating datasets encoded in CSML.

On the contrary, the OPM, though conceptually applicable to the CRU datasets, was deemed too generic and uncommon within the geospatial community to be effectively applied to the CRU datasets.

6 http://csml.badc.rl.ac.uk/

7 http://badc.nerc.ac.uk/home/index.html

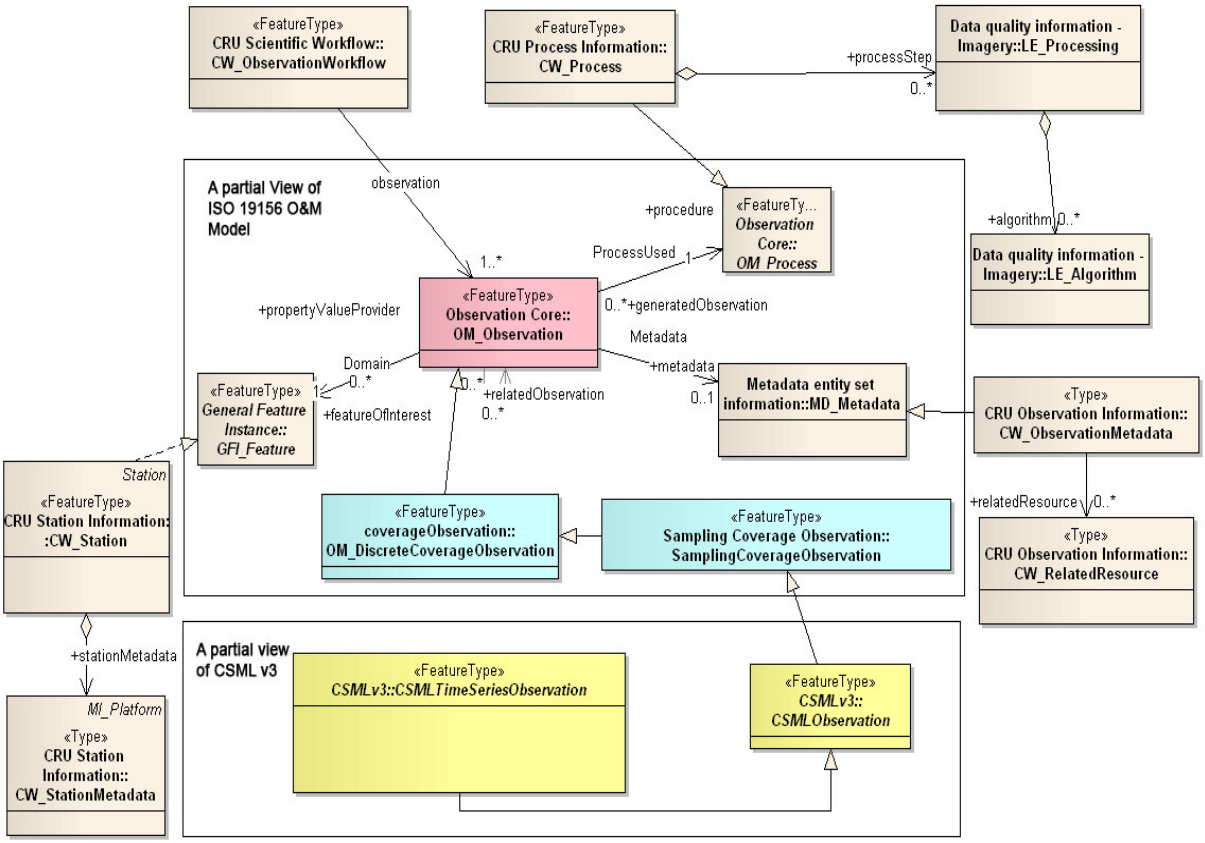

Figure 1: The ACRID Workflow Information Model

Therefore, we developed the CRU information model as an application schema of the ISO O&M Model with the observation related concepts derived from the *CSMLTimeSeriesObservation* classes (Figure 1). The model was mainly developed in UML with the underlying concepts additionally represented in RDF for facilitating linked-data representations of the CRU workflows, and GML for enabling compatibility with the CSML and other related tools. A complete description of the ACRID information model is provided in [4].

