XMC Cat: An Adaptive Catalog for Scientific Metadata

Scott Jensen and Beth Plale

School of Informatics and Computing
Indiana University-Bloomington

Enabling the Discovery and Reuse of Scientific Data

Researchers have noted the deluge of scientific data being generated and the need for detailed discovery metadata to find relevant data. Additionally, as noted in a study by the UK e-Science Core Programme, “metadata is key to being able to share results.” This need for detailed metadata has lead scientific communities to develop XML schemes to describe their data products.

One focus of our research is on the characteristics of scientific metadata schemata – with an important characteristic being that they are composed of complex concepts that describe a data object. Exploring this characteristic allows the XMC Cat metadata catalog to communicate using the schema of the community in which it is implemented, while storing metadata using a generalized relational schema that can be applied across domains and is loosely coupled to the domain schema.

This approach is possible due to a partitioning of the schema based on its metadata concepts, which allows for a static global ordering of the concept elements. Paired with a hybrid XML-Relational approach to shredding and storing metadata, XMC Cat provides a scalable solution with fast query response rates.

The XMC Cat metadata catalog was initially developed in the context of the Linked Environments for Atmospheric Discovery (LEAD) project and is deployed as a web service using a lightweight software stack based on Apache Tomcat, Axis2, and MySQL.

Features & Benefits of XMC Cat

- Identify metadata concepts and create configuration files using a point-and-click web interface.
- XMC Cat adapts to the scientific community instead of scientists adapting to it.
- Searching is performed using a point-and-click web interface that walks the scientist through creating a query.
- Search definitions can be saved and shared – allowing scientist to share searches for experiment configurations or data known to create model instability or anomalous results – before they publish.
- Metadata capture can be automated to harvest additional metadata from data objects as they are added to the metadata catalog by registering post-processing plugins.
- A scientist’s data remains in their private workspace until they decide to publish it to the community.
- Metadata is easily and efficiently added incrementally – allowing scientists to monitor long-running experiments.

Initially developed for atmospheric metadata, we are now implementing XMC Cat in other scientific communities. We are exploring using XMC Cat to capture astronomical metadata as well as metadata from other scientific domains. Science gateways often use a profile or extension of an existing metadata standard, so we are working on providing configurations based on publicly available domain metadata schemata.

Although the focus of XMC Cat has been on managing metadata describing scientific data captured as files – particularly in binary formats, the Data to Insight Center is exploring using XMC Cat to assist scientists with managing legacy research data. In support of a long-term collaborative research network with decades of data on forest management and ecological data, we are working with researchers to implement XMC Cat as a metadata search tool over legacy databases containing decades of measurements and survey results.

This research was funded in part by NSF CNS0720580, ATM-0331480 and EIA-0202048.

Photo credit: NOAO/AURA/NSF. Copyright WIYN Consortium, Inc., all rights reserved.