The International Journal of Digital Curation
Volume 7, Issue 1 | 2012

Opening Up Climate Research: A Linked Data Approach to Publishing Data Provenance

Arif Shaon, Sarah Callaghan, Bryan N. Lawrence and Brian Matthews,
STFC Rutherford Appleton Laboratory, UK

Timothy Osborn and Colin Harpham,
The Climatic Research Unit,
University of East Anglia, UK

Andrew Woolf,
The Bureau of Meteorology, Canberra, Australia

Abstract

Traditionally, the formal scientific output in most fields of natural science has been limited to peer-reviewed academic journal publications, with less attention 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. Detailed knowledge of a dataset's provenance is essential to establish the pedigree of the data for its effective re-use, and to avoid redundant re-enactment of the experiment or computation involved. It 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. To address these issues, we present an approach that combines the Digital Object Identifier (DOI) – a widely adopted citation technique – with existing, widely adopted climate science data standards to formally publish detailed provenance of a climate research dataset as an associated scientific workflow. This is integrated with linked-data compliant data re-use standards (e.g. OAI-ORE) to enable a seamless link between a publication and the complete trail of lineage of the corresponding dataset, including the dataset itself.

 doi:10.2218/ijdc.v7i1.223

Introduction and 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 paid to the chain of intermediate data results and their associated metadata, including provenance. This has effectively constrained the representation and verification of the data provenance to the confines of the related publications. However, this culture has started to change owing to initiatives such as the OJIMS[1] and CLADDIER[2] 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.

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.

Detailed knowledge of a dataset's provenance is essential to establish the pedigree of the data for its effective re-use, to avoid redundant re-enactment of the experiment or computation involved. For example, when sharing the result of an analysis of a set of global temperature records, the presence of the assumptions or decisions made during the analysis gives a context in which it can be re-used and also credits the scientists(s) involved. Additionally, this level of granularity of data provenance is also important for scientific workflows, particularly for ensuring repeatability as well as validation of the related scientific claims made. It 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 provenance 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 highlighted in the UK's 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 (House of Commons Science and Technology Committee, 2010), 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."

[1] The OJIMS: http://proj.badc.rl.ac.uk/ojims

[2] The CLADDIER Project: http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2005/claddier.aspx

A key motivator driving the citation and publication of environmental datasets is the requirement that the creators of those datasets receive academic credit for the considerable work they put into 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.

There have been a number of projects working on data publication per se, and an extensive assessment of these approaches can be found in (Lawrence et al., 2011), so is not repeated here.

In this paper, we present the outcomes of the Advanced Climate Research Infrastructure for Data (ACRID) project, which has taken the climate/research datasets held by CRU as exemplars to address the issues of publishing detailed provenance associated with complex environmental datasets. In essence, the ACRID project has developed a linked data approach to exposing detailed scientific workflows, including the key concepts needed to describe both the important steps in data production and the final products, thereby providing greater transparency of the provenance of the corresponding dataset.

The Core Requirements for Publishing Data Provenance

The task of publishing the complete record of provenance associated with complex scientific datasets needs to meet a number of core requirements, as identified in (Bechhofer et al., 2010). In particular, for publishing the provenance records of geospatial datasets (the premise of the work presented here), these requirements are the following:

Repeatability and Reusability of Scientific Workflows

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, the published provenance information about a scientific dataset should be adequate to facilitate accurate re-enactment or repetition of the associated workflow to help verify the evidential basis of the claims in the publications.

It is also a common practice for the components of a scientific 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), for example, the height at which the temperature is measured. Therefore, if possible, the provenance record associated with a scientific dataset should contain sufficient information about the constituents of its workflow to facilitate their re-usability.

Common Information (Provenance) Model and Citability

To ensure greater re-usability, a provenance record associated with a scientific dataset should be underpinned by an information model that is understood by the wider user community. In the case of dynamic or evolving datasets, such an information model would need to address data versioning and other related aspects.

Driven by the INSPIRE Directive in Europe, the ISO 19100 series information of standards (such as ISO 19156 Observations and Measurements[3]) 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 developed based on these ISO standards (as appropriate) would have the potential to be more widely applicable and shareable than any bespoke description of that workflow. In addition, an instance of such a common standard approach would be more easily consumable as a citation in scholarly communications.