3.2.3.3 Publishing Linked Workflows using OAI-ORE and DOI

To publish the workflows described by the workflow model outlined above as linked-data, we have developed an RDF/OWL[8] ontology representation of the model. This has also involved creating unofficial ontology representations of the ISO O&M model and CSML as well as a number of other related ISO models (e.g. ISO 19115-2:2009) because no formal ontologies for these models currently exist.

Dissemination of the linked-data instances of the workflows is done using the OAI-ORE[9] technology. The OAI-ORE defines standards for the description and exchange of aggregations of Web-based resources in a linked-data compliant way. The key OAI-ORE concepts are:

- **Aggregation (A)**: a set of web-based resources.
- **Aggregated Resource (AR)**: a web-based resource that constitutes (by itself or together with other resources) an Aggregation. Examples include a workflow instance and a related publication.
- **Resource Map (ReM):** a brief description of an Aggregation.

[8] "RDF is a standard model for data interchange on the Web" that "extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a "triple")" - http://www.w3.org/RDF/. OWL, developed as a vocabulary extension of RDF, is a semantic markup language for publishing and sharing ontologies on the World Wide Web - http://www.w3.org/TR/owl-ref/.

[9] Open Archives Initiative Object Reuse and Exchange (OAI-ORE) - http://www.openarchives.org/ore/

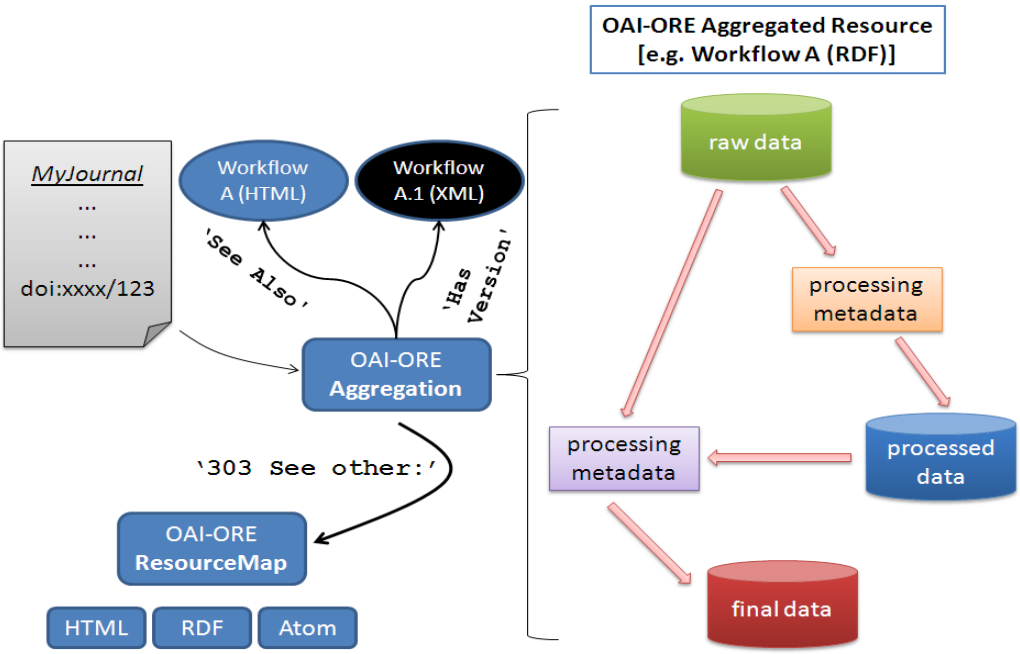

Figure 2: An OAI-ORE representation of linked workflows

So, as illustrated in Figure 2, a CRU workflow instance described by the workflow model would be encapsulated within an OAI-ORE Aggregation as an Aggregated Resource.

