DataFinder: Semantically Informed Search in Metadata Repositories

Thomas Russ and Hans Chalupsky
University of Southern California Information Sciences Institute

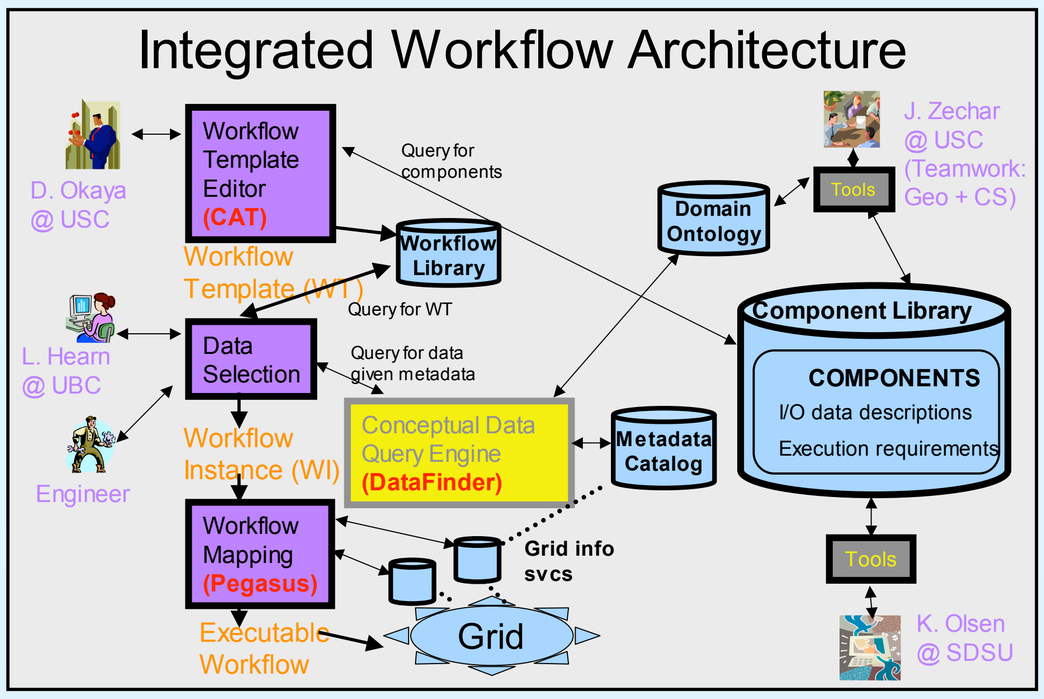
Computational Workflows and the Problem of Finding Data Files

DataFinder is part of the 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 will involve running various software modules which pass data via often large data files.

When a workflow is being instantiated, a decision needs to be made about whether to use an existing data file or to generate the required file by running computational models. 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 files, 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 the workflow instantiation process, DataFinder can also provide a way to locate end data products for users. This capability currently exists in proof-of-concept form for locating map files. Developing a web-based interface to DataFinder for such end products will make them accessible to a wider range of users.



Architecture of the SCEC CME integrated workflow process. Workflow templates, which describe an abstract workflow, are instantiated to produce a concrete plan which can be executed on the computational grid. DataFinder's role in workflow instantiation is to locate already existing data files which would otherwise need to be computed.

Problem with Current Approach

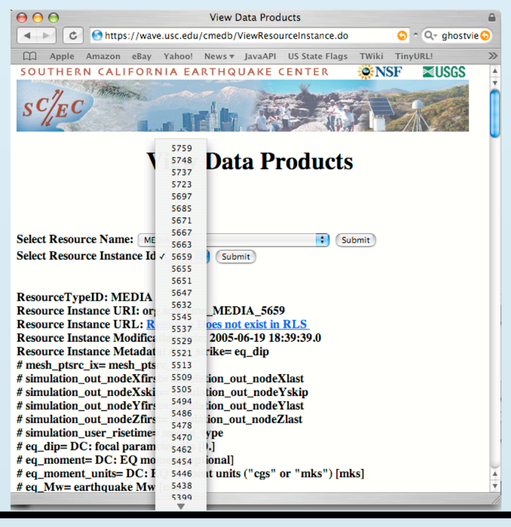
It is not easy to find data files that match the requirements of a given workflow. The existing SCEC CME metadata browsing tools were not up to this task. They present the metadata on a per-file basis which just does not scale.