Efficient Metadata Curation Strategy

The ability to publish detailed provenance-related information about a dataset is heavily reliant on the effective curation of the associated metadata. This would need to involve capturing accurate metadata at crucial junctures of the data lifecycle, quality assurance, efficient management (e.g. versioning) of the metadata captured, and finally storing it, ideally in a medium that is suitable for efficient querying and dissemination of the metadata. Without effective curation, the metadata may become out of step with the data, which may lead to inaccurate and/or incomplete provenance description of the data.

The ACRID Approach

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 (see Figure 1 or Osborn et al., 2011).

[3] ISO 19156:2010 Geographic information — Observations and measurements: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32574

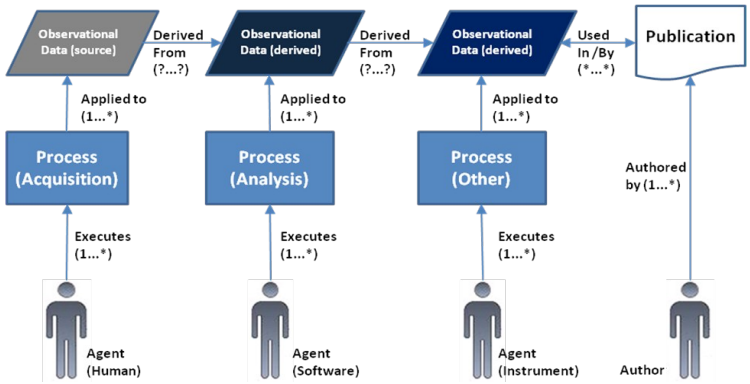

Figure 1. CRU dataset workflows

These workflow constituents can be generalised into the following concepts:

- **Observation**: The act of measuring or calculating a particular property (such as temperature) associated with a certain feature of interest (such as 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** (such as the 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.
- **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. As is the case with the workflows discussed earlier, a process used for one observation may be used for another, though the observation-specific parameters – such as process inputs and outputs – may be different and hence, non-reusable.
- **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.

[4] CRU time-series dataset: 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.

 doi:10.2218/ijdc.v7i1.223

The ACRID Provenance Model for Geospatial Workflows

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 the Open Provenance Model (OPM) (Moreau et al., 2010), the ISO 19156 Observations and Measurements (O&M) model and Climate Science Modelling Language (CSML)[6]. The review was conducted in consultation with a number of domain experts 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. In essence, CSML is 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[7], 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.

Therefore, we developed an information model as an application schema of the ISO O&M model with the observation related concepts derived from the *CSMLTimeSeriesObservation* classes (see Figure 2). This model is primarily intended to facilitate detailed and accurate description of the three main aspects of climate research data outlined before. The model was mainly developed in UML, with the underlying concepts additionally represented in RDF to facilitate linked data representations of the associated workflows (described later), and GML to enable compatibility with the CSML and other related tools. A complete description of the ACRID information model is provided by Shaon, Ventouras and Tandy (2011).

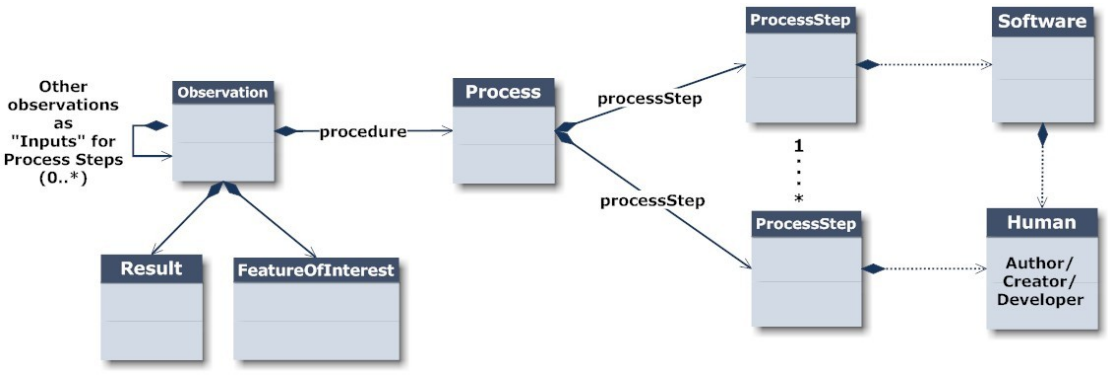

Figure 2. An overview of the ACRID provenance model.