In order to publish the workflow instance, we assign a DOI to the corresponding OAI-ORE Aggregation (identified by an OAI-ORE Aggregation URI). So, when the DOI is de-referenced, the following sequence of events may occur:

- The client is redirected (using HTTP 303 re-direct as recommended by the linked-data principles) from the Aggregation URI to the URI of the Resource Map that describes the Aggregation.
- The Resource Map serves as a *landing or splash page* providing a description of the Aggregation (*not Aggregated Resource*), which includes the URI for the Aggregated Resource (e.g. a workflow instance). The client is then able to de-reference the URI for the Aggregated Resource to retrieve it. ***It is important that the contents and format of the Aggregated Resource remain static for an indefinite period of time in order to adhere to the DOI rules***.

The Aggregation description contained within a Resource Map may also include information about other static or non-static resources related to the Aggregated Resource. For example, the link to a newer version of the workflow instance may be provided in the Aggregation using an appropriate vocabulary (e.g. RDFS 'seeAlso' – Figure 2). In effect, this enables the provider of a workflow instance to be able to seamlessly link to other related resources that he or she may not have control over – one of the principle advantages of linked-data.

In addition, a Resource Map may be provided in multiple formats (e.g. HTML, RDF, atom – Figure 2) based on the client's request. So, if an Aggregation URI is de-referenced in an RDF browser, the client should expect an RDF representation of the corresponding Resource Map. If the same URI is de-referenced in an HTML browser, then the same Resource Map should be provided in HTML and so on. However, as mentioned before, it is vital that the actual Aggregated Resource to which a DOI corresponds remains static in terms of both contents and format. Additional representations of the Aggregated Resource may be made available to the users through its Aggregation description using an appropriate vocabulary (e.g. Dublin Core 'hasVersion' – Figure 2).

3.2.3.4 Managing 'Live' and Published Workflows

As highlighted before, a published workflow should remain static in terms of both contents and integrity for an indefinite period of time. Therefore, as aptly identified in [7], the published workflows

should be managed separately from the 'Live' workflows, which typically represent volatile datasets. For example, the HadCET dataset[10] held by the UK Met Office is updated daily – the workflow associated with this would be a 'Live' workflow. The management of these two types of workflows could be conducted in either logically or physically separate environments. In ACRID, we have adopted the latter approach to avoid inadvertent changes to the published workflows, and thereby facilitating more effective management of both types of workflows.

3.2.4 Validation and Prototype

We have tested our linked-data approach using three distinct datasets published by CRU: (i) CRUTEM land-surface air temperature data (specifically version CRUTEM3); (ii) CRU TS land-surface high-resolution data for multiple variables (specifically version CRU TS 3.1); and (iii) a tree-ring chronology from the Yamal region of northern Siberia[11]. In addition, we have also applied the ACRID linked-data approach to the Hadley Centre's Central England Temperature dataset (HadCET) published by the UK Met Office.