The example of searching for a velocity mesh covering a particular region shown at right illustrates the problems with a browsing approach.

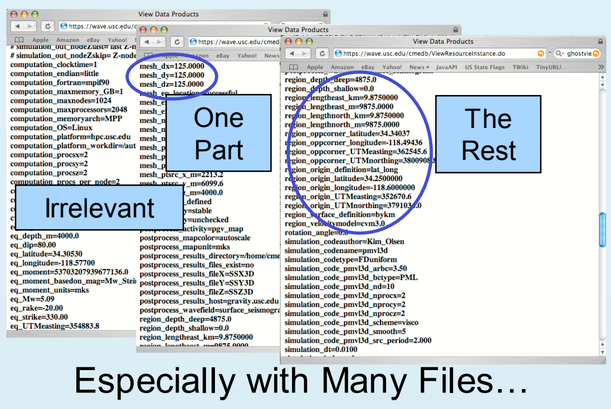
The underlying Metadata Catolog System has a limited search function, but it still requires working at the low-level attribute-name value level and it doesn't have as expressive a query mechanism as the DataFinder's interface.

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

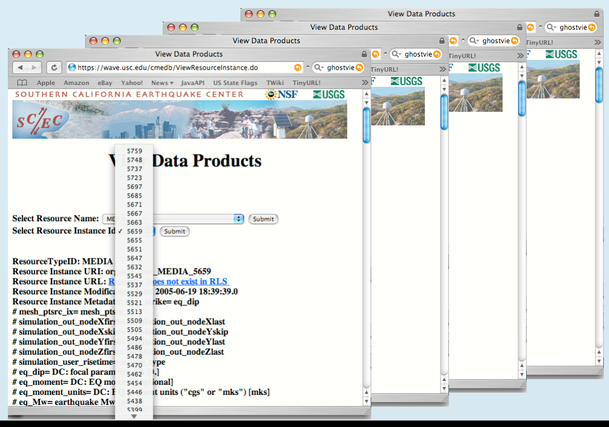
Before DataFinder, the only way to find files in the repository was through an interface that allowed browsing the metadata associated with individual files.