[6] CSML: http://csml.badc.rl.ac.uk/

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

Publishing Linked Provenance Records using OAI-ORE and DOI

To publish provenance records defined by the ACRID provenance model, outlined above as linked data, we have developed an RDF/OWL ontology representation of the model.[8] This has also involved creating unofficial ontology representations[9] 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 provenance records 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, and
- **Resource Map (ReM):** A brief description of an Aggregation.

So, as illustrated in Figure 3, the provenance record for a CRU workflow would be encapsulated within an OAI-ORE Aggregation as an Aggregated Resource. In order to publish the record, 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 client is redirected (using an HTTP 303 re-direct as recommended by linked data principles) from the Aggregation URI to the URI of the Resource Map that describes the Aggregation.

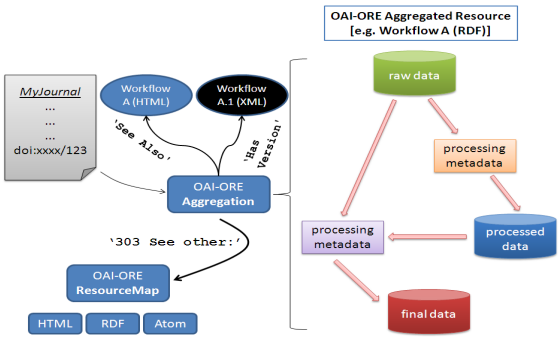

Figure 3. An OAI-ORE representation of linked provenance records for climate research workflows.

The Resource Map serves as a landing or splash page, providing a description[10] of the Aggregation (not the Aggregated Resource), which includes the URI for the Aggregated Resource (for example, a provenance record). The client is then able to

[8] ACRID Workflow Ontology: http://www.cru.uea.ac.uk/cru/projects/acrid/ontologies/cw/cru_workflow.owl

[9] ACRID ISO O&M Ontology: http://www.cru.uea.ac.uk/cru/projects/acrid/ontologies/om/iso-19156-om.owl

[10] The level of detail of an OAI-ORE Aggregation provided in the corresponding Resource Map is left open to specific implementation approaches.

 doi:10.2218/ijdc.v7i1.223

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 using an appropriate vocabulary (e.g., the RDF Schema 'seeAlso' shown in Figure 3). 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, such as HTML, RDF, or Atom (see Figure 3), based on the client's request.

Prototype Implementation and Validation

We have tested our linked data approach using three distinct datasets published by CRU:

1. CRUTEM land-surface air temperature data (specifically version CRUTEM3);
2. CRU TS land-surface high-resolution data for multiple variables (specifically version CRU TS 3.1); and
3. 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.

An Improved Data Curation Infrastructure for the CRU Datasets

To this end, the ACRID project first assessed the current data management infrastructure of CRU and identified the lack of efficient mechanisms for capturing important information about the processes executed on their datasets. Based on this assessment, the project has made a number of significant improvements to the existing data management infrastructure of CRU to accurately and efficiently capture and manage provenance-related information (as defined by the provenance model) about the workflows associated with the three aforementioned datasets. In particular, various metadata capturing tools and scripts have been implemented and integrated within the CRU data management infrastructure to capture important metadata at various important stages of the lifecycles of the datasets. While the mechanisms employed for capturing metadata vary between different stages of the data lifecycle, they are either fully or semi automated wherever possible. For example, the scripts are able to automatically capture the inputs, outputs and other related parameters of a software run that is a part of an analytical process conducted on a dataset. This has significantly improved the curation of the CRU's data and the associated metadata.

