

# Towards Quantification of Limits in Automated Curation of e-Science Data

Beth Plale<sup>1</sup>, You-Wei Cheah, and Yiming Sun  
Department of Computer Science, Indiana University

## Abstract

*Workflow systems are an increasingly popular e-Science tool for executing complex sequences of tasks. The large volumes of data created during the course of these computationally intense and data-driven scientific investigations drives research in techniques to automate metadata capture to relieve the burden on the user of manual annotation. In this paper we describe our experience to date in quantifying the limits of automated metadata collection in e-Science workflow systems.*

## 1. Motivation

The scientific knowledge discovery process increasingly utilizes the vast number of information sources available on the Web giving rise to new forms of knowledge derived through synthesis, analysis, modeling, and mining of vast volumes of data. The availability and accessibility of real time data acquired through sensing the environment, for instance, opens new vistas for predicting future behavior. Current readings of atmospheric conditions from the Doppler radar, for instance, can derive more accurate regional weather forecasts than can the continental forecasts that are routinely run over the country.

The creation of new data at high volumes over short periods of time, such as can occur in scientific workflow systems that can easily run hundreds of large scale jobs simultaneously [1,6], demands equally aggressive measures at metadata capture [5] to enable discovery, sharing, and reuse of the data. Without some form of automated metadata capture, however, either metadata description becomes largely a manual task, which is difficult if not impossible under high-volume conditions, or the searchability and manageability of the resulting data products is disappointingly low.

In this position paper we discuss our experiences with automated metadata generation drawn from five years of experience with e-Science workflows to execute dynamically adaptive regional weather forecast and analysis tasks on-demand in response to severe regional weather. The forecast and analysis workflows are executed with a BPEL based orchestration system running in the LEAD service oriented architecture (SOA) framework [3]. The nodes of the workflow are web services that “wrap” application tasks (e.g., models, analysis tools, etc.). LEAD is distinct in that the application tasks are often computationally intensive so are deployed remotely, specifically to the TeraGrid [7], where they run on parallel computing resources. Data collection is under control of the XMC Cat metadata catalog [8]. Our experience shows that *metadata that can be generated at the source, that is, either at the portal when the user sets up a workflow, or during workflow execution is sufficient to describe access, attribution, and distribution metadata but is insufficient for conducting all but the simplest data discovery.* Data discovery, or the ability to query for and find data products and collections after the fact, is important because data reuse is facilitated when discovery metadata is rich. Data preservation becomes easier as well.

## 2. Core metadata collection

The framework used to test different scenarios for metadata generation is illustrated in Figure 1. Through the portal, a scientist creates a BPEL script using a visual interface. The script is passed to a workflow engine for execution. Our tools capture high-level metadata about the workflow (e.g. center latitude and center longitude of the bounding box, and grid increments in both X and Y directions) by watching portal traffic. A workflow is composed of tasks (called service 1, service 2, etc. in the figure); tasks are wrapped in service wrappers that expose interfaces through WSDL definitions. The web service wrapper is instrumented to collect metadata about the data products consumed and produced by a particular service

---

<sup>1</sup> plale@cs.indiana.edu. Work funded in part through NSF Cooperative Agreement 0331480.

during workflow execution. This information is formatted using a domain specific XML schema and passed to the metadata collection tool. The schema used to describe data and collections as they are passed between services in the LEAD SOA is a profile of FGDC [4] for representing data granules and collections.

The initial design for metadata collection gathered metadata from two sources: experiment data collected at the portal and automated collection at the workflow task or node during runtime. In practice, the metadata we collected from this approach exhibited a number of deficiencies:

- *Inheritance of geospatial data* – the geospatial boundaries collected at the portal during experiment setup define the geospatial bounds of the full multi-step experiment. It was initially assumed that this geospatial information would be sufficient to describe the geospatial characteristics of the individual output data products of models such as the WRF forecast model. But models are complex and the initial geographic boundaries are often too coarse if for instance nested results are produced. Too, the output products are described by a special coordinate system. As a result, the coordinates defined during workflow construction are insufficient for file-level data discovery.

