7th International Digital Curation Conference
December 2011

Opening up Climate Research: a linked-data approach to exposing data provenance

Arif Shaon[1], Sarah Callaghan[2], Bryan Lawrence[2], Brian Mathews[1]

[1]E-Science Centre,

[2]NCAS British Atmospheric Data Centre

STFC Rutherford Appleton Laboratory

Timothy Osborn, Colin Harpham

The Climatic Research Unit, The University of East Anglia

July 2011

Extended Abstract

Traditionally, the formal scientific output in most fields of natural science has been limited to peer-reviewed publications. Of course datasets have been archived, and continue to be archived, but most communities have concentrated on the final output, with less attention paid to the chain of intermediate data results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. In effect, this has inadequately limited 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 and CLADDIER projects, developing mechanisms for formally publishing scientific datasets as scientific resources in their own right, rather than, as traditionally regarded, merely as an adjunct to the publication.

Publishing a dataset by itself, however, is not sufficient for providing a complete account of its provenance. Typically, 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 other related metadata together 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, thereby avoiding 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, presence of e.g. 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. This is also relevant from the perspective of regulatory mechanisms to protect intellectual property. Additionally, this level of granularity of data provenance is also important for non-deterministic scientific workflows, particularly for ensuring repeatability as well as validation of the related scientific claims made. This is increasingly important for the 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).

Further, the need 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 Climatic Research Units (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" 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".

In this paper, we present the outcomes of the Advanced Climate Research Infrastructure for Data (ACRID) project that has taken the climate/research datasets held by CRU, as exemplars to address the issues of publishing detailed provenance associated with complex environmental datasets. To this end, ACRID 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.

In our analysis, we identified the need for a common information model to describe the workflow associated with a dataset in order ensure its greater re-usability. Driven by the INSPIRE Directive in Europe, the ISO 19100 series information models and standards (e.g. ISO 19156 O&M model) 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. This observation also applies to other existing provenance models, such as the Open Provenance Model, which, albeit conceptually applicable, is too generic and uncommon within the geospatial community to be effectively applicable to geospatial datasets like the CRU datasets.

We also identified that the ability to link resources (using linked-data techniques) 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) , 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) 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. 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.

To address the aforementioned issues, the ACRID approach combines the Digital Object Identifier (DOI) - a widely adopted citation technique - with existing widely adopted climate science data models (e.g. ISO 19156 O&M model and CSML) to formally publish detailed provenance of a climate research dataset as the 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.