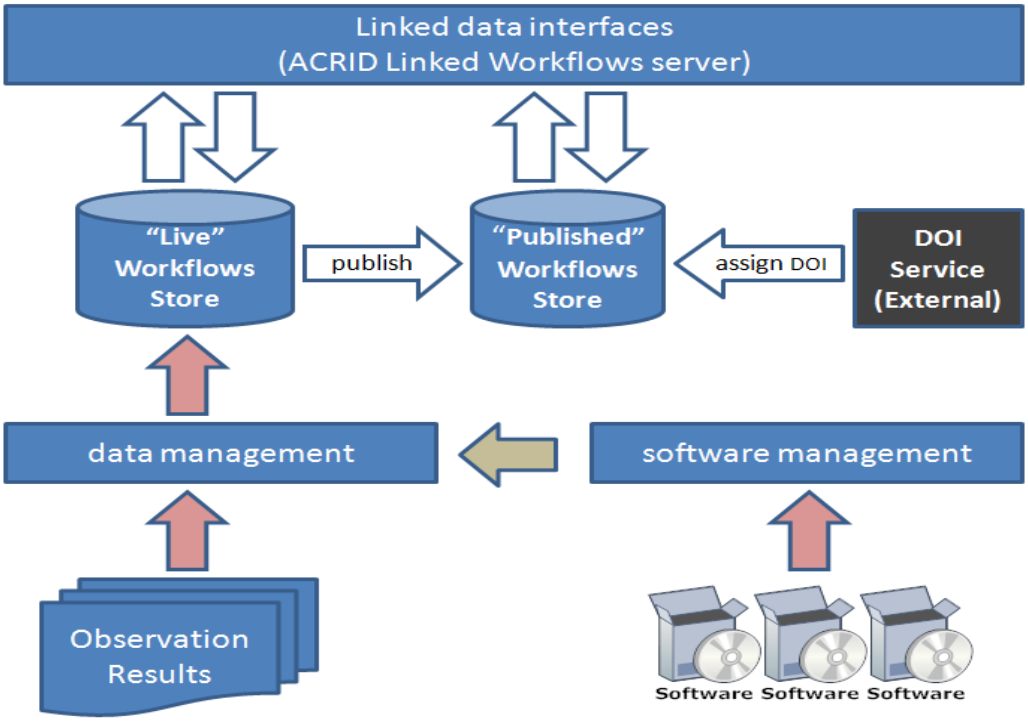

Figure 3: CRU Data Management and Publishing Infrastructure

To this end, we first designed a data management infrastructure (Figure 3) for CRU to accurately and efficiently capture and manage provenance-related information (as defined by the workflow model) about the workflows associated with the three aforementioned datasets [4]. The information captured is then stored and exposed as linked-data in accordance with the approach described in (3.2.3.4) through a linked-data server, namely the ACRID Linked Workflows Server [5]. Two separate data stores (based on the PostgreSQL relational database – Figure 3) are used to store and manage the published and "live" workflows to ensure the integrity of the published workflows and effective management of different versions of the "work in progress" workflows respectively [6].

We have also developed an infrastructure to enable citation of the "published" workflows within the context of scholarly communication. This involves formally publishing the OAI-ORE aggregation of a workflow in the "Published" workflows store, using the Digital Object Identifier (DOI) technique (Figure 2). A key aspect of this citation infrastructure is a "data publishing" function incorporated within the

[10] Met Office Hadley Centre HadCET observations dataset - http://www.metoffice.gov.uk/hadobs/hadcet/graphs/index.html

[11] CRU Yamal tree-ring data - http://www.cru.uea.ac.uk/cru/people/briffa/yamal2009/data/

ACRID Linked Workflows Server that is accessible through a secure, user-friendly and intuitive web interface. This enables taking a snapshot of a workflow to be published from the "Live" workflows store and storing it in the "Published" workflows store (Figure 3) in order to preserve the integrity of both the contents and the format of a published workflow. In addition, unique URIs are assigned to the published workflows in order to distinctly identify a workflow and the format in which it has been published.

Notably, the ACRID Linked Workflows Server is based on GeoTOD[12] - an open-source linked-data infrastructure that implements the draft UK Cabinet Office guidelines [10] for exposing geospatial data as linked-data. These draft guidelines for geospatial data extend more general guidelines for publishing UK public sector data (under data.gov.uk), and have been proposed by the UK Government in specific recognition of the importance of geospatial data, and also recognising parallel work at the European level on deploying the INSPIRE [8] SDI (which currently uses web services, but not linked-data principles). We therefore envisage that the adoption of GeoTOD for publishing CRU's datasets would have the future potential for sharing these datasets through the INSPIRE SDI (should it adopt linked-data approaches to data sharing).