Manual Search is No Solution

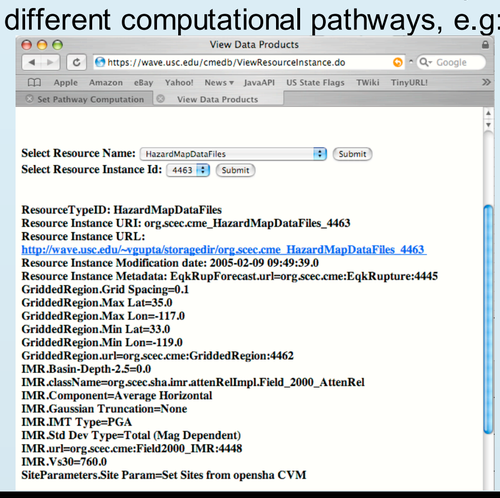


Especially with Many Files…



Additional Problems

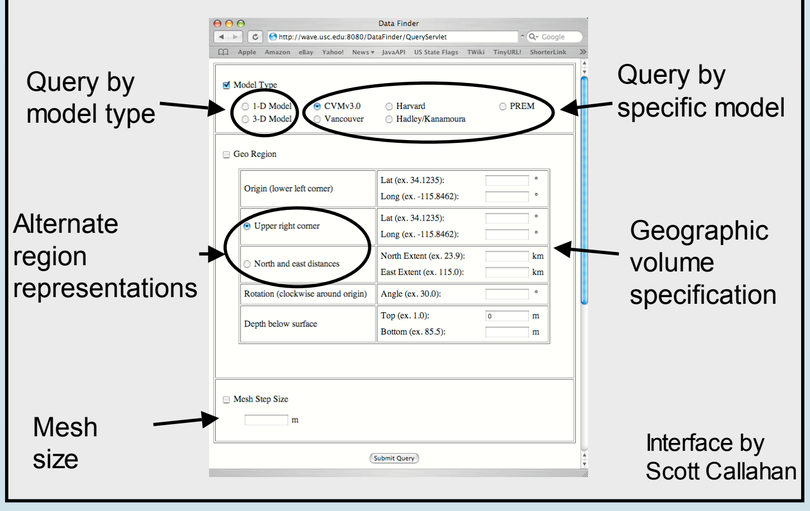
- Desn't use useful abstractions, e.g.
  - Only specific models, not model types
  - The mesh spacing is specified separately in all three dimensions in the metadata
  - Individual components of the region definition are not aggregated
- Lack of regional comparisons
  - Individual components must be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



DataFinder: Adding Semantic Search

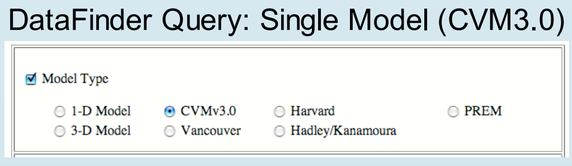
DataFinder translates the domain concepts specified by the user into the low-level SCEC CME metadata attribute names. Users can identify the logical file names of the files needed to execute a workflow that solves their geophysical problem. DataFinder also allows users to locate end data products using domain-level terms instead of program-specific and varied metadata attributes.

DataFinder Velocity Mesh Interface



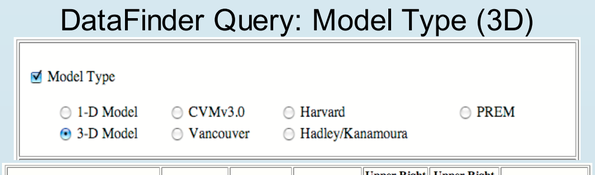
This interface was created to show the capabilities of DataFinder in locating velocity mesh data files. The interface allows querying by type or model as well as individual models, and abstracts multiaxial gird spacing parameters into a single value. It also provides containment queries for geographic regions.

DataFinder Query: Single Model (CVM3.0)



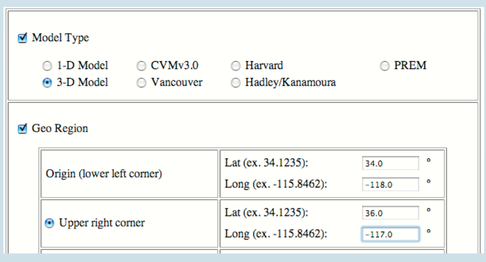
Filename	Data Model	Origin Lat	Origin Long	Upper Right Lat	Upper Right Long	Mesh Step Size
org.scec.cme_MEDIA_5804	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5759	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5748	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
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	-118.38716	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)

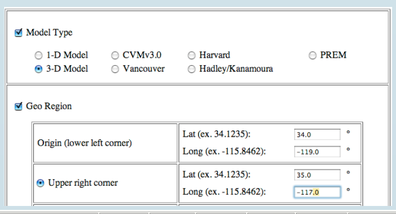


Filename	Data Model	Origin Lat	Origin Long	Upper Right Lat	Upper Right Long	Mesh Step Size
org.scec.cme_MEDIA_5804	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5759	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5748	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
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	-118.38716	125.0
org.scec.cme_MEDIA_5663	harvard	33.0	-119.0	33.27407	-118.68549	125.0
org.scec.cme_MEDIA_5647	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



