A linked data approach to publishing complex scientific workflows

Conference or Workshop Item

Accepted Version

A Linked Data Approach to Publishing Complex Scientific Workflows

Arif Shaon[1], Sarah Callaghan[2], Bryan Lawrence[2], Brian Matthews[1]
[1]e-Science Centre
[2]British Atmospheric Data Centre (BADC)
Science and Technology Facilities Council
Didcot, United Kingdom
{arif.shaon, sarah.callaghan, bryan.lawrence, brian.mathews}@stfc.ac.uk

Andrew Woolf
The Bureau of Meteorology
Canberra, Australia
A.Woolf@bom.gov.au

Timothy Osborn, Colin Harpham
Climatic Research Unit
The University of East Anglia, United Kingdom
{T.Osborn, C.Harpham}@uea.ac.uk

Abstract—**Past data management practices in many fields of natural science, including climate research, have focused primarily on the final research output – the research publication – with less attention paid to the chain of intermediate data results and their associated metadata, including provenance. Data were often regarded merely as an adjunct to the publication, rather than a scientific resource in their own right. In this paper, we attempt to address the issues of capturing and publishing detailed workflows associated with the climate/research datasets held by the Climatic Research Unit (CRU) at the University of East Anglia. To this end, we present a customisable approach to exposing climate research workflows for the effective re-use of the associated data, through the adoption of linked-data principles, existing widely adopted citation techniques (Digital Object Identifier) and data exchange mechanisms (Open Archives Initiative Object Reuse and Exchange).**

Linked-data; Scientific workflow; ISO 19156; Provenance; OAI-ORE, CSML, Climate Research, Geospatial

I. INTRODUCTION AND RATIONALE

The formal scientific output in most fields of natural science has been limited to peer reviewed publications. Datasets have been archived, and continue to be archived, but most communities have concentrated on the final output, with less attention paid to the chain of intermediate data results and their associated metadata (including provenance) – the workflow associated with the data. Even where archived, data were often regarded merely as an adjunct to the publication, rather than a scientific resource in their own right.

In this paper, we take the climate/research datasets held by the Climatic Research Unit (CRU) at the University of East Anglia[1], as exemplars to address the issues of capturing and publishing scientific data and the associated workflows for re-use. We present a customisable approach (developed by the ACRID project[2]) to exposing climate research data for re-use, through the adoption of linked-data principles for the data themselves.

In essence, the approach presented here combines the Digital Object Identifier (DOI)[3] - a widely adopted citation technique - with existing widely adopted climate science data models (e.g. ISO 19156 Observations & Measurements model [7] and CSML[4]). This is integrated with linked-data compliant data re-use standards (e.g. OAI-ORE[5]) to enable a seamless link between a publication and the detailed workflow associated with the corresponding datasets.

II. THE MAIN CHALLENEGES

The task of publishing complex scientific workflows needs to address a number of challenges, as identified in [4]. In particular, for publishing workflows associated with geospatial/environmental datasets (the premise of the work presented here), these challenges are the following:

A. Repeatability and Reusability

The main purpose of publishing a scientific dataset is often to support publications written based on that dataset. However, the dataset by itself may not always be sufficient for verifying or validating the related claims/statements made in the corresponding publications. Detailed information about the processes used and the interim results generated, if applicable, is also needed. In other words, published scientific workflows should contain sufficient information in order to facilitate their accurate re-enactment

[1] Climatic Research Unit (CRU), The University of East Anglia - http://www.cru.uea.ac.uk/

[2] Advanced Climate Research Infrastructure for Data (ACRID) - http://www.cru.uea.ac.uk/cru/projects/acrid/
[3] The Digital Object Identifier (DOI) System - http://www.doi.org/
[4] Climate Science Modelling Language (CSML) - http://csml.badc.rl.ac.uk/
[5] Open Archives Initiative Object Reuse and Exchange (OAI-ORE) - http://www.openarchives.org/ore/

The work presented in this paper has been funded by the JISC Managing Research Data (JISCMRD) programme - http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx

or repetition to help verify the evidential basis of the claims in the publications.

The motivation behind publishing scientific workflows is not only limited to verification of provenance; it is also a common practice for the components of a workflow to be re-used in other related workflows. For example, a process for measuring air temperature (e.g. holding a thermometer in the air for a certain time at a certain height) could be applied to measuring air temperatures of two different locations for two different environmental observations. In both cases, the basic function of the process would remain unchanged– what could be changed is the related parameter instance(s), e.g. height at which the temperature is measured. Therefore, if possible, a publishable workflow should contain sufficient information about its constituents to facilitate their re-usability.