3.2.5 Dissemination

Work on the project has been presented at a number of conferences and workshops:

- Arif Shaon (Presentation): *Advanced Climate Research Infrastructure for Data*, The Met Office, UK 1 September 2011
- Arif Shaon (Presentation): *Advanced Climate Research Infrastructure for Data*, Linked-data Workshop, The University of Oxford, 2 August 2011
- Arif Shaon (Presentation): *Advanced Climate Research Infrastructure for Data*, JISC Managing Research Data (International) programme workshop, 28-29 Mar 2011.
- Arif Shaon, Sarah Callaghan, Bryan Lawrence, Brian Matthews, Andrew Woolf, Tim Osborn and Colin Harpham (Paper): *A Linked Data Approach to Publishing Complex Scientific Workflows*, Accepted for publication and presentation at the IEEE e-Science Conference 2011
- Arif Shaon, Sarah Callaghan, Bryan Lawrence, Brian Matthews, Andrew Woolf, Tim Osborn and Colin Harpham (Paper): *Opening up Climate Research: a linked data approach to publishing data provenance*, Submitted to the 7th International Digital Curation Conference 2011

3.3 What did you learn?

We encountered the following issues with the existing standards and techniques adopted in ACRID:

3.3.1 Linking vs Exchanging

The linked-data principles [9] offer an excellent means of seamlessly linking geospatial workflows to their corresponding publications as well as other related resources. However, the ability to link resources may not necessarily translate into the ability to effectively exchange and share these resources, unless the linking and exchange formats are either the same or equally common within the associated community. The Resource Description Framework (RDF)[13], the principal linked-data format, though gaining increased adoption, is not a commonly used format for exchanging data within the geospatial community. Instead it predominantly relies on the Geography Markup Language (GML)[14] representations of the ISO 19100 series models along with other geographical data formats, such as netCDF, for encoding and exchanging environmental data. For example, GML is the official data exchange format for the INSPIRE community.

Therefore, a greater awareness by the research community itself of data publishing motivations and technologies will be required before the benefits can fully be realised of an approach like ours (which enables related, but unconnected, data resources to be linked). Until that is achieved, a linked-data

12 Geospatial Transformation with OGSA-DAI (GeoTOD-II) on SourceForge - http://geotod.sourceforge.net/about.html

13 RDF-Semantic Web Standard - http://www.w3.org/RDF/

14 Geography Markup Language http://www.opengeospatial.org/standards/gml

approach to describing and publishing geospatial workflows should support commonly used data exchange formats, such as GML, in addition to RDF.

3.3.2 RDF/OWL ontology representations of the ISO 1900 series standards

The ACRID information model is a specialisation of the ISO O&M model and CSML, none of which currently has any official RDF ontology representations. Therefore, it was necessary to develop unofficial ontologies based on the ISO O&M, CSML and other related ISO 1900 series standards to enable a comprehensive ontology representation of the ACRID information model. However, ontology representation of some of the complex geospatial aspects defined in the GML encoding specification (ISO 19139 [3]) was deemed outside the scope of ACRID as it would have required deeper collaboration with the wider GML user community to ensure correct interpretation and representation of these aspects. Consequently, it would have likely exceeded the project's time scope. Hence, it was decided to use GML representations of these aspects of the CRU datasets as embedded XML literals in the corresponding ACRID ontology instance. However, this approach, though effective, is by no means ideal. With interest in linked-data approaches rising within the geospatial community, this identifies a significant need for official ontology representations of the ISO 1900 series standards that are used to describe and share the geospatial datasets.

3.3.3 Spatial-temporal Coordinate Reference System (CRS)

Currently there is a need for suitable spatial-temporal CRS definitions for representing data grids larger than 2-D. Commonly used work-around solutions involve the use of custom 2-D spatial CRS for representing 3-D or larger grids. ACRID had to adopt such a work-around for the CRU 3-D gridded datasets, such as CRUTEM, by modifying the EPSG:6:6:4326 CRS to list the time values of the "time" axis of the grid being described as the number of days since a particular year[15]. Exploring this issue in more detail with a view to developing a more efficient solution would have been ideal but was outside the remit of ACRID.

