SCIENCE AND TECHNOLOGY FACILITIES COUNCIL

Advanced Climate Research Infrastructure for Data (ACRID)

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

Arif Shaon, Sarah Callaghan (STFC)

1 August 2011

Science & Technology Facilities Council
e-Science

University of East Anglia

Revision History:

Version	Date	Authors	Sections Affected / Comments
0.1	01/08/2011	AS & SAC	Outline and Main Body

Contents

1 Introduction and Rationale

In ACRID, we have developed a linked-data approach to formally publishing detailed workflows associated with complex climate research datasets, such as the datasets held by the Climatic Research Unit (CRU) of the University of East Anglia. In effect, our approach involves the adoption of the OAI-ORE[1] technology to disseminate linked-data representations of a workflow described by the ACRID workflow model [1]. In essence, OAI-ORE defines standards for the description and exchange of aggregations of Web-based resources (e.g. a CRU dataset workflow instance) in a linked-data compliant way. Additionally, we have also developed and deployed a linked-data server, the ACRID Linked Workflows Server (ALWS), to expose the dataset workflows encapsulated within OAI-ORE aggregations through linked-data compliant HTTP URIs [2].

In this report, we detail the development of an infrastructure to enable citation of these workflows within the context of scholarly communication. This involves formally publishing the OAI-ORE aggregation of a scientific workflow, using the Digital Object Identifier (DOI) technique. Notably, the DOI mechanism has been chosen by the NERC data citation and publication project to enable citation of scientific datasets held in NERC data centres and the ACRID project draws from this project. Additionally, we demonstrate the infrastructure by exposing a number of distinct datasets held by the CRU and the UK Met Office.