B. *Common Workflow Model*

To ensure greater re-usability, a publishable workflow needs to be described using an information model that is understood by the wider user community. Driven by the INSPIRE Directive [6] in Europe, the ISO 19100 series information models and standards (e.g. ISO 19156 O&M model [7]) are increasingly being adopted within the geospatial community for describing geospatial operations and the datasets that result from them. From this perspective, a geospatial workflow model developed based on these ISO standards (as appropriate) would have the potential to be more widely applicable and shareable than any bespoke model for that workflow.

C. *Linking vs Exchanging*

The linked-data principles [5] 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)[7], the recommended 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)[8] representations of the ISO 19100 series models along with other geographical data formats, such as NetCDF for encoding and exchanging environmental data. This analogy also applies to the workflow description formats used by various popular workflow engines/tools, such as Taverna[9]. While these tools are very useful for (semi-) automatically re-enacting workflows (e.g. to verify provenance, confirm repeatability), the formats used for describing the workflows have yet to garner major uptake within the geospatial community. So, a linked-data approach to describing and publishing geospatial workflows should support commonly used data exchange formats, such as GML, in addition to RDF.

III. Existing Related Data Publication Approaches

A key motivator driving the citation and publication of environmental data sets is the requirement that the creators of those datasets receive academic credit for the considerable work they put in to creating or collecting the data, and ensuring they are in an appropriate format, have complete metadata, and are stored in a data repository where they will be archived and curated properly.

Another motivator is providing a process for the validation of scientific datasets through peer-review. For the scientific work presented in academic journal articles, the peer-review process ensures the quality of the work reported in the article, while the publication process produces an article which is fixed and citable, and provides its author(s) with academic credit. An analogous process for data publication would provide benefits to the wider scientific community, allowing for ease of discovery and re-use of the data, while also allowing the conclusions drawn from a given dataset to be independently verified.

In traditional academic publishing, the object to be cited (i.e. the article) is written and peer-reviewed before it becomes citable. In the case of data publication, it makes more sense to allow citation of the dataset before full peer-review, as, by citing a dataset, the host repository is confirming that the dataset is complete and frozen. If dataset citation was to occur after peer-review, there is no guarantee that the dataset would still contain the same data as it did when it was reviewed. Citation before publication also gives some credit to the dataset authors as soon as the dataset is complete, without it ever having to go through peer-review. If scientific peer-review of a dataset is considered the “gold standard”, a method of citation which carries with it connotations of authority and permanence such as DOI therefore becomes a “silver standard”, confirming that the dataset is complete, unlikely to change, in an appropriate format and has sufficient metadata (at least as far as the host repository is concerned).

The UK’s Natural Environment Research Council (NERC) funds six data centres which between them have responsibility for the long-term management of NERC's environmental data holdings. The NERC Science Information Strategy (SIS)[10] has been created to provide the framework for NERC to work more closely and effectively with its scientific communities in delivering data and information management services.

The NERC SIS data citation and publication project is a cross data centre project with the primary aim of implementing the publication and citation of datasets held within the NERC data centres. It builds on previous work funded by NERC and JISC which investigated and

[6] European Commission, INSPIRE web site http://inspire.jrc.ec.europa.eu/
[7] RDF-Semantic Web Standard - http://www.w3.org/RDF/
[8] Geography Markup Language http://www.opengeospatial.org/standards/gml
[9] Taverna - http://www.taverna.org.uk/
[10] NERC SIS: http://www.nerc.ac.uk/research/sites/data/sis.asp

developed methods to form human-readable citation strings (the CLADDIER project [3]) and to demonstrate the mechanics for an overlay data journal (the OJIMS project [1][2]). The project team is working in collaboration with the British Library and DataCite[11] to use DOIs to identify and cite datasets.

The approaches outlined in the OJIMS and CLADDIER projects are very general. CLADDIER, as well as proposing a structure for human readable citation strings, investigated differing methods for publishing datasets, and discussed the requirements for the peer-review of data. OJIMS took the case of an overlay journal for data publication and created a demonstrator journal, investigating the business case for operating it on a long-term basis, as well as surveying the proposed user community about their opinions on data publication and their use of data repositories. The ACRID project extends the work done in both of these previous projects, and takes it down to a more detailed level, focusing on the very specific cases of key CRU datasets and workflows. Different datasets will require different methods to enable them to meet the citation and publication requirements outlined in CLADDIER and OJIMS. The ACRID project goes into these details, using linked data to collect the data, metadata and workflows required to publish the CRU datasets.