3.4 Immediate Impact

The ACRID project has stimulated and supported a number of improvements to the management of research data within the Climatic Research Unit at the lead institution of this project. For example, version control software is now being used (for data, software and documents) more widely within CRU than previously and ACRID has supported the transition from the older, less capable system (RCS) to a more modern and flexible system (Subversion). Beyond CRU, the Research Computing Service of the lead institution (UEA) has now implemented an institution-wide UEA Subversion and Trac[16] service, prompted in part by the needs of the ACRID project. This will facilitate wider uptake of version control for managing research data and associated software/documents. The ACRID project has also supported improvements in the internal recording and managing of the metadata and workflows associated with some climate datasets within CRU. Although sufficient information to allow derived datasets to be reproduced was already held, some aspects have now been collated and/or restructured to support more efficient management and to facilitate easier replication. These changes, together with the information available via the deliverable reports and the linked-data server, should also benefit the wider community by providing more information about source data and statistical analysis methods (i.e. workflow) that underpins widely used climate datasets.

3.5 Future Impact

Regardless of the questions/issues above (Section 3.3), the use of the techniques presented in this paper should significantly help in the scientific process itself – CRU is not the only organisation with complex workflows migrating "raw" data to "published" data. It is not uncommon for researchers to fail to record key details in this process, necessitating the expensive and time-consuming re-construction of thoughts and processes to reproduce pre-existing results.

[15] See also the Climate and Forecast (CF) metadata convention: http://cf-pcmdi.llnl.gov/documents/cf-conventions/1.6/ch04s04.html

[16] http://en.wikipedia.org/wiki/Trac

The methodology presented here should be deployable elsewhere within the climate and other environmental sciences, and (with suitable adaptation to the data model used) could also be applied to publish data in wider areas of science. For example, while the O&M model has been designed for geospatial observations, the underlying concepts have the potential for application across wider domains of the science. This should be investigated in future work.

In addition, it would also be possible to develop suitable mechanisms for mapping the Workflow Model presented above (Section 3.2.3.2) on to the workflow description languages used by some of the widely used workflow execution engines, such as Taverna. This should effectively enable (semi-)automated re-enactment, and thus, validation of the workflows described by the workflow model.

Further, the use of linked-data techniques coupled with content negotiation will also be of significant benefit in ensuring that the information can be consumed by a variety of clients, not just by browsers displaying HTML. To that end, the lessons learned here will be explored further in the context of the wider roll-out of DOIs linking citation descriptions to data in the data centres funded by the UK's Natural Environment Research Council (NERC).

We also envisage that our approach will become increasingly important as the semantic web and linked-data compete with existing Spatial Data Infrastructures (SDIs) like INSPIRE as web platforms for publishing geo-scientific data. With growing political sensitivity over the need for openness in research data, technical approaches like ours are being sought that support alignment with national government transparency agendas.

4 Conclusions

The project success criteria were if the project:

- developed an information model capturing key climate data workflow metadata
  - Completed and reported in deliverables D2.2 and D4.1.
- developed and implemented reference climate data and software management tooling
  - Completed and reported in deliverables D4.2 and D4.3.
- deployed linked-data prototypes for one or more key climate research datasets
  - Completed and reported in deliverables D5.1, D5.2, D5.3 and D5.4.
- enabled the recreation of published versions of processed climate datasets from source
  - This has been attempted and successfully achieved for several instances of published climate data (CRUTEM3, Yamal tree-ring ring chronologies), though further population of the database underlying the ACRID linked-data server database is required to enable this to be done via the linked-workflow system.

