IN43A-0226

DataFinder: Using Ontologies and Reasoning to Enhance Metadata Search

Thomas Russ and Hans Chalupsky
University of Southern California, Information Sciences Institute
4676 Admiralty Way, Marina del Rey, CA 90292
tar@isi.edu, hans@isi.edu

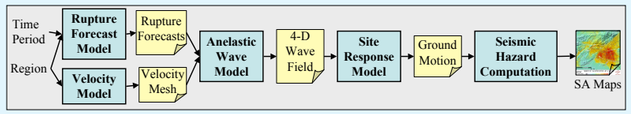
Computational Workflows and the Problem of Finding Data Files

DataFinder is part of the Southern California Earthquake Center (SCEC) Community Modeling Environment's suite of tools for generating computational workflows and locating the resulting data products. A workflow is a specific series of computations that produce data files based on simulation models. A given workflow involves running various software modules which pass data via often large data files.

When a workflow is being instantiated, a decision needs to be made whether to use an existing data file or to generate the required file by running a computational model. Since it is often more efficient to use existing data, locating previously computed results can produce a better workflow.

The files produced by SCEC computational models have descriptive metadata stored as pairs of attribute names and values. These attributes describe the content and provenance of the files and can be used to identify files of interest. But depending on which software was used to prepare the file, different attribute names and different organizational schemes are used for the metadata. The low-level nature of the metadata descriptions and the variety of metadata schemata present challenges for finding files needed for workflows which DataFinder addresses.

In addition to its part in workflow instantiation process, DataFinder also provides a way to locate end data products for users. Currently in proof-of-concept form for locating map files, such an extension of the current web-based interface to DataFinder will make them accessible to a wider range of users.



A sample SCEC CME workflow process. This is part of a "Pathway 2" computation which uses wave propagation models to compute the effects of an earthquake rupture on the ground and structures. (In contrast "Pathway 1" computations which use empirically-derived intensity measure relationships) DataFinder's role in workflow instantiation is to locate already existing data files which would otherwise need to be computed.

The Southern California Earthquake Center's Community Modeling Environment uses computer codes for simulation and hazard analysis computations. The process of running workflows using several computational models produces numerous intermediate and final data files. These files have descriptive metadata stored as pairs of attribute names and values. Depending on which software was used to prepare the files, different attribute names and different organizational schemes are used for the metadata. Previous search tools for this metadata repository rely on the user knowing the structure and names of the metadata attributes in order to find stored information. Matters are made even harder because sometimes the type of information in a data file must be inferred. For example, seismic hazard maps are described simply as "JPEGFile", with the domain content of the file inferable only by looking at the workflow that produced the file. This greatly limits the ability to actually find data of interest.

DataFinder translates the domain concepts specified by the user into the low-level SCEC CME metadata attribute names. Computational pathway builders can identify the logical file names of the files needed to execute a workflow that solves their geophysical problem.

DataFinder also allows users to locate end data products using domain-level terms instead of program-specific and varied metadata attributes. The domain-level terms can also include hierarchies not present in the metadata itself, but rather *inferred* from other attributes.

Why Not Just Extend the Metadata?
The model type information could just be added to the file metadata, but our approach is more flexible. It is easier to have a deeper hierarchy of classes, our approach can be applied to existing metadata stores, and it is easy to add new distinctions as the need arises. Handling multiple schemata is also easier.

DataFinder: Semantic Enhancement with Ontologies

DataFinder uses ontologies to provide a semantic overlay for the metadata attributes that are used to index data files. A domain ontology is combined with a metadata attribute ontology to link geophysical and seismic hazard domain concepts with the metadata attributes that describe the computational products. DataFinder uses a domain ontology and additional rules expressed in first-order logic to provide this semantic enhancement. The domain and metadata attribute ontology is represented in the PowerLoom representation language. DataFinder is implemented using a hybrid reasoning approach based on combining the strengths of the PowerLoom logical reasoning engine with the database technology underlying the metadata repository to provide scalability.