- *Minimal contextual information at workflow nodes* – in the LEAD SOA, application functionality is “wrapped” as a web service. During invocation of a service, some information about the workflow context is passed to the web service wrapper. However, the contextual information proved minimal. For instance, a service wrapping an atmospheric assimilation code may only see a directory generated, but lack contextual information for interpreting the directory’s contents.

- *Important search parameters buried in files* - Critical search parameters often reside as configuration parameters in opaque “containers” (i.e., files).

While the first approximation to metadata collection gave useful metadata, it took more aggressive forms of collection to extend the metadata to include support for discovery.

## 2.2 Enhanced metadata collection

To support richer discovery for data object reuse, we turned our attention to server side curation to enhance metadata collection. That is, we

designed into the metadata catalog an extensibility mechanism for the addition of asynchronous curation routines. The primary sources of additional metadata were in 1.) the self-describing binary format used by the atmospheric community (binary representations such as NetCDF and HDF are widely used in computationally oriented communities) and 2.) in model configuration files. The server side solution extends the XMC Cat catalog with plug-ins to process or “curate” the data products after they are registered to the XMC Cat catalog.

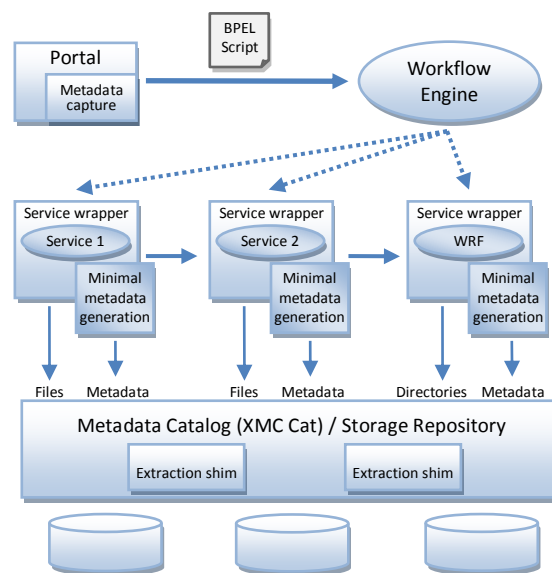


Figure 1. Points at which metadata generation and capture occur: at the portal, during workflow execution, and during ingestion at the metadata catalog.

One plug-in (called a “shim” [6]) extracts metadata from WRF output files by examining the file type to determine if it is a NetCDF file and the file name to identify whether or not it is the product of a WRF model run. The shim contacts an OPeNDAP [2] server and retrieves the NetCDF header through a NetCDF Java API. The retrieved metadata includes geospatial bounds, grid spacing, initialization time, start date, end date, and initialization offset for hourly files. The shim also attempts to retrieve the file offset (e.g., “hour 4 of a 36 hour forecast”) by first examining the experiment to which the file belongs to see if the experiment metadata contains the “history interval” and “frames per outfile” parameters. If so, then the file’s offset can be calculated. Additional shims have been created to extract task (i.e., model, analysis tool) configuration parameters from Fortran namelist

files. These attributes are stored at the workflow level, and contribute significantly to the discovery metadata for that workflow run.

### 3. Evaluation

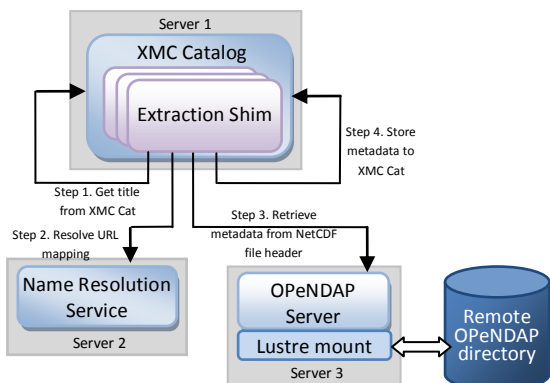


Figure 2. Steps in curation of WRF output files. Calls outside XMC Catalog are synchronous.

