DataFinder: Semantically Informed Search in Metadata Repositories

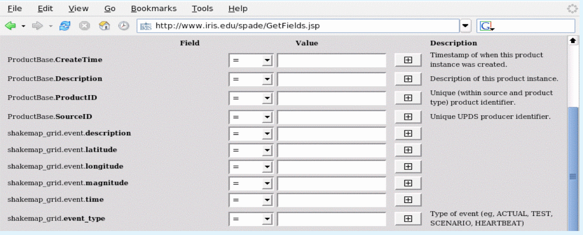
Thomas Russ and Hans Chalupsky
University of Southern California Information Sciences Institute

The Data Finding Problem

- Pathway computations generate large number of data products (stored in digital libraries)
- If there is metadata, it is complex, low-level, non-uniform, and constantly evolving
- Finding results or reusing previously computed products is hard
  - Very large space of potentially relevant files
  - Which metadata attributes describe what I want?
  - Complex search interfaces
- Example query
  - Find PGA hazard maps within rectangular region bounded by 34°N, 122°W and 37°N ,118°W
  - Looks simple
    - Go to MySRB or other metadata search interface
    - Specify metadata attribute constraints
    - Get results
  - But how do you know which attributes to use?
  - How do you phrase the query?
    - Attributes are different for pathway 1 and 2
    - Type information is implicit
    - Some joins are complex

```
SELECT ?mapfile  WHERE …
 AND file_logical_type = "JPEGFile"
AND   IMR.IMT Type = “PGA”
AND   EqkRupForecast.url = ?url
AND THERE EXISTS ?forecast_file
WHERE file_logical_name = ?url
AND   EqkRupForecast.NumRuptures > 1
```

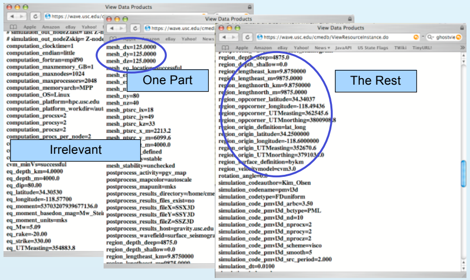
Standard Attribute-based Search



http://www.iris.edu/spade/GetFields.jsp

Field	Value			Description
ProductBase.**CreateTime**	=		⊞	Timestamp of when this product instance was created.
ProductBase.**Description**	=		⊞	Description of this product instance.
ProductBase.**ProductID**	=		⊞	Unique (within source and product type) product identifier.
ProductBase.**SourceID**	=		⊞	Unique UPDS producer identifier.
shakemap_grid.event.**description**	=		⊞	
shakemap_grid.event.**latitude**	=		⊞	
shakemap_grid.event.**longitude**	=		⊞	
shakemap_grid.event.**magnitude**	=		⊞	
shakemap_grid.event.**time**	=		⊞	
shakemap_grid.**event_type**	=		⊞	Type of event (eg, ACTUAL, TEST, SCENARIO, HEARTBEAT)

But Which Attributes are Relevant?



https://wave.usc.edu/cmedb/ViewResourceInstance.do

```
# simulation_out_nodeZskip= Z-node
computation_clocktime=1
computation_endian=little
computation_fortran=mpif90
computation_maxmemory_GB=1
computation_maxnodes=1024
computation_maxprocessors=2048
computation_memoryarch=MPP
computation_OS=Linux
computation_platform=hpc.usc.edu
computation_platform_workdir=/aut
computation_procsx=2
computation_procsy=2
computation_procsz=2
computation_procs_per_node=2
cvm_minVs=successful
eq_depth_km=4.0000
eq_depth_m=4000.0
eq_dip=80.00
eq_latitude=34.30530
eq_longitude=-118.57700
eq_moment=53703207939677136.0
eq_moment_basedon_mag=Mw_Stei
eq_moment_units=mks
eq_Mw=5.09
eq_rake=-20.00
eq_strike=330.00
eq_UTMeasting=354883.8
```

```
mesh_dx=125.0000
mesh_dy=125.0000
mesh_dz=125.0000
mesh_ny=80
mesh_nz=40
mesh_ptsrc_ix=18
mesh_ptsrc_jy=49
mesh_ptsrc_kz=33
mesh_ptsrc_x_m=2213.2
mesh_ptsrc_y_m=6099.6
mesh_stability=unchecked
postprocess_activity=pgv_map
postprocess_mapcolor=autoscale
postprocess_mapunit=mks
postprocess_results_directory=/home/cme
postprocess_results_files_exist=no
postprocess_results_fileX=SSX3D
postprocess_results_fileY=SSY3D
postprocess_results_fileZ=SSZ3D
postprocess_results_host=gravity.usc.edu
postprocess_wavefield=surface_seismogra
region_depth_deep=4875.0
region_depth_shallow=0.0
region_lengtheast_km=9.8750000
```