The PowerLoom reasoning engine allows adding semantic enhancements by overlaying the raw metadata with a hierarchy of concepts, providing more abstract views of the data collection. For example, a velocity mesh is one of the intermediate data products used in seismic wave propagation simulations. It is the output of a 3-dimensional wave propagation velocity model of the subsurface geology. This mesh can be computed using any one of several models which have different characteristics. Each velocity mesh has metadata that records the specific model used. But the models also fall into general classes such as 1-dimensional and 3-dimensional models. Such classification of models can be used to allow retrieval of velocity meshes created by any 3-dimensional model, without the user being required to specify all of the particular model names. This provides an abstraction over the attributes and values stored with the dataset. Other mappings are used to present a uniform interface to metadata information that is stored using different attribute names or even different organizational schemes. This abstraction layer gives DataFinder the ability to allow users to locate end data products using domain-level descriptions instead of program-specific and varied metadata attributes.

DataFinder uses ontologies to provide a semantic overlay that enriches the raw metadata. A simple example is the class and instance structure that describes velocity models. This structure lets a model type query work even though type information is not present in the metadata attributes itself. It is *inferred* based on the relationships in the ontology.

Velocity Model Classes and Instances

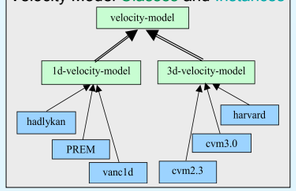
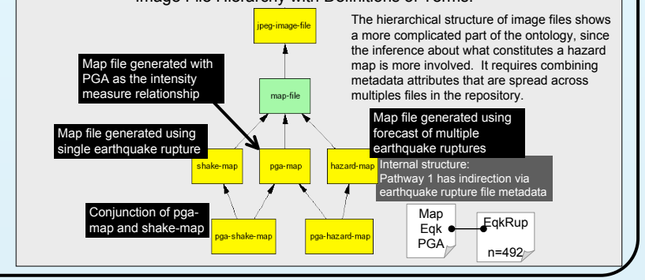


Image File Hierarchy with Definitions of Terms.



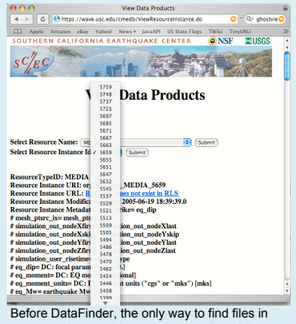
The hierarchical structure of image files shows a more complicated part of the ontology, since the inference about what constitutes a hazard map is more involved. It requires combining metadata attributes that are spread across multiples files in the repository.

Available Metadata and Current Approach

It is not easy to find data files that match the requirements of a given workflow. The existing SCEC CME metadata browsing tools were not up to this task. They present the metadata on a per-file basis which just does not scale. The example of searching for a velocity mesh covering a particular region shown at right illustrates the problems with a browsing approach. In addition to the browsing problems, there are additional shortcomings of using low-level attribute-value pairs:

- **Doesn't use useful abstractions, e.g.**
  - Only specific models, not model types
  - The mesh spacing is specified separately in all three dimensions in the metadata
  - Individual components of the region definition are not aggregated
- **Lack of regional comparisons**
  - Individual components must be found, combined and compared separately
- **Alternate metadata schema used by different computational pathways, e.g:**

Pathway 1 Seismic Hazard Computations use a different organizational approach to the metadata. Instead of copying all values forward to derived products, they use pointers to the input file metadata.

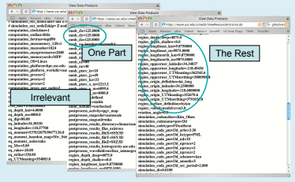


Before DataFinder, the only way to find files in the repository was through an interface that allowed browsing the metadata associated with individual files.

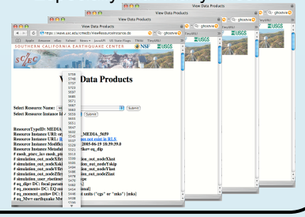
The metadata information is stored in the computational GRID's Metadata Catalog System (MCS). It has a limited search function and comparison function, but it still requires working at the low-level attribute-name value level. The MCS doesn't have as expressive a query mechanism as DataFinder's PowerLoom engine, and it doesn't support a rich a modeling language.