Due to the limited scope of this paper, a broader review of the existing data publication approaches has not been provided here. An extensive assessment of these approaches can be found in [3].

IV. ACRID METHODOLOGY

A. Analysis of the CRU Datasets

An analysis of the scientific workflows associated with a number of CRU datasets indicates that these workflows typically consist of a chain of intermediate data results and their associated metadata including the processes used (i.e. provenance) to generate the results [6]. These workflow constituents can be generalised into the following concepts:

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[12]) 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[13] with varying structures.

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) 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.

B. A Workflow Information Model for Geospatial Datasets

In view of the above analysis of the CRU dataset workflows, we have developed a generic information model to enable detailed and accurate description of such workflows, particularly in terms of the three main concepts mentioned above.

Development of the information model involved a comparative review of the following three widely adopted models:

1) Open Provenance Model (OPM)

OPM [11] is a generic model intended to enable digital representation of provenance for any object, whether it is digital or physical. A comparison of the OPM concepts (Fig. 1) with the main concepts (IV, Section A) of the workflows associated with the CRU datasets indicates a close parallel between these concepts. Conceptually, the OPM Artefact (A), Process (P) and Agent (Ag) concepts are analogous to the CRU Dataset, Process and Processor concepts respectively. However, the OPM concepts are too generic and uncommon within the geospatial community to be effectively applicable to geospatial datasets like the CRU datasets.

2) ISO 19156 Observations and Measurements (O&M) Model

The ISO 19156 O&M model [7] defines a conceptual schema for describing environmental observations and the features involved in the sampling associated with such observations. This conceptual schema could also be used to exchange information describing observation acts and their results between communities. In contrast with the OPM, the ISO O&M Model is specifically designed for describing environmental observations (Fig.2), such as those represented by the CRU datasets. However, in common with the OPM, the ISO O&M model, too, is intended to be generic, albeit offering a few example specialised

[11] DataCite: http://www.datacite.org
[12] CRU Time-Series Dataset - http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__dataent_1256223773328276

[13] A series of values measured at different points of time as the result of an observation.

observation types, such as Temporal Coverage Observation [7] for time-series.

3) *Climate Science Modelling Language (CSML)*

CSML was originally developed as part of the NERC Data Grid (NDG) project in the UK as an application schema of GML to describe complex feature types for the atmospheric and oceanographic domain. However, it has recently been re-modelled as an application schema (i.e. profile) of the ISO O&M model specialised for representing time-series datasets (such as the CRU datasets). There is also a growing user community lead by BADC[14] developing and providing tools and software support for understanding and manipulating data encoded in CSML.

In light of the above review of the existing related information models, we have developed the workflow information model as a profile of the ISO O&M model with the observation-related concepts derived from the CSML *TimeSeriesObservation* classes [8]. The rationale here is to enable the model to be generally interoperable with both CSML and the ISO O&M model. The former would enable existing tools that support CSML to also understand the workflow model, thereby facilitating processing and manipulating the datasets described. The latter (i.e. interoperability with the ISO O&M model) on the other hand, would facilitate observational datasets such as the CRU datasets being shared with a wider geospatial community, potentially through a global Spatial Data Infrastructure (SDI), such as INSPIRE.

As illustrated in Fig 2, The Workflow Information Model has been developed in UML using the Model Driven Architecture (MDA) [15] adopted by INSPIRE with the following principle concepts/classes:

a) *CW_ObservationWorkflow*[16]

This class is effectively a wrapper class that encapsulates observation instances to provide a coherent and structured view of the workflow associated with an observation dataset. By definition, this class can be used to encapsulate an instance of the ISO O&M *OM_Observation* class or any of its subclasses, such as the CSML *TimeSeries* observation classes. Therefore, it provides flexibility in terms of defining new observation types according to the types and structures of the corresponding datasets, if necessary. Furthermore, it defines a number of properties (e.g. title, ownership information etc.) to record additional metadata about a workflow.

b) *CW_Process*

This class specialises the core O&M class *OM_Process* to describe various aspects the process associated an environmental observation. In particular, it uses the ISO 19115-2:2009 Metadata-imagery class *LE_Processing* to add information about the different steps in a process through the ***processStep*** property. This facilitates provision of a comprehensive description of the various aspects of a process step including inputs and outputs, algorithm employed and processor information, such as software used and its parameters.