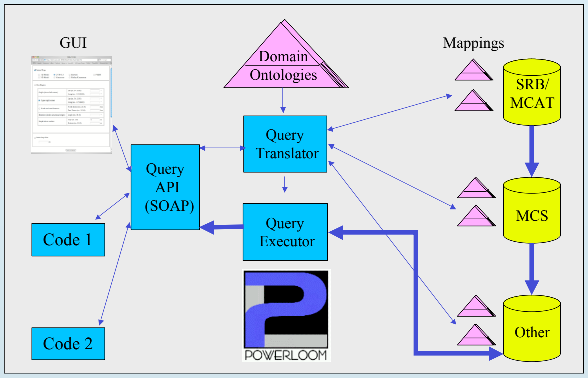
```
region_depth_deep=4875.0
region_depth_shallow=0.0
region_lengtheast_km=9.8750000
region_lengtheast_m=9875.0000
region_lengthnorth_km=9.8750000
region_lengthnorth_m=9875.0000
region_oppcorner_latitude=34.34037
region_oppcorner_longitude=-118.49436
region_oppcorner_UTMeasting=362545.6
region_oppcorner_UTMnorthing=3800908.8
region_origin_definition=lat_long
region_origin_latitude=34.2500000
region_origin_longitude=-118.6000000
region_origin_UTMeasting=352670.6
region_origin_UTMnorthing=3791030.0
region_surface_definition=bykm
region_velocitymodel=cvm3.0
rotation_angle=0.0
simulation_codeauthor=Kim_Olsen
simulation_codename=pmvl3d
simulation_codetype=FDuniform
simulation_code_pmvl3d_arbc=3.50
simulation_code_pmvl3d_bctype=PML
simulation_code_pmvl3d_nd=10
simulation_code_pmvl3d_nprocx=2
simulation_code_pmvl3d_nprocy=2
simulation_code_pmvl3d_nprocz=2
simulation_code_pmvl3d_scheme=visco
simulation_code_pmvl3d_smooth=5
simulation_code_pmvl3d_src_period=2.000
simulation_dt=0.0100
```

And how does one know which ones they are?

The DataFinder Solution

- Semantically informed search tool
  - Semantic search based on semantic description of data products
  - Ontologies & rules define meaning of relevant terms in a domain
    - e.g., PGA map, Hazard map, multi-event rupture, etc.
  - Mappings define meaning of metadata attributes in domain terms
  - Query translator rewrites domain queries into metadata attribute queries
  - Query executor runs translated queries against repositories and combines results
- Query translation
  - Semantic query cannot be run directly on repositories
    - MCS, SRB, etc. only understand attribute/value queries
  - Must translate semantic language to low-level attribute language
    - using ontologies, mappings and logical inference
    - potentially creating multiple queries
    - well-researched problem in data integration
- New query translation algorithm
  - allows **very expressive domain language** and rules
  - results in very **compact conjunctive queries**
  - offloads “heavy lifting” to external repository
  - “patent almost pending”

DataFinder Architecture



Translation of a domain query into an attribute-value query (MCS)

