Metadata in the Wild

An Empirical Survey of OPeNDAP-accessible Metadata and its Implications for Discovery

Darren Hardy
University of California, Santa Barbara

and the *DAPADL* project: Greg Janée & James Frew at UCSB,
James Gallagher & Peter Cornillon at OPeNDAP, Inc. & U Rhode Island.
Funding from NSF.

2006 AGU Fall Meeting, San Francisco, CA, 15 December 2006

Outline

- OPeNDAP and metadata
- Search problem and approach
- Empirical survey results

Scientific Data Access

- What is OPeNDAP?
  - Standard, protocol, and software
  - Data Access Protocol (DAP) designed for remote access to science data

Metadata

- What metadata does OPeNDAP provide?
  - *Syntactic* via DDS (i.e., to plot data)
  - *Semantics* via DAS (i.e., to label data)

Metadata Example

Syntactic

a variable that is a 3 dimensional grid of floating point numbers

```
Grid {
 ARRAY:
    Float32 zeta_rg[time = 109]
                   [latitude_rg = 144]
                   [longitude_rg = 160];
 MAPS:
    Float32 time[time = 109];
    Float32 latitude_rg[latitude_rg = 144];
    Float32 longitude_rg[longitude_rg = 160];
} zeta_rg;
```

Semantic

data are surface elevation measurements in meters from UNC model

```
zeta_rg {
    String long_name "Surface Elevation (Regular Grid)";
    String standard_name "sea_surface_elevation_regular_grid";
    String units "m";
}
```

Title
"Q51 SEACOOS SNFS netCDF file"

Model Name
"QUODDY"

Institution
"Uni. of NC @ Chapel Hill,
Dept of Marine Science."

Conventions
"CF-1.0"

Source: SouthEast U.S. Atlantic Coastal Ocean Observing System (SEACOOS)
http://nemo.isis.unc.edu/cgi-bin/nph-dods/data/nc-coos/model_data/quoddy/forecast/2004/SAB_Q51_2004_03_18.nc.html

Search Problem

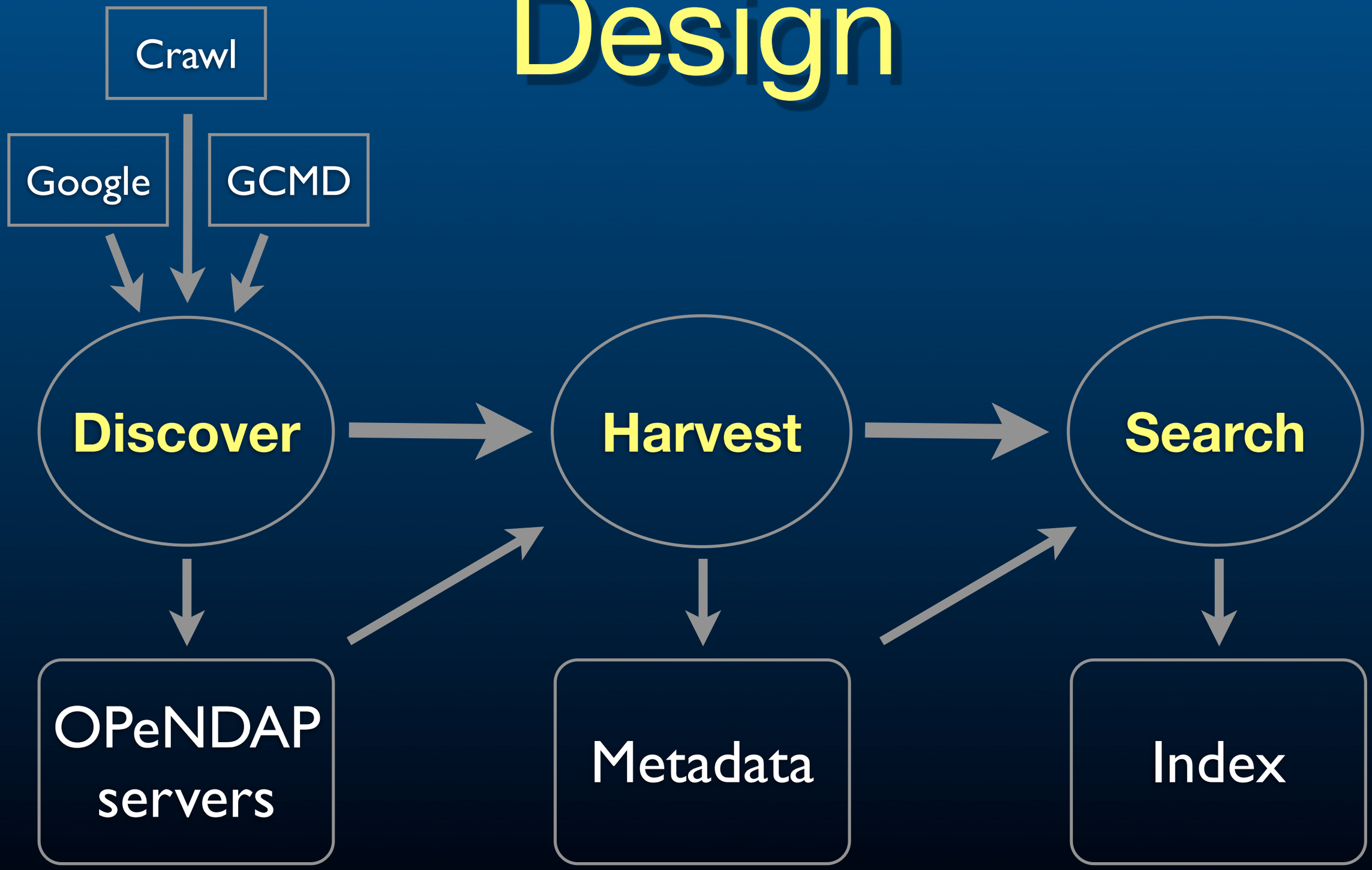
- *Where* are the data available via OPeNDAP?
  - Discovery via hand-crafted directories (e.g., NASA's GCMD)
- But, no global search service for data