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

2 ACRID Dataset Workflow Citation Infrastructure

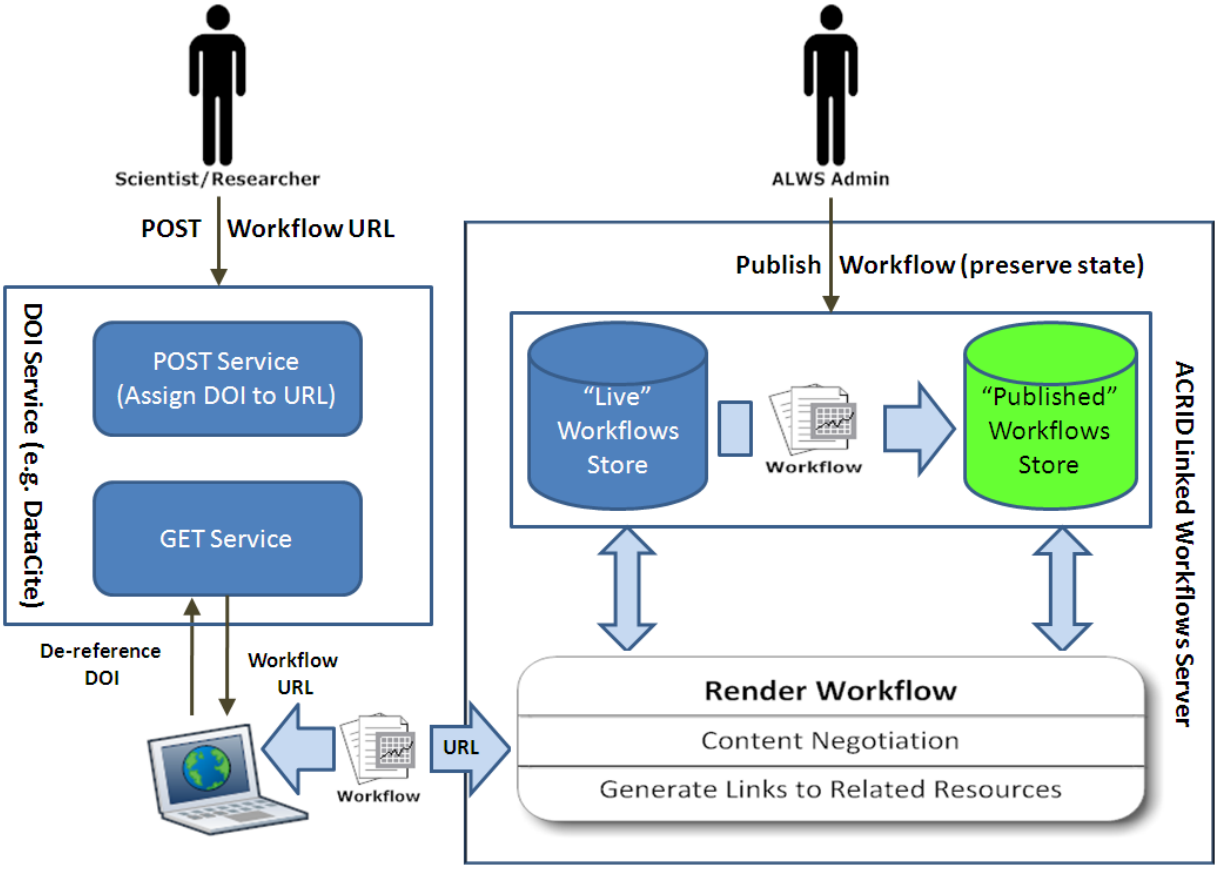

Figure 1: ACRID workflow citation infrastructure

As illustrated in Figure 1, publishing a citable workflow using the ACRID approach comprises the following steps:

2.1 Preserving the state of a published workflow

To comply with the DOI rules, a published workflow should remain static in terms of both contents and integrity for an indefinite period of time. Therefore, the published workflows should be managed separately from the 'Live' workflows, which typically represent volatile datasets. For example, the HadCET dataset[2] held by the UK Met Office is updated daily – the workflow associated with it would be a 'Live' workflow. The management of these two types of workflows could be conducted in either logically or physically separate environments. In

[2] Met Office Hadley Centre HadCET Observations Dataset - http://www.metoffice.gov.uk/hadobs/hadcet/graphs/index.html

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.