Example: Find a velocity mesh with
Grid spacing = 125m
In region bounded by
34°N, 118°W and 36°N, 117°W

Manual Search is No Solution



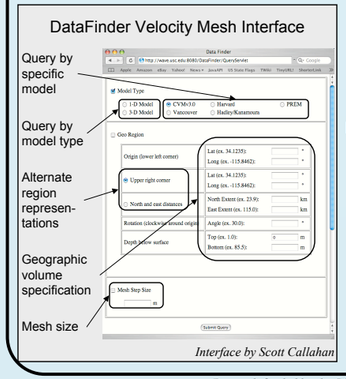
Especially with Many Files…



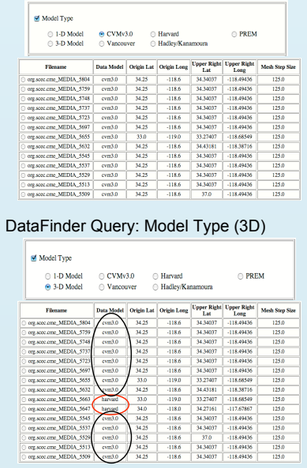
DataFinder Interface and Examples

This interface was created to show the capabilities of DataFinder in locating velocity mesh data files. The interface allows querying by type or model as well as individual models, and abstracts multi-axial gird spacing parameters into a single value.
Geographic regions can be specified using one of two representations. Containment queries for geographic regions are also implemented.

DataFinder Velocity Mesh Interface



DataFinder Query: Single Model (CVM3.0)



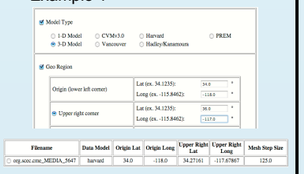
Filename	Data Model	Origin Lat	Origin Long	Upper Right Lat	Upper Right Long	Mesh Step Size
org.scec.cme_MEDIA_5804	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5739	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5748	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5737	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5723	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5697	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5655	cvm3.0	33.0	-119.0	33.27407	-118.68549	125.0
org.scec.cme_MEDIA_5632	cvm3.0	34.25	-118.6	34.43181	-118.38715	125.0
org.scec.cme_MEDIA_5545	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5537	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5529	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5513	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5509	cvm3.0	34.25	-118.6	37.0	-118.49436	125.0

DataFinder Query: Model Type (3D)

Filename	Data Model	Origin Lat	Origin Long	Upper Right Lat	Upper Right Long	Mesh Step Size
org.scec.cme_MEDIA_5804	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5739	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5748	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5737	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5723	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5697	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5655	cvm3.0	33.0	-119.0	33.27407	-118.68549	125.0
org.scec.cme_MEDIA_5632	cvm3.0	34.25	-118.6	34.43181	-118.38715	125.0
org.scec.cme_MEDIA_5682	harvard	33.0	-119.0	33.27407	-118.68549	125.0
org.scec.cme_MEDIA_5645	harvard	34.0	-118.0	34.27161	-117.67867	125.0
org.scec.cme_MEDIA_5545	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5537	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5529	cvm3.0	34.25	-118.6	37.0	-118.49436	125.0
org.scec.cme_MEDIA_5513	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5509	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0

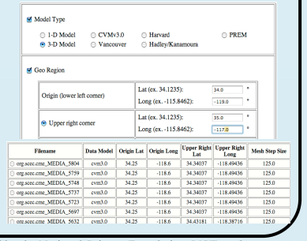
DataFinder Queries with Regional Bounds Added

Example 1



Filename	Data Model	Origin Lat	Origin Long	Upper Right Lat	Upper Right Long	Mesh Step Size
org.scec.cme_MEDIA_5647	harvard	34.0	-118.0	34.27161	-117.67867	125.0

Example 2



Filename	Data Model	Origin Lat	Origin Long	Upper Right Lat	Upper Right Long	Mesh Step Size
org.scec.cme_MEDIA_5804	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5739	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5748	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5737	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5723	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5697	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5632	cvm3.0	34.25	-118.6	34.43181	-118.38716	125.0

2005 AGU Fall Meeting

Research funded by the SCEC Community Modeling Environment Project which is funded by the National Science Foundation (NSF) under contract EAR-0122464 (The SCEC Community Modeling Environment (SCEC/CME): An Information Infrastructure for System-Level Earthquake Research)