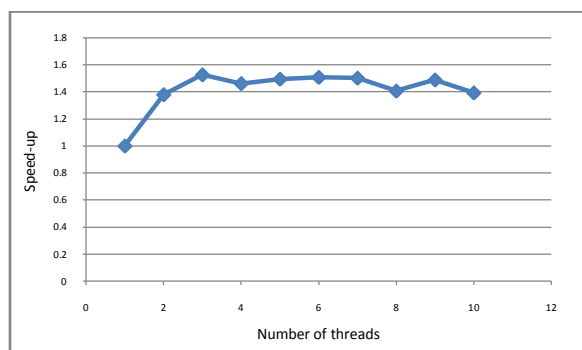


Figure 3. Initial parallelization of the curation plug-in for WRF files shows room for optimization, with the likely cause shown in Fig. 4.

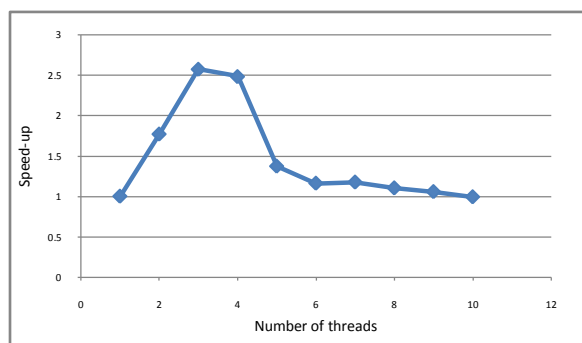


Figure 4. Contention occurs at a relatively low number of threads due in part to the remote location of the OPeNDAP directory.

We measured speed-up by post processing 200 WRF output files. The four-step process includes contacting a name resolution service to resolve a logical name, and contacting a remote

OPeNDAP server to extract the metadata from the NetCDF header (see Fig. 2). While the speed-up results shown in Fig. 3 indicate room for significant improvement, Fig. 4 points to the likely point of contention at the remote OPeNDAP server. We have not yet distinguished latency due to network delay from latency at the server. This is ongoing work.

### 4. Conclusion

Research to-date indicates that discovery metadata can be substantially strengthened through automated metadata generation. But these gains come at the cost of embedding domain-specific routines at the server. Ongoing research examines additional kinds of metadata that can be collected automatically and examines the generalizability of these results to other workflow systems.

### Acknowledgements

We thank Scott Jensen, author of XMC Cat (<http://www.cs.indiana.edu/~scjensen/catalog/catalog.html>) and Suresh Marru, both of IU.

### References

- [1] Altintas, I., C. Berkely, et al, Kepler, an extensible system for design and execution of scientific workflows. In *SSDBM*, pages 423-424, 2004.
- [2] Cornillon, P., Gallagher, J., and Sgourosy, T. OPeNDAP: accessing data in a distributed, heterogeneous environment. *Data Science Journal*, 2(5), 164-174, Nov. 2003.
- [3] Droegemeier, K., K. Brewseter, et al., Service-oriented environments for dynamically interacting with mesoscale weather, In *Computing in Science and Engineering*, IEEE Computer Society Press, Vol. 7, No. 6, pp. 12-29, 2005.
- [4] Federal Geographic Data Committee, <http://www.fgdc.gov/metadata/csdlgm>
- [5] Gray, J., D.T. Liu, et al, Scientific data management in the coming decade, In *Technical Report, MSR-TR-2005-10*, Microsoft Research, July 2005.
- [6] Oinn, T, M. Addis et al., Taverna: a tool for the composition and enactment of bioinformatics workflows, In *Bioinformatics*, 20(17): 3045-3054, Oxford University Press, London, UK, 2004.
- [7] Reed, D.A., Grids, the teragrid, and beyond, *Computer*, 36(1), 62-68, 2003.
- [8] Sun, Y., S. Jensen, S.L. Pallickara, and B. Plale, Personal workspace for large-scale data-driven computational experimentation, In *7th IEEE/ACM Int'l conf. on grid computing (Grid'06)*, Barcelona, September 2006.