Approach

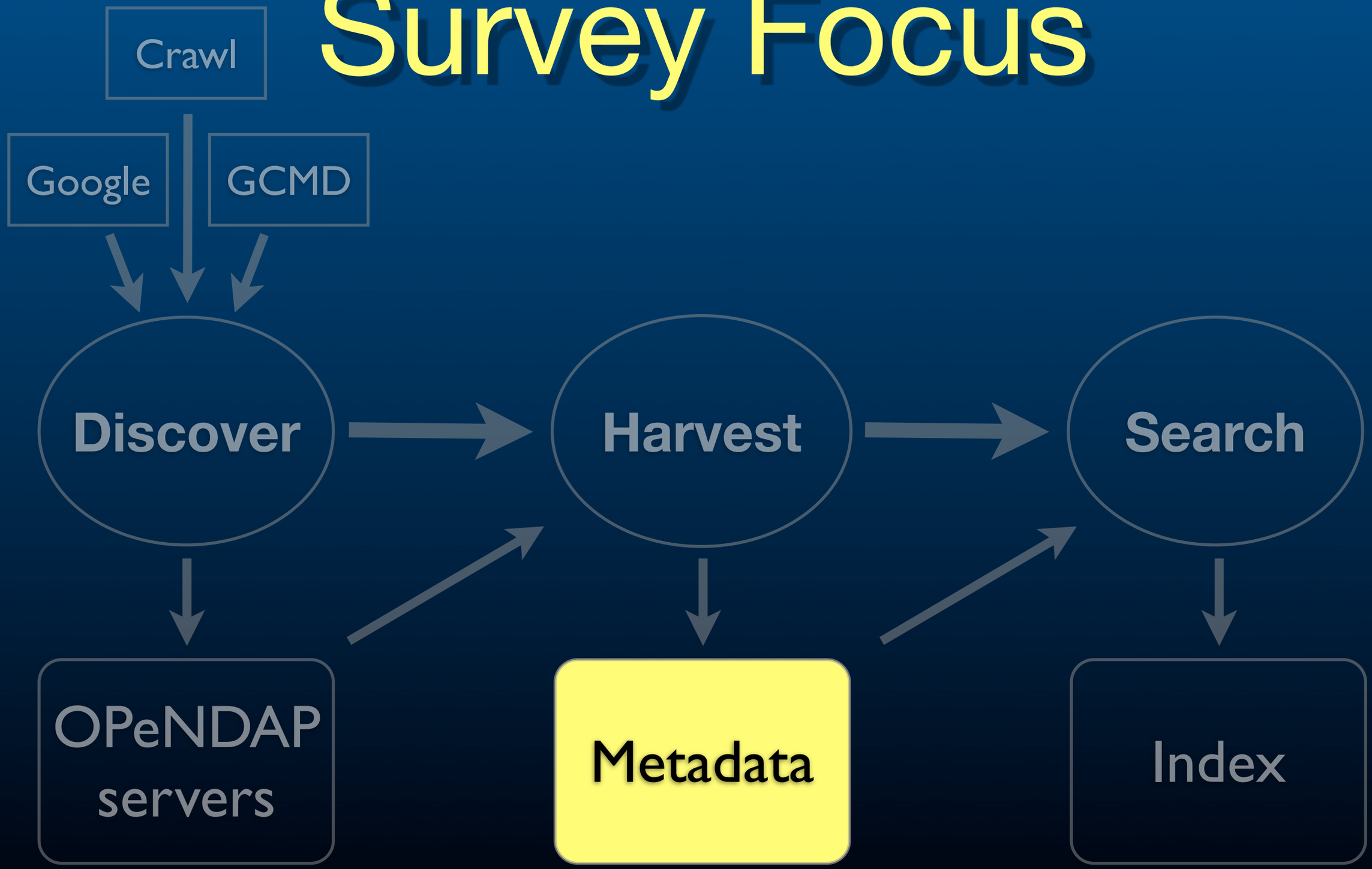
Unified search service for OPeNDAP

1. Text
   “sea surface temperatures”
2. Spatiotemporal coverage
   “in North Atlantic during last year”
3. Spatiotemporal resolution
   “daily averages at 6km”

Design


Crawl
Google
GCMD
Discover
Harvest
Search
OPeNDAP servers
Metadata
Index

Survey Focus



Survey Questions

- Goal: Characterize existing metadata
- Can we automate search services for...
  - Text?
  - Spatiotemporal coverage?
  - Spatiotemporal resolution?
- Are metadata conventions helping?