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



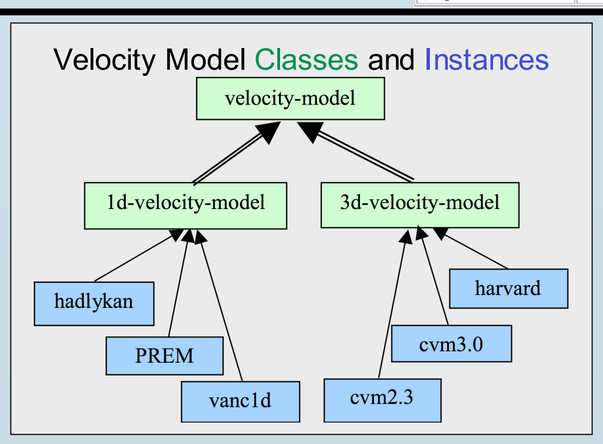
Filename	Data Model	Origin Lat	Origin Long	Upper Right Lat	Upper Right Long	Mesh Step Size
org.scec.cme_MEDIA_5804	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5759	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
org.scec.cme_MEDIA_5748	cvm3.0	34.25	-118.6	34.34037	-118.49436	125.0
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

Semantic Enhancement with Ontologies

DataFinder uses ontologies, or formal definitions of terms to provide a semantic overlay that enriches the structure of the raw metadata. One simple example is the classes and instances describing velocity models. This structure is what enables the model type query to work even though the model type information is not present in the metadata attributes itself. It is *inferred* based on the relationships captured in the ontology.

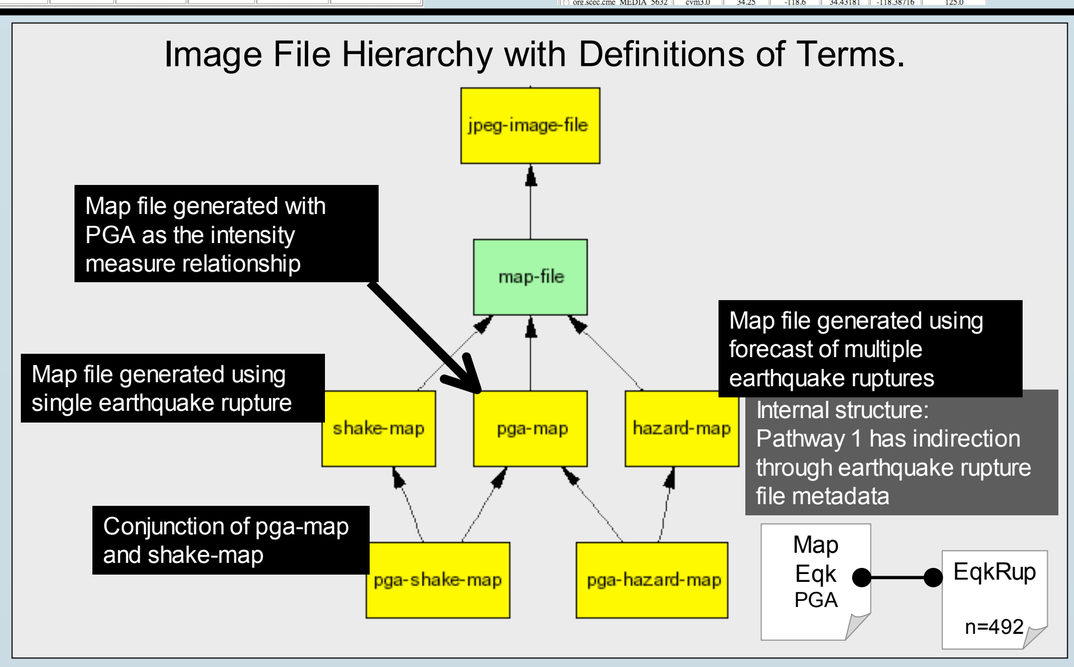
The hierarchical structure of image files shows a more complicated part of the ontology, since the inference about what constitutes a hazard map is more involved. It requires combining metadata attributes that are spread across multiples files in the repository. By capturing that information in the ontology, a user looking for data products is relieved of the need to know such details.

PowerLoom maps such terms onto the underlying attributes using its efficient database interface layer to provide a scalable and convenient solution.



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's PowerLoom reasoning engine uses the definitions and their underlying links to metadata attributes to map queries using domain terms to queries against the underlying metadata repository. Hazard maps are found by looking for map files generated using multiple earthquake forecasts. For pathway 1, this means the map file has a link to an earthquake rupture forecast file containing more than one forecast.