

# The Big Dig

## Global DiGIR Monitor

Dave Vieglais

Biodiversity Research Center  
University of Kansas

# What is DiGIR?

- “Distributed Generic Information Retrieval”
- Simple protocol for federated database access
- Messages as XML over HTTP
- “Federation Schemas” in XMLSchema
- Widely deployed for specimen collections
- Started about year 2001
- Many contributing developers

# What is the “Big Dig”

Automated (“hands off”) system for discovering and monitoring DiGIR data providers, their resources, and federation schemas.

Located at <http://bigdig.ecoforge.net>

Not to be confused with..

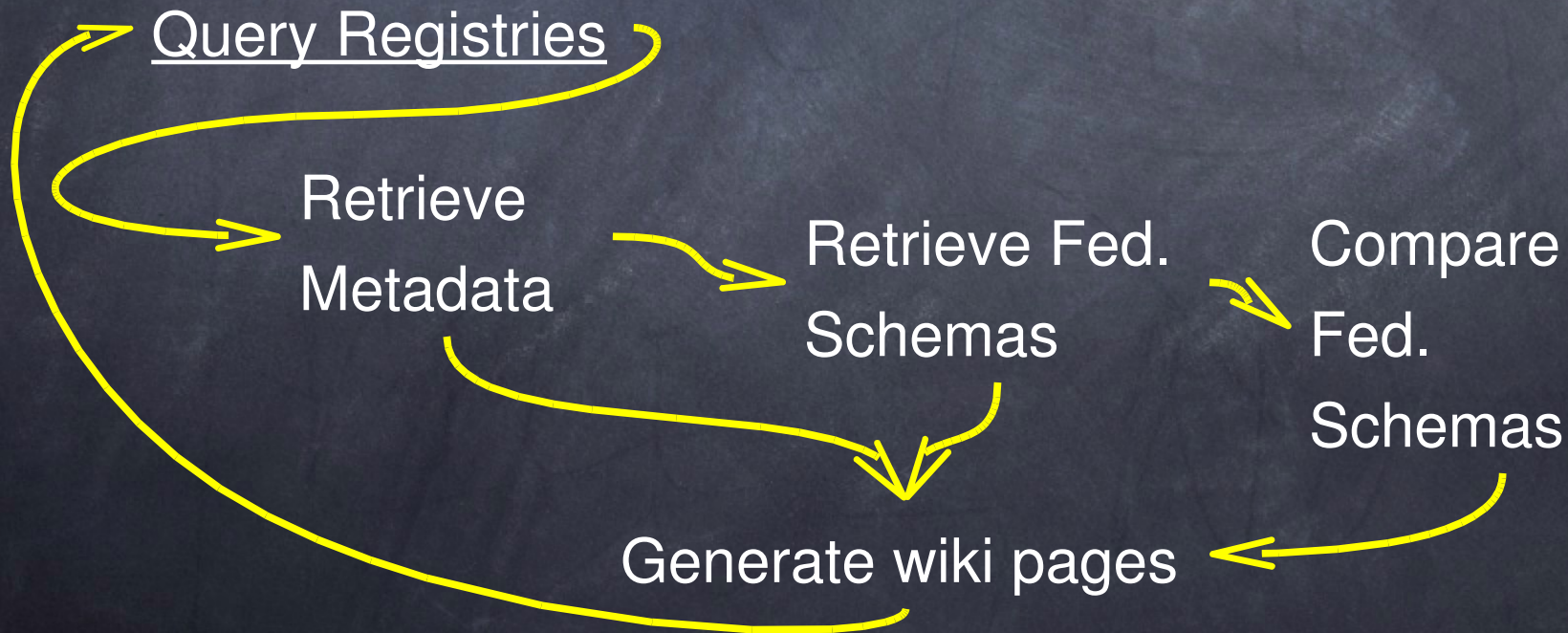
- “The most expensive highway project in America”, approx \$14 billion.
- This Big Dig cost about 116 triple lattes, no

# Why?

- Originally to find all the federation schemas, of which there seems to be considerable (overwhelming?) proliferation
- Quickly realized there are lots of problems with existing DiGIR installations, many could be easily fixed, leading to significantly improved data accessibility
- Evolved into basis for a monitoring service
- **Addresses a need that doesn't seem to be filled**

# Implementation

- Was: bash + curl + xsltproc (naive!)
- Now: python + MySQL + wiki
- Should of: Jython + RDF/WASABI



# The Product...

- Daily status checks of data providers
- Weekly updates of registered providers
- Weekly updates of schema cross-reference
- Automated comparison between schemas  
(normalized sum of weighted comparison  
between schema properties.)