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

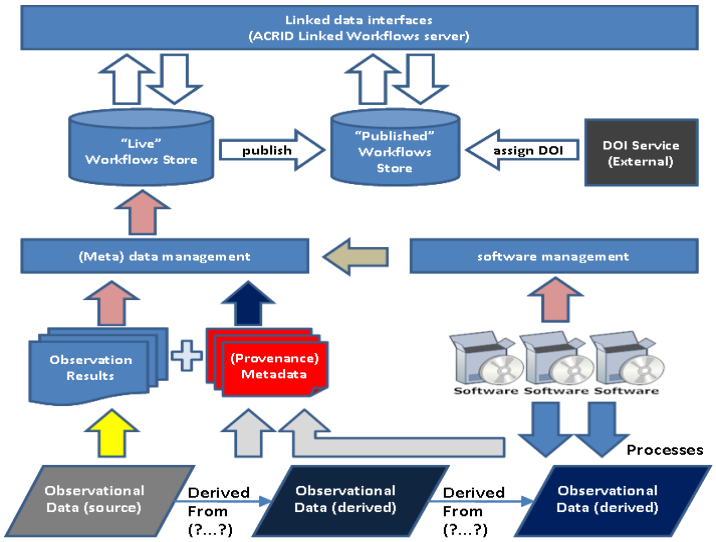

Figure 4. CRU data management and publishing infrastructure.

The ACRID Linked Workflows Server

The information captured is quality assured, versioned, finally stored, and exposed as linked data in accordance with the approach described before through a linked data server, namely the ACRID Linked Workflows Server[12]. Two separate data stores (see Figure 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.

The ACRID Data Citation Infrastructure

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) (see Figure 4). 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 (see Figure 4) 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 workflows to be published in order to distinctly identify a workflow and the format in which it has been published.

Conclusions and Future Directions

In summary, the ACRID approach combines the Digital Object Identifier (DOI) with existing, widely adopted, climate science data standards (and profiles thereof, such as the ISO 19156 O&M model and CSML) to formally publish detailed provenance of a climate research dataset as an associated scientific workflow. This is integrated with linked data compliant data re-use standards (e.g. OAI-ORE) to enable a seamless link

[12] The ACRID Linked Workflows Server: http://westerly.badc.rl.ac.uk:8080/alws/index.html

 doi:10.2218/ijdc.v7i1.223

between a publication and the complete trail of lineage of the corresponding dataset, including the dataset itself. From a wider perspective, the ACRID approach has the potential to facilitate greater transparency and traceability of the lifecycle of climate research data, and thereby open up climate research.

More importantly, the work presented here identifies the need for effective metadata curation to enable accurate capturing, quality assurance and management of the provenance and other related information associated with complex scientific datasets. The ACRID project has stimulated and supported a number of improvements to the management and curation of research data within the CRU. 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 (i.e. Revision Control System) to a more modern and flexible system (i.e. Subversion). 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. workflows) that underpin widely used climate datasets.

Finally, 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.

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 ISO 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.

Acknowledgements

The work presented in this paper has been funded by the JISC Managing Research Data (JISCMRD) programme.[13] We also 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 Provenance Model.

[13] JISC MRD programme: http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx

References

Bechhofer, S., Ainsworth, J., Bhagat, J., Buchan, I. Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., Gamble, M., Goble, C., Michaelides, D., Missier, P., Owen, S., Newman, D., De Roure, S. & Sufi S. (2010). Why linked data is not enough for scientists. Paper presented at the 10th IEEE e-Science Conference, Brisbane, Australia. Retrieved from http://eprints.ecs.soton.ac.uk/21587/5/research-objects-final.pdf

House of Commons Science and Technology Committee. (2010). *The disclosure of climate data from the Climatic Research Unit at the University of East Anglia in the Eighth Report of Session 2009–10, 31 March 2010.* House of Commons Science and Technology Committee, UK. Retrieved from http://www.publications.parliament.uk/pa/cm200910/cmselect/cmsctech/387/387i.pdf

Lawrence, B.N., Jones, C.M., Matthews, B.M., Peplar, S.J. & Callaghan, S.A. (2011). Citation and Peer Review of Data: Moving Towards Formal Data Publication. *International Journal of Digital Curation, 6*(2). Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/181

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. & Van den Bussche, J. (2010). The Open Provenance Model core specification (v1.1). *Future Generation Computer Systems*. Retrieved from http://eprints.ecs.soton.ac.uk/id/eprint/21449

Osborn, T., Harpham, C., Harris, I., Shaon, A. & Callaghan, S. (2011). *Description of scientific workflows.* JISC Project Report for ACRID. Retrieved from http://www.cru.uea.ac.uk/cru/projects/acrid/ACRID_D2.1_scientificworkflows.pdf

Shaon, A., Ventouras, S. & Tandy, T. (2011). *Report II: Work package 2.2 information architecture.* Retrieved from http://www.cru.uea.ac.uk/cru/projects/acrid/ACRID_D2.2_informationarchitecture.pdf