It should be noted that a **CW_Process** instance (and the instances of other related classes, e.g. *LE_Processing*) is re-usable as it does not record any information specific to a particular instance of a process or an observation. This effectively facilitates re-usability of a process as discussed earlier in the paper (see II). The process instance-specific information, such as the associated parameter instances are recorded in the observation instance (as *parameter* properties) to which the process corresponds.

c) *CW_Station*

This class describes a climate monitoring station as an ISO O&M Sampling Feature[17] and is based on the definition provided by the World Meteorological Organisation (WMO) [18]. By definition, the class CW_Station is a specialisation of the ISO O&M class *SF_SamplingPoint*, which defines a number of properties to describe the geospatial aspects (e.g. geographical location) of the climate monitoring stations at which the source/raw observation datasets are collected.

d) *CW_StationMetadata*

This is an extension of the ISO 19115-2:2009 class *MI_Platform*, used by the instances of the **CW_Station** class for describing various non-geospatial aspects of a Climate Monitoring Station, such as station identifier and ownership information.

e) *CW_ObservationMetadata*

This class is an extension of the class *MD_Metadata* from the ISO 19115:2010 Metadata model. In particular, this class uses the “contact” and “identificationInfo” properties of the class *MD_Metadata* to describe ownership and constraints (e.g. for use and access) related information associated with an observation. Additionally, it enables recording (through instances of the class **CW_RelatedResource**) of references or links (i.e. URLs) to publications or related observations or any other resources that are of relevance to the observation being described but exist externally to it.

C. *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

[14] The British Atmospheric Data Centre (BADC) - http://badc.nerc.ac.uk/home/index.html

[15] A platform-independent modelling technique that uses a common but domain-specific modelling language, such as the Unified Modelling Language (UML). More information about MDA is available at: http://en.wikipedia.org/wiki/Model-driven_architecture

[16] In keeping with the ISO UML class naming convention, the UML class names in the ACRID Workflow UML model contains a two-letter prefix “CW”, which is an abbreviation of “CRU Workflow”.

[17] A feature, such as a station, transect, section or specimen, which is involved in making observations concerning a particular application domain. [7, definition 4.16]

[18] World Meteorological Organization (WMO) - http://www.wmo.int/pages/index_en.html

RDF/OWL ontology representation of the model[19]. This has also involved creating unofficial ontology representations[20] of the ISO O&M model and CSML as well as a number of other related ISO models (e.g. ISO 19115-2:2009) as no formal ontologies for these models currently exist.

Dissemination of the linked-data instances of the workflows is done using the OAI-ORE 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 Resource that constitutes (together with other resources) an Aggregation. Examples include a workflow instance and a related publication.
- **Resource Map (ReM):** a brief description of an Aggregation.

So, as illustrated in Fig 3, 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[21] 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. RDF Schema '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 crucial 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).

V. 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[22]. 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.

For example, the construction of the gridded CRUTEM monthly temperature dataset (current version CRUTEM3[23]), including the various processing steps and a gross description of the data sources, is described in a sequence of papers published over the last 25 years (see [9] and references therein). However, information on the precise provenance of each individual value within the underlying CRUTEM station monthly temperature database is not easily accessible, though with access to internal records and time to make comparisons between original data and the current version, this would be possible in most if not all cases. The workflow model and the linked-data approach presented in this paper should enable providing more transparent provenance between source data and final published results concerning CRUTEM3.

To this end, we first designed a data management infrastructure (Fig. 4) 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. The information captured is then stored and exposed as linked-data in accordance with the approach described in (IV, Section C) through a linked-data server, namely the ACRID Linked Workflows Server (ALWS)[24]. Two separate data

[19] ACRID Workflow Ontology - http://www.cru.uea.ac.uk/cru/projects/acrid/ontologies/cw/cru_workflow.owl
[20] ACRID ISO O&M Ontology - http://www.cru.uea.ac.uk/cru/projects/acrid/ontologies/om/iso-19156-om.owl
[21] The level of detail of an OAI-ORE Aggregation provided in the corresponding Resource Map is left open to specific implementation approaches.
[22] CRU Yamal tree-ring data - http://www.cru.uea.ac.uk/cru/people/briffa/yamal2009/data/
[23] Temperate data (HadCRUT3 and CRUTEM3) - http://www.cru.uea.ac.uk/cru/data/temperature/
[24] ACRID Linked Workflows Server - http://westerly.badc.rl.ac.uk:8080/alws/index.html

stores (based on the PostgreSQL relational database – Fig. 4) 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.

