DataFinder: Semantically Informed Search in Metadata Repositories

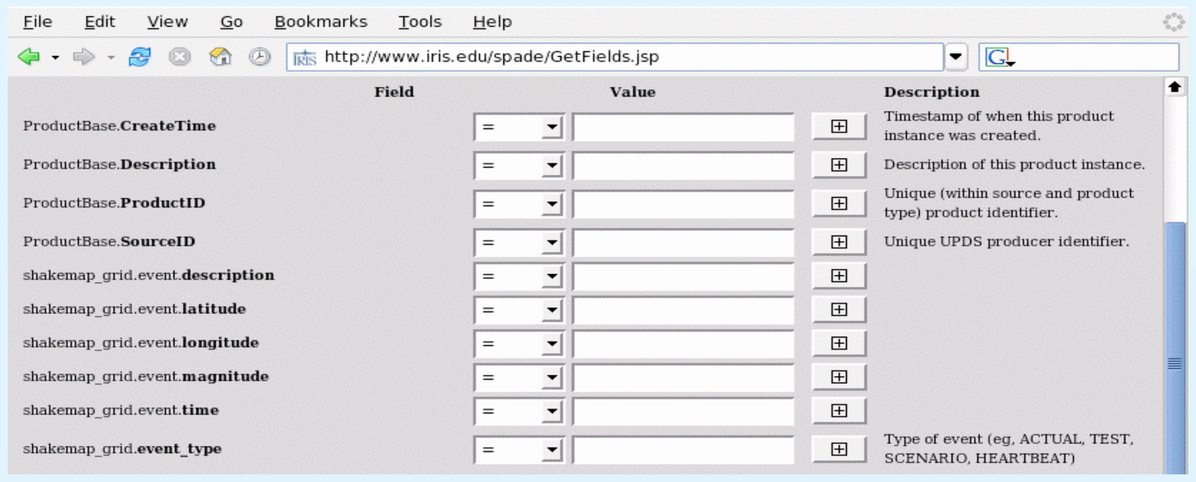
Thomas Russ and Hans Chalupsky
University of Southern California Information Sciences Institute

The Data Finding Problem

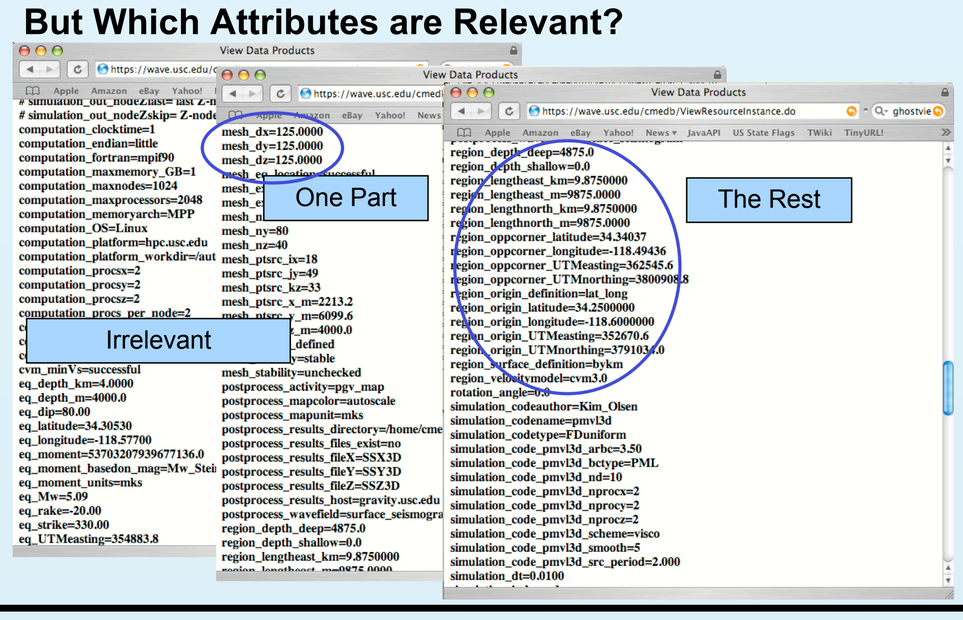
- Pathway computations generate large number of data products (stored in digital libraries)
- If there is metadata, it is complex, low-level, non-uniform, and constantly evolving
- Finding results or reusing previously computed products is hard
  - Very large space of potentially relevant files
  - Which 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



But Which Attributes are Relevant?