# The Product cont'd...

## The Big Dig

The Big Dig is a service that generates an automated report on the current status of [DIGIR](#) data providers than can be found by examining a number of different registries.

A list of available data providers is available on the [ProviderStatus](#) page. Links to information about individual data providers are available from there.

Information about the Federation Schemas used by all the DIGIR providers can be found on the [SchemaStatus](#) page. Likewise, summary information about each schema is linked from that page. A comparison between all federation schema documents is also provided on the [SchemaStatus](#) page. The schema cross-reference is an automated attempt to provide a metric of the (dis)similarity between all schemas.

The information on data providers is updated approximately once a day. Information about schemas is updated approximately once per week.

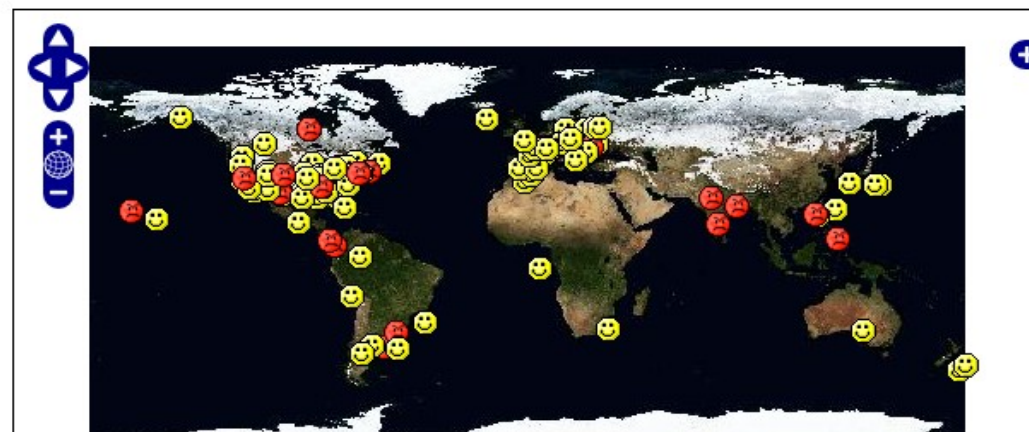
Quick Links:

- [ProviderStatus](#): see a list of all data providers
- [SchemaStatus](#): see a list of all federation schemas
- [ProvidersOffline](#): list of inaccessible providers

Quick Stats as of 2006-10-14 06:17:29+0000:

Number of federation schemas	29
Number of data providers	180
Total number of resources	964
Total number of records	76719383
Number of inaccessible providers	38

Map showing the approximate distribution of DIGIR providers (georeferenced by IP address). Yellow markers indicate providers known to be online, red indicates the provider was offline when last checked. If you have [Google Earth](#) then the KML file can be downloaded [here](#). Note that the provider locations are modified by a small random factor so that icons from the same institution don't fall on exactly the same spot.



# Results

- Apparently 17 unique fed schemas out there in the wild, definitely 12 that are accessible
- About 180 registered providers. About 25 have never responded with metadata, a couple block public access
- Some 80 million records (varies between 69 and 82 million)
- There are problems, but most (all?) easily

# Results Cont'd...

- Character encoding (utf-8 please!)
- Namespace duplication (expletives deleted)
- Federation schema proliferation (inheritance)
- Typos (configuration files)
- Installed versions (maintenance)
- Assign LSIDs to providers
- Operation consistency

# Conclusions / Future

- In general, provides useful feedback for new and existing networks
- Pretty much done, add some charts
- Perhaps expose a “provider lint” service
- Add feed-back to providers (email provider or network admins)
- Monitor other protocols?