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 (Fig. 3). A key aspect of this citation infrastructure is a "data publishing" function incorporated within the 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.

The linked-data server used for ACRID is based on GeoTOD[25] - 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 [4] 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).

VI. Conclusions and Future Directions

A. *Opening up Climate Research*

The requirement for the publication of data provenance has been highlighted in the UK's House of Commons Science and Technology Committee report into the release of private emails at the CRU [12] which noted that although CRU's *"(data sharing) actions were in line with common practice in the climate science community"* they 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 even so, *"it is not standard practice in climate science to publish the raw data and the computer code in academic papers*". The work discussed in this paper has aimed to address this issue directly, by developing a linked-data approach to exposing the key concepts needed to describe both the important steps in data production and the final products. In essence, this is achieved by combining the widely adopted DOI mechanism with pre-existing domain specific models, such as the ISO 19156 O&M and CSML, for describing climate phenomena and their measurement.

B. *Lessons Learned*

This work represents only the start of a journey towards developing an operational approach to publishing research data with associated provenance and workflow metadata. A number of issues remain to be addressed, chief amongst them: whether or not the use of the ISO19115-2 process step formalism can capture enough information for workflow re-use (or is it best for capturing descriptive information, and limited key run-time parameters); and whether or not workflow re-use is desirable, necessary, or redundant in this context. The answers are probably domain dependent, but the work we have done here could serve as an exemplar for further investigation in climate science, and for extension and answers in other domains.

Further, we indicated earlier the tension between DOI requirements for static resources and the often dynamic, versioned nature of scientific data. As well, relevant information models and ontologies must be developed and agreed by domain-specific research communities. Such community agreement alone, however, will not suffice without uptake also by academic publishers (in turn this requires a sustainable ecosystem of institutional and domain-specific data repositories). On a technical level, the data publishing approach must be supported by robust tooling and software. Not least, 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).

C. *Future Directions*

Regardless of the questions/issues above, 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 atypical 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.

The methodology presented here should be deployable elsewhere within the climate and other environmental sciences, and with suitable adaptation to the model of data used, could also be used 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 should also be useful to develop suitable mechanisms for mapping the Workflow Model presented above (see IV, Section B) on to the workflow description languages used by some of the widely used workflow execution engines, such as Taverna. This should effectively

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

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 must 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 NERC data centres (see III).

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.

ACKNOWLEDGMENT

We sincerely thank Spiros Ventouras, Dominic Lowe and Ag Stephens of the British Atmospheric Data Centre (BADC), and Jeremy Tandy of the UK Met Office for their expert advice and guidance on the development and validation of the ACRID Workflow Model.

REFERENCES

[1] S. Callaghan, S. Pepler, F. Hewer, P. Hardaker, and A. Gadian "How to publish data using overlay journals: the OJIMS project" Ariadne Issue 61, October 2009, Uniform Resource Locator (URL): http://www.ariadne.ac.uk/issue61/callaghan-et-al/ Last accessed: 1-Jul-2011

[2] S. Callaghan, F. Hewer, S. Pepler, P. Hardaker and A. Gadian "Overlay Journals and Data Publishing in the Meteorological Sciences " Ariadne Issue 60, July 2009, Uniform Resource Locator (URL):http://www.ariadne.ac.uk/issue60/callaghan-et-al/ Last accessed: 1-Jul-2011

[3] B.N. Lawrence, C.M. Jones, B.M. Matthews, S.J. Pepler and S.A. Callaghan "Data publication", in press, International Journal of Digital Curation, 2011.