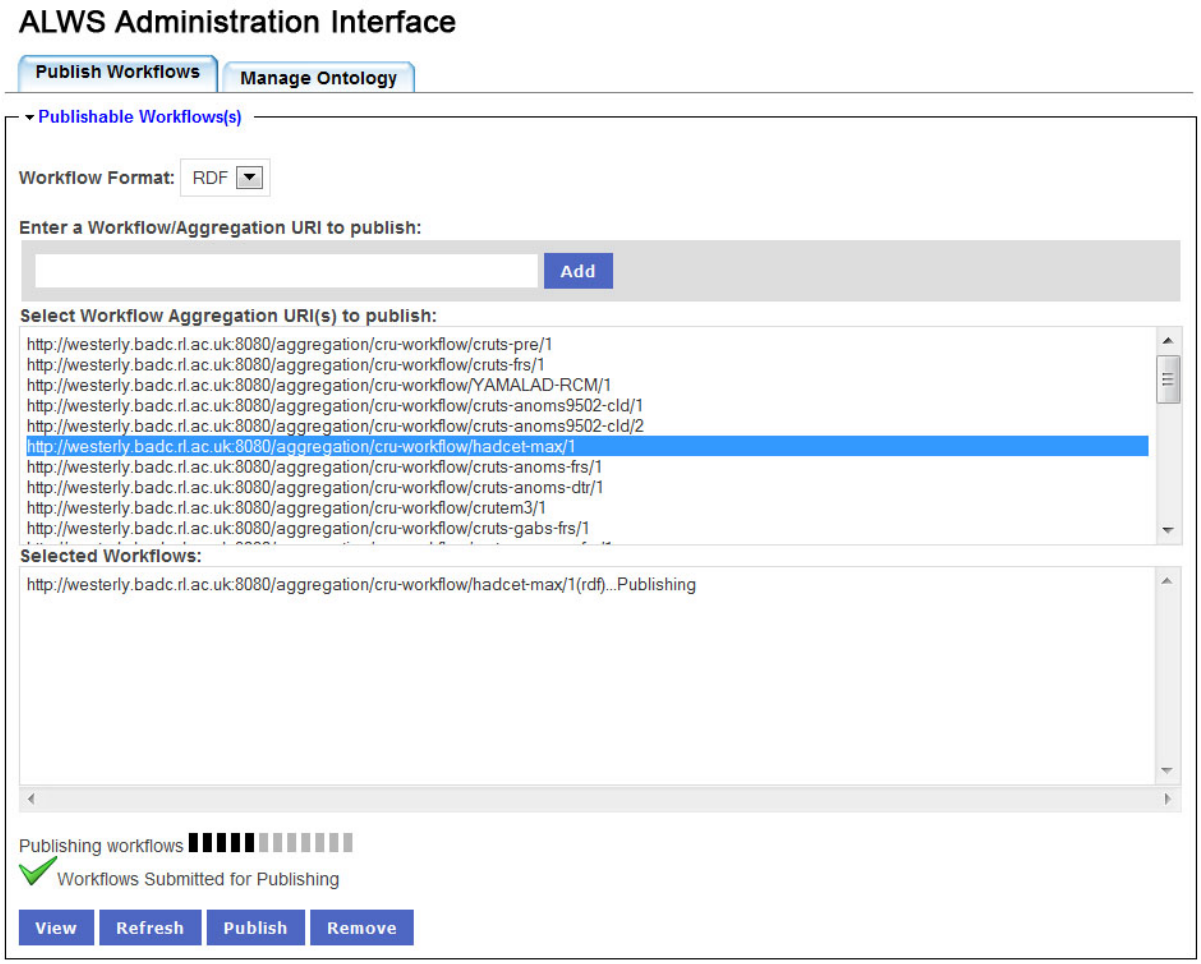

Figure 2: ALWS Workflow publishing interface

To that end, we have added a "data publishing" function to ALWS that is accessible through a secure, user-friendly and intuitive web interface (Figure 2). 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 1). In effect, this involves automatically traversing through all the linked constituents (e.g. process description) of a workflow and copying them across to the "Published" workflows store. In addition, unique URIs are assigned to the published workflows in order distinctly identify a workflow and the format in which it has been published. So, for example, the URI for a workflow published in RDF would be,

*http://domain/so/**published**/CW_Workflow/cw/origin/version*

while the same workflow published in GML would be:

*http://domain/so/**published-gml**/CW_Workflow/cw/origin/version*

The current version of ALWS supports publishing workflows only in RDF and GML, but can be extended to support other formats, if needed.

2.2 Assigning a DOI to a published workflow

To assign a DOI to a published workflow, the publisher must first have a DOI minting account. In the UK these are provided through the British Library, who are a member institute of the DataCite, itself a member of the International DOI Foundation. DataCite aims to "enable data citation and facilitate easier access to scientific research data on the Internet." NERC has been assigned a DOI minting account and prefix to enable its data centres to assign DOIs to datasets held in their repositories.

DataCite has provided RESTful API to allow data centres to mint DOIs for their datasets (https://mds.datacite.org/static/apidoc). It is recommended that users integrate minting and updating DOIs with their metadata management infrastructure. So if, for example, a URL changes, automatic notification will be pushed to the MDS service and the updated URL will be used for resolving the DOI. Metadata about the datasets must also be submitted when the DOI is minted, and this metadata must conform to the DOI metadata schema available from http://schema.datacite.org/ .

DOIs can also be minted by hand via the MDS web form, though this is not advised for large numbers of datasets. At its simplest, this requires registering the DOI and its http landing page (see Figure 3), and then uploading the xml file containing the DOI-specific metadata (Figure 4)