5 Recommendations

- At present, there is a need for a suitable spatial-temporal CRS for representing data grids larger than 2-D . Therefore, efforts should be dedicated to developing and standardising a suitable spatial-temporal CRS for 3-D or larger data grids.
- With interest in linked-data approaches rising within the geospatial community, there is a significant need for official ontology representations of the ISO 1900 series standards that are used to describe and share the geospatial datasets. The domain experts including the related standardisation bodies, such as the Open Geospatial Consortium (OGC), should seek to address this need.

6 Implications for the future

There has been an increasing focus recently on the reliability of climate datasets and consequently on the confidence which can be placed in our interpretation of those datasets and in the implications for understanding climate change. Reliability and confidence arise from a range of factors, including the provision of source data and a transparent description of the scientific workflow that enables derived climate datasets to be reproduced. Reproducibility (and in some cases repeatability) is a key element of science and is an area of active investigation – not only in climate science but also in other scientific fields. The broader climate science community are addressing these issues, including appropriate ways to publish more detailed scientific workflows and raw data, and more transparent links between related datasets and between scientific findings and the data that underpin them. The various information technologies used to construct the ACRID prototypes, and the trialling of the case studies, are contributions to this broader and ongoing activity of the user community that should help

identify the relative merits of different approaches to managing and publishing climate data and associated workflows.

We will finish populating the URLs to which the linked data points. We also plan to "publish" with a DOI the CRUTEM3 version that we've been working with and will do the same for the CRUTEM4 dataset when it is ready for release. However, some consultation with the user community needs to occur before this is done because it might do a disservice to those users who want to work with these datasets in their traditional format if the "official" DOI for CRUTEM3 (or CRUTEM4) pointed only to the linked-data representations. The linked-data representation will be unfamiliar to many users and may only provide benefits (over the form in which these data have previously been made available) to some of them, while most may prefer to go direct to the representation that they are familiar with. However, there is no reason why those pre-existing representations may not be cited with a (different) DOI in their own right, or the representations could be recorded as related information to avoid confusion.

The long-term contact for the linked-data server is Arif Shaon, while Tim Osborn is the long-term contact for the scientific workflows and data. Contact details are available at the ACRID project website: http://www.cru.uea.ac.uk/cru/projects/acrid/people.htm

7 References

[1] ISO 19156:2010 - Geographic information — Observations and measurements

[2] Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E. and Van den Bussche, J. (2010) The Open Provenance Model core specification (v1.1). *Future Generation Computer Systems*, last accessed 26 February 11 - http://eprints.ecs.soton.ac.uk/id/eprint/21449

[3] ISO 19107:2003 Geographic information -- Metadata -- XML schema implementation

[4] Report II: Work Package 2.2 - Information Architecture

[5] Report IV: Work Package 4.2 & 4.3 – Linked-data Server for exposing Climate Research Data

[6] Report V: Work Package 4.4, 5.1, 5.2, 5.3 & 5.4 – Data Citation Infrastructure & Example Datasets

[7] S. Bechhofer, J. Ainsworth, J. Bhagat, I. Buchan, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop, M. Gamble, C. Goble, D. Michaelides, P. Missier, S. Owen, D. Newman, S. De Roure, and S. Sufi "Why Linked Data is Not Enough for Scientists", in press, e-Science, December 2010, Brisbane, Australia. Originating URL: http://eprints.ecs.soton.ac.uk/21587/5/research-objects-final.pdf Last accessed: 1-Jul-2011

[8] European Commission, INSPIRE web site http://inspire.jrc.ec.europa.eu/

[9] T. Berners-Lee "Linked data – Design Issues", W3C Document, 2007. Originating URL: http://www.w3.org/DesignIssues/LinkedData.html Last accessed: 1-Jul-2011

8 Appendices

No appendixes have been included because the technical information arising from this project has already been included in the individual project deliverables. These are available from the project website: http://www.cru.uea.ac.uk/cru/projects/acrid/