[4] 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. Uniform Resource Locator (URL): http://eprints.ecs.soton.ac.uk/21587/5/research-objects-final.pdf Last accessed: 1-Jul-2011

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

[6] T. Osborn, C. Harpham, I. Harris, A. Shaon and S. Callaghan "Description of Scientifc Workflows", JISC Project Report for ACRID. Uniform Resource Locator (URL): http://www.cru.uea.ac.uk/cru/projects/acrid/ACRID_D2.1_scientific workflows.pdf Last accessed : 1-Jul-2011

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

[8] D. Lowe and A. Woolf, "CSML 3: Climate Science Modelling Language - MetOcean DWG" Presentation, 76th OGC Technical Committee Bonn, Germany, 2011. Uniform Resource Locator (URL): http://external.opengis.org/twiki_public/pub/MetOceanDWG/MetOceanDWGBonn/CSMLV3_Lowe.pdf Last accessed: 1-Jul-2011

[9] P. Brohan, J. Kennedy, I. Harris, S.F.B. Tett and P.D. Jones, "Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850" J. Geophys. Res. 111, D12106, 2006. doi:10.1029/2005JD006548.

[10] Chief Technology Officer Council "Designing URI Sets for Location", Cabinet Office, v0.5, July 2010, Uniform Resource Locator (URL): http://location.defra.gov.uk/wp-content/uploads/2010/04/Designing_URI_Sets_for_Location-Ver0.5.pdf Last accessed: 1-Jul-2011

[11] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan and J. Van den Bussche, "The Open Provenance Model core specification (v1.1)" Future Generation Computer Systems, Volume 27, Issue 6, June 2011.

[12] House of Commons Science and Technology Committee "The disclosure of climate data from the Climatic Research Unit at the University of East Anglia" House of Commons Science and Technology Committee, Eighth Report of Session 2009–10, 31 March 2010, Uniform Resource Locator (URL): http://www.publications.parliament.uk/pa/cm200910/cmselect/cmsctech/387/387i.pdf Last accessed: 1-Jul-2011