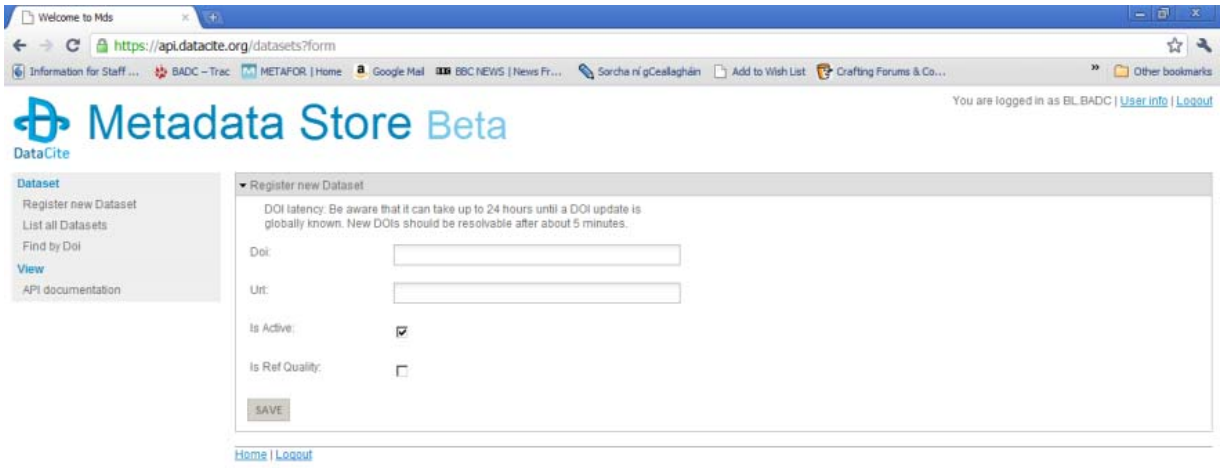

Figure 3: The MDS webform for minting DOIs by hand – registering a DOI and accompanying url.

Figure 4: The MDS webform for minting DOIs by hand – uploading the xml document containing the DOI specific metadata

DOIs must resolve to a landing page which is open – this is discussed further in the next section for the specific case of ACRID.

2.3 Exposing a published workflow through DOI

The DOI for a published workflow de-references to the URL of its OAI-ORE Aggregation, which itself is re-directed to an OAI-ORE Resource Map (ReM) document, in accordance with the OAI-ORE specification. In effect, the ReM document serves as a *landing or splash page* providing a description of the Aggregation (not the Workflow itself), which includes the URI for the Aggregated Resource i.e. the published workflow. This enables a client to de-reference the workflow URI to retrieve it. To ensure the integrity of both the contents and the format of a published workflow, the corresponding ReM document always points to (through its unique URI – Section 2.1) the format (e.g. RDF, GML) in which the workflow was originally published. A ReM document itself however, may be provided in any supported format (e.g. HTML, RDF) requested by the client, independent of the format of the corresponding published workflow.