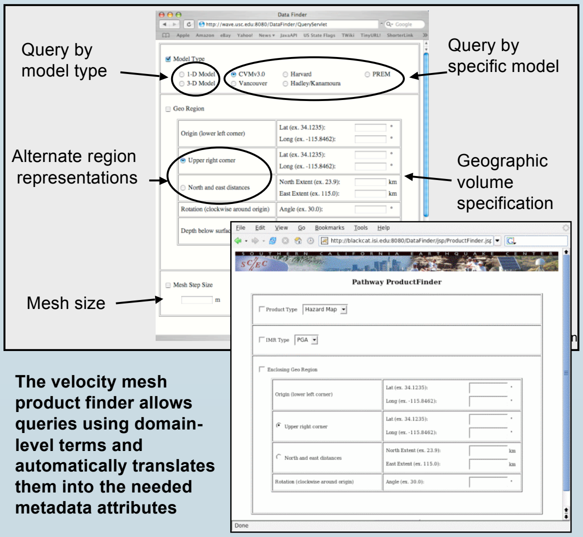
FIND *?file*
WHERE *?file* is PGA-Hazard-Map
AND *?file* describes *?region*
AND *?region* is enclosed by
 bounding box 34, -122, 37, -118

```
SELECT DISTINCT _T1_.Object_id,
  _T1_.Attribute_value,_T2_.Attribute_value, _T3_.Attribute_value,
  _T4_.Attribute_value,_T5_.Attribute_value, _T6_.Attribute_value,
  _T10_.Attribute_value,_T11_.Logical_name, _T11_.Data_type
  FROM mcs_file_string_attributes _T1_, mcs_file_string_attributes
  _T2_, mcs_file_string_attributes _T3_, mcs_file_string_attributes
  _T4_, mcs_file_string_attributes _T5_, mcs_file_string_attributes
  _T6_, mcs_file_string_attributes _T7_, mcs_file_string_attributes
  _T8_, mcs_file_string_attributes _T9_, mcs_file_string_attributes
  _T10_, mcs_logical_file _T11_ WHERE 1=1 AND
  _T1_.Object_id=_T2_.Object_id AND
  _T1_.Object_id=_T3_.Object_id AND …
```

DataFinder Features

- Semantic querying
  - Find data & products based on their meaning, not low-level features
  - Containment reasoning (subsumption) for types and regions
- Aggregation of metadata distributed over multiple files
  - combine metadata from multiple objects
  - “pathway-1-style” vs. “pathway-2-style”
- Transparent support of different metadata schemata NEW
  - different metadata used by different researchers/codes
  - different metadata standards (SCEC vs. FGDC)
- Scalability
  - use query rewriting to translate domain level queries NEW
  - push conjunctive queries and constraints to MCS, SRB backend NEW
- Extensibility without recoding
  - add new abstractions to the ontology
  - add new attributes and mappings without affecting existing ones NEW
- Integrated with MCS **and SRB** NEW
  - provides **semantic layer on top of Metadata Catalog**
  - leverages PowerLoom^{TM} inference & RDBMS interface
  - query multiple repositories simultaneously NEW
- Generic technology—widely applicable
  - Example: Velocity meshes, and **Pathway-1 products** NEW
  - Other domains

Velocity Mesh and Product Finder Interfaces



http://wave.usc.edu:8080/DataFinder/QueryServlet

☑ Model Type
○ 1-D Model ○ 3-D Model
● CVMv3.0 ○ Harvard ○ PREM
○ Vancouver ○ Hadley/Kanamoura

☐ Geo Region

Origin (lower left corner)	Lat (ex. 34.1235):
	Long (ex. -115.8462):
		°
		°

● Upper right corner	Lat (ex. 34.1235):
	Long (ex. -115.8462):
		°
		°

○ North and east distances	North Extent (ex. 23.9):
	East Extent (ex. 115.0):
		km
		km

Rotation (clockwise around origin)	Angle (ex. 30.0):	°
Depth below surfac		

☐ Mesh Step Size
m

http://blackcat.isi.edu:8080/DataFinder/jsp/ProductFinder.jsp

Pathway ProductFinder

☐ Product Type Hazard Map

☐ IMR Type PGA

☐ Enclosing Geo Region

Origin (lower left corner)	Lat (ex. 34.1235):
	Long (ex. -115.8462):
		°
		°

● Upper right corner	Lat (ex. 34.1235):
	Long (ex. -115.8462):
		°
		°

○ North and east distances	North Extent (ex. 23.9):
	East Extent (ex. 115.0):
		km
		km

Rotation (clockwise around origin)	Angle (ex. 30.0):	°

The velocity mesh product finder allows queries using domain-level terms and automatically translates them into the needed metadata attributes