Figure 1: The main concepts of the Open Provenance Model

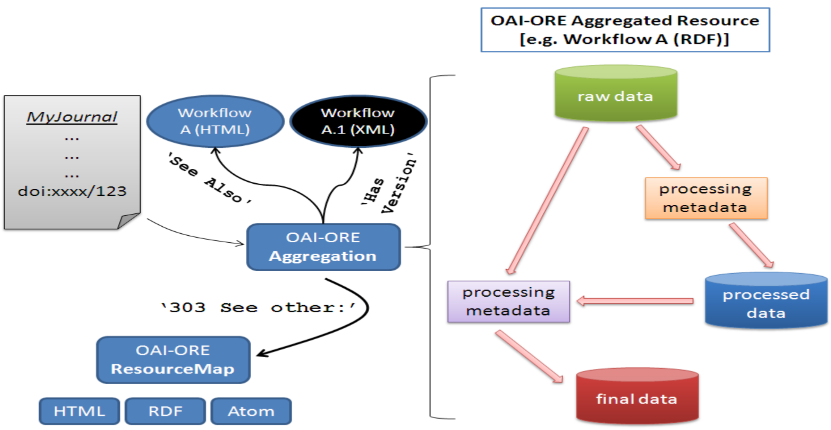

Figure 3: An OAI-ORE representation of linked workflows

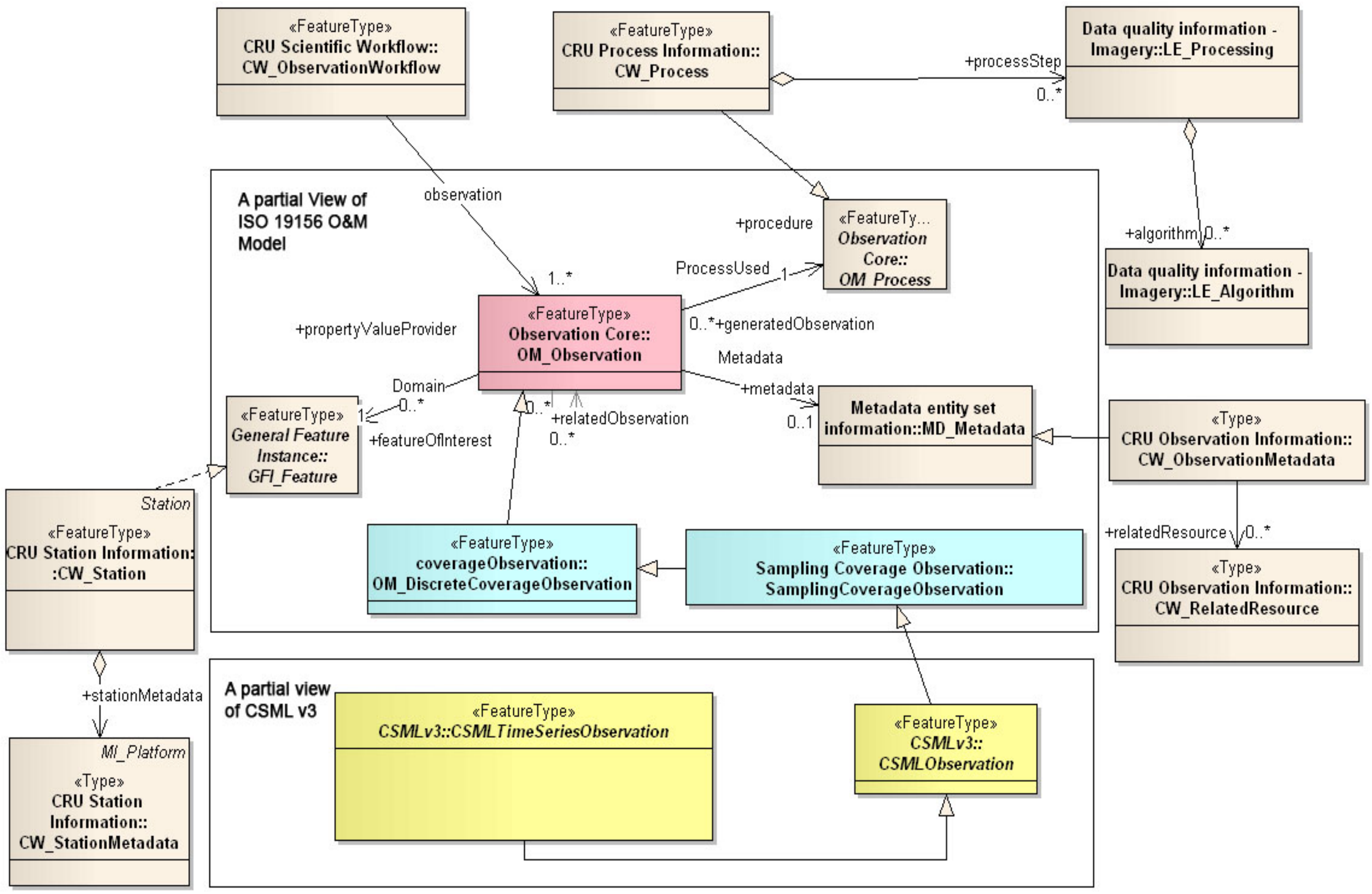

Figure 2: A Workflow Model for Geospatial Datasets

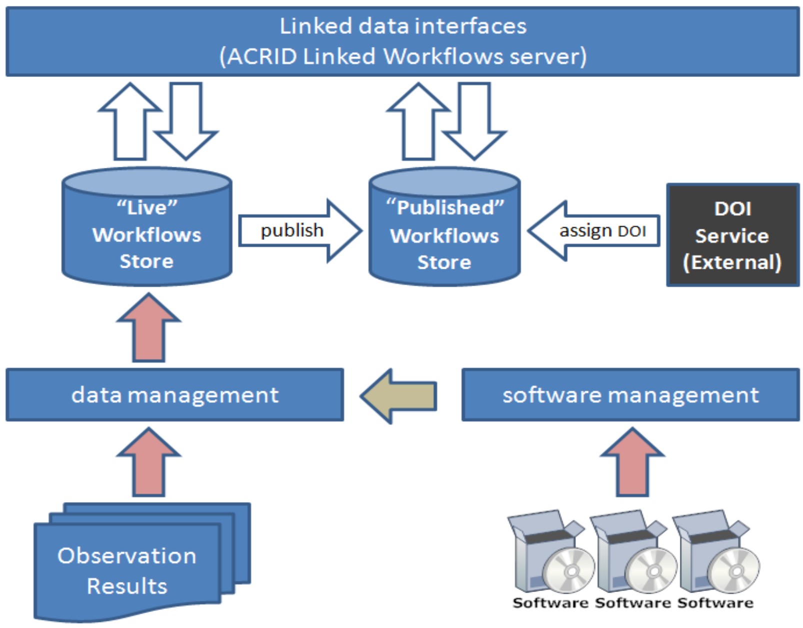

Figure 4: CRU Data Management Infrastructure