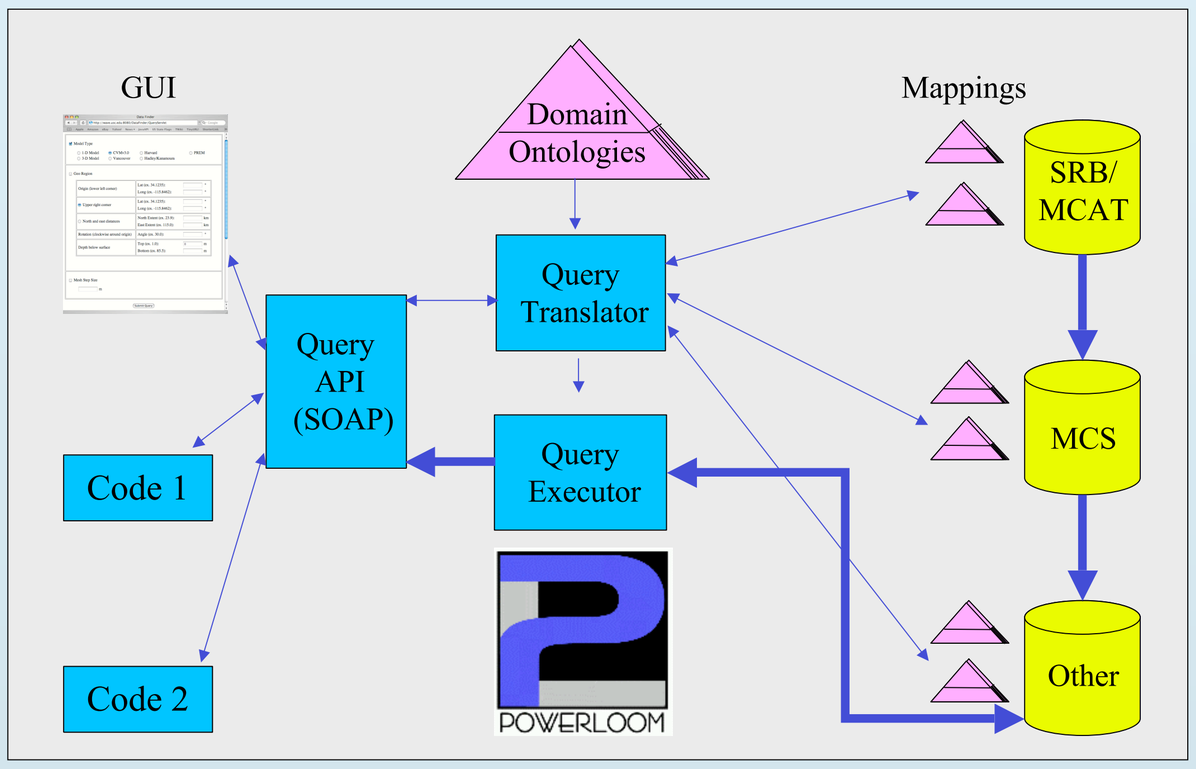


And how does one know which ones they are?

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
```

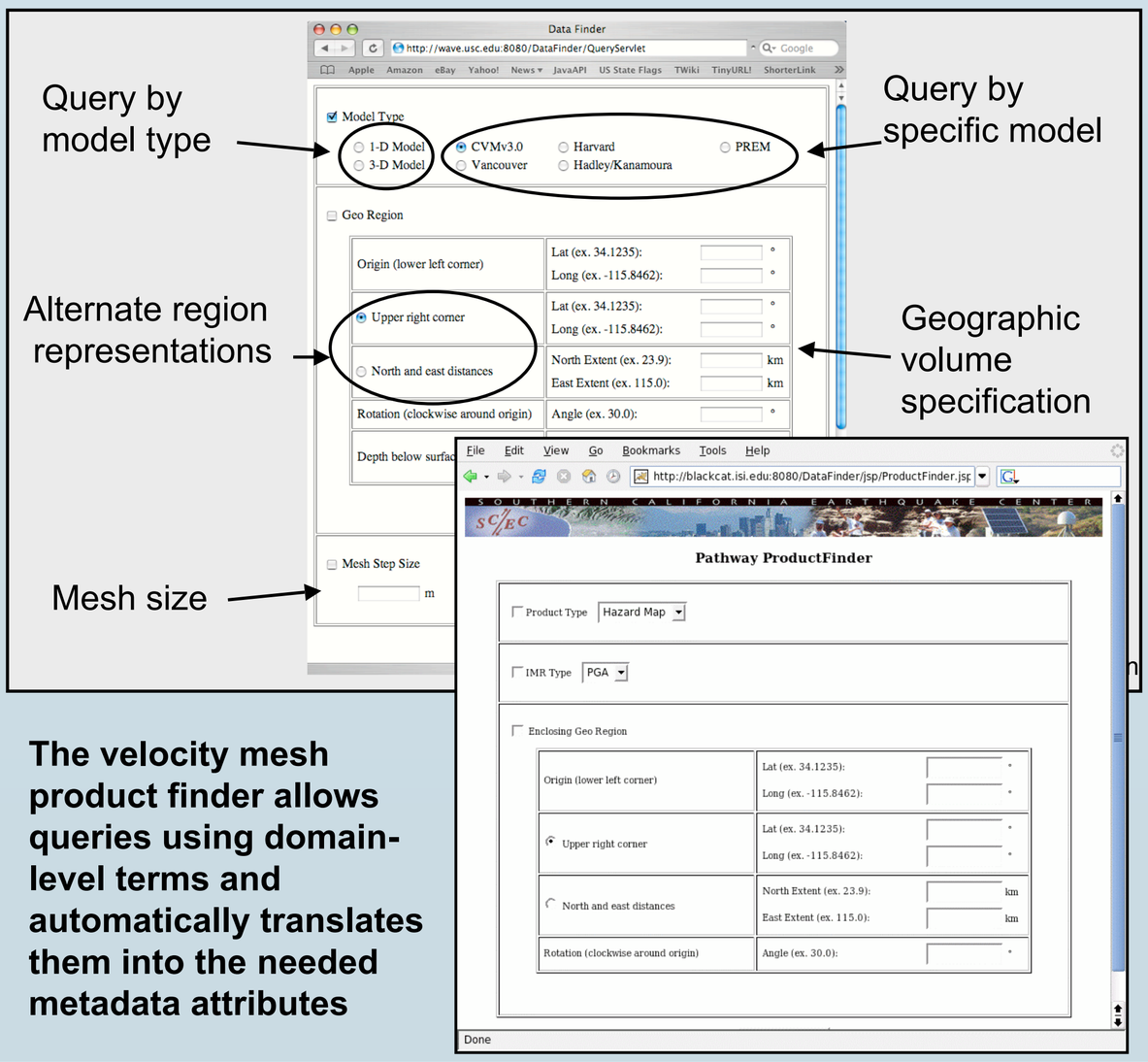
→

```
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™ 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



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