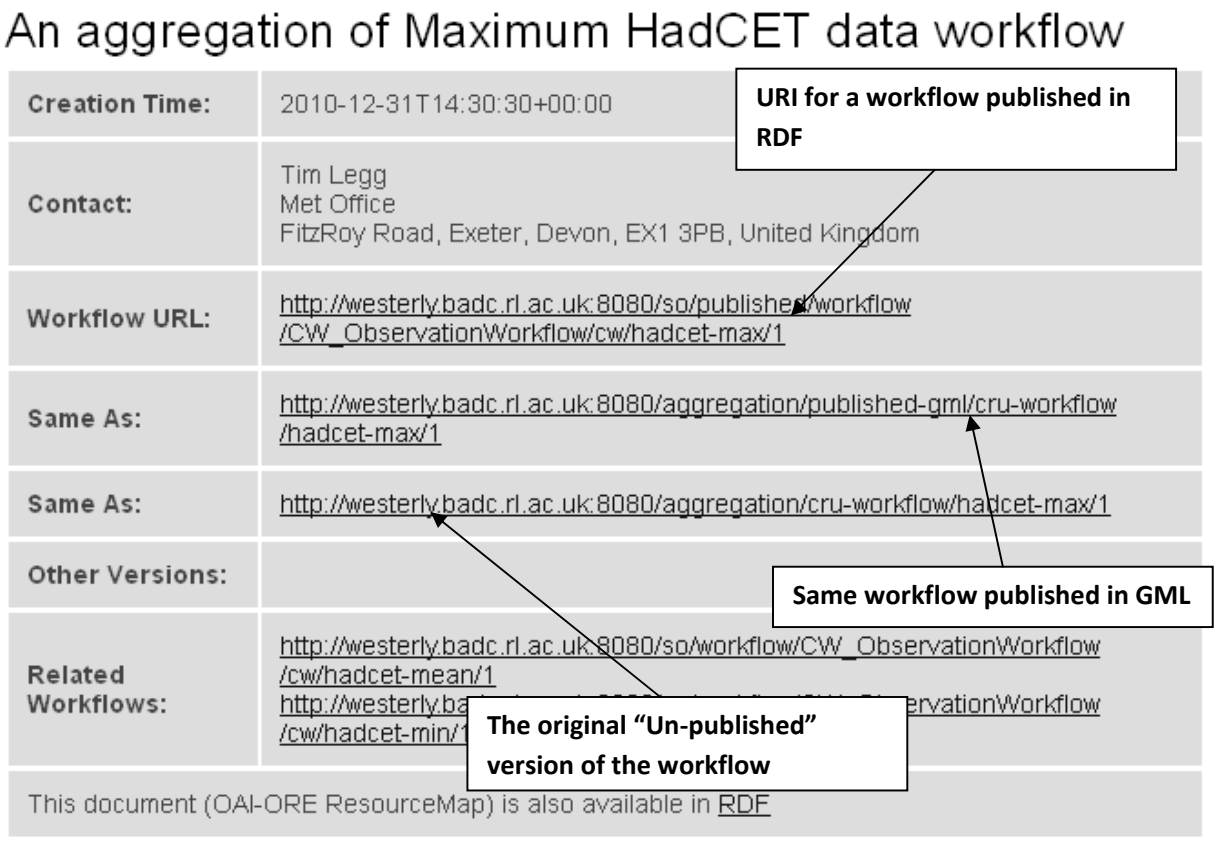

Figure 5: The OAI-ORE Aggregation description (Resource Map) of an example workflow published through ALWS

The Aggregation description contained within a ReM document may also include information about other static or non-static resources related to the Aggregated Resource. For example, the link to a newer version or a different representation (published or un-published) of the workflow instance may be provided in the Aggregation using an appropriate vocabulary (e.g. RDFS 'seeAlso', OWL 'sameAs'). This effectively 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. To facilitate this, the ALWS citation infrastructure incorporates an "on-demand" process (Figure 1) that automatically determines and generates link(s) to the resource(s) related to a published workflow and embeds them in its corresponding OAI-ORE Aggregation description (Figure 5), when requested by a client.

3 Example Workflows

We have tested our linked-data approach using three distinct datasets published by the CRU. These datasets are:

i. CRUTEM land-surface air temperature data; the OAI-ORE Aggregation URI for this dataset is: http://westerly.badc.rl.ac.uk:8080/aggregation/cru-workflow/crutem3/1

ii. CRU TS land-surface high-resolution data for multiple variables; the OAI-ORE Aggregation for CRU TS: http://westerly.badc.rl.ac.uk:8080/aggregation/cru-workflow/cruts-cld/1

iii. a network of tree-ring chronologies[3] across the Northern Hemisphere; the corresponding workflow URI is: http://westerly.badc.rl.ac.uk:8080/aggregation/cru-workflow/YAMALAD-STD/1

In addition, we have also applied the ACRID linked-data approach to the Hadley Centre's Central England Timeseries dataset (HadCET) published by the UK Met Office. The OAI-ORE Aggregation URI for this dataset is: http://westerly.badc.rl.ac.uk:8080/aggregation/cru-workflow/hadcet-mean/1

4 Conclusions

In the current and final phase of ACRID, we have developed an infrastructure to enable citation of the workflows associated with climate research datasets within the context of scholarly communication. In essence, this involves formally publishing the OAI-ORE aggregation of a scientific workflow, using the Digital Object Identifier (DOI) technique.

A key aspect of this citation infrastructure is a "data publishing" function incorporated within the ACRID Linked Workflows Server (ALWS) 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 1) in order to preserve the integrity of both the contents and the format of a published workflow – fundamental to the DOI technique. 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. We have also demonstrated the infrastructure by exposing a number of distinct datasets held by the CRU and the UK Met Office through the ACRID linked-data server.

References

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

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

[3] CRU Tree-ring data - http://www.cru.uea.ac.uk/~timo/datapages/mxdtrw.htm