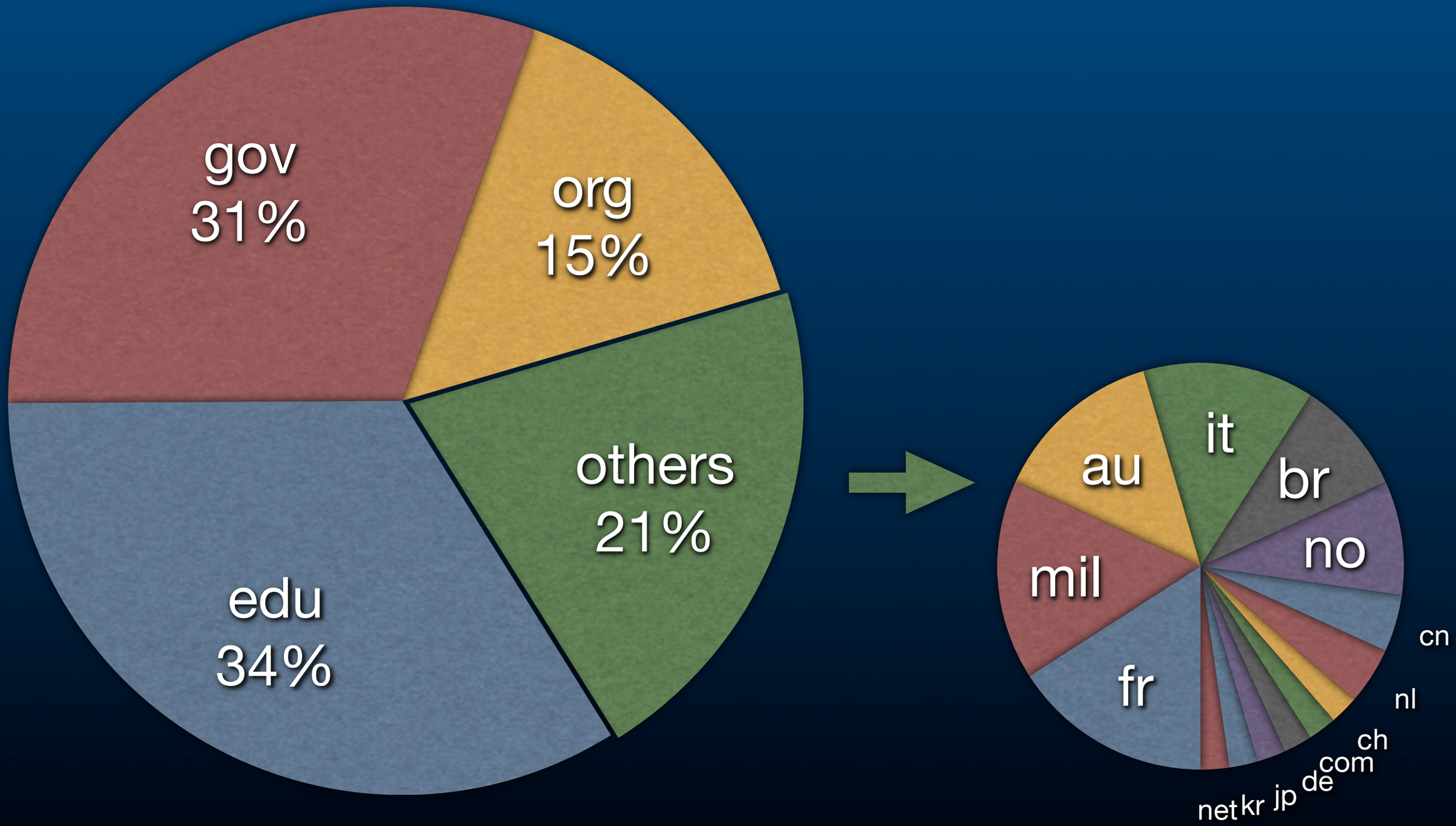
Initial Expectations

- Some expected diverse metadata
- Some expected similarity
- So, go do the survey...

Sample Size

	Total	Study Set *Min 10; omit Top 3*
Servers	213	162
Datasources	1,408,996	396,638
Variables	49,711,772	18,359,268
Attributes	349,319,571	57,833,469

Distribution of servers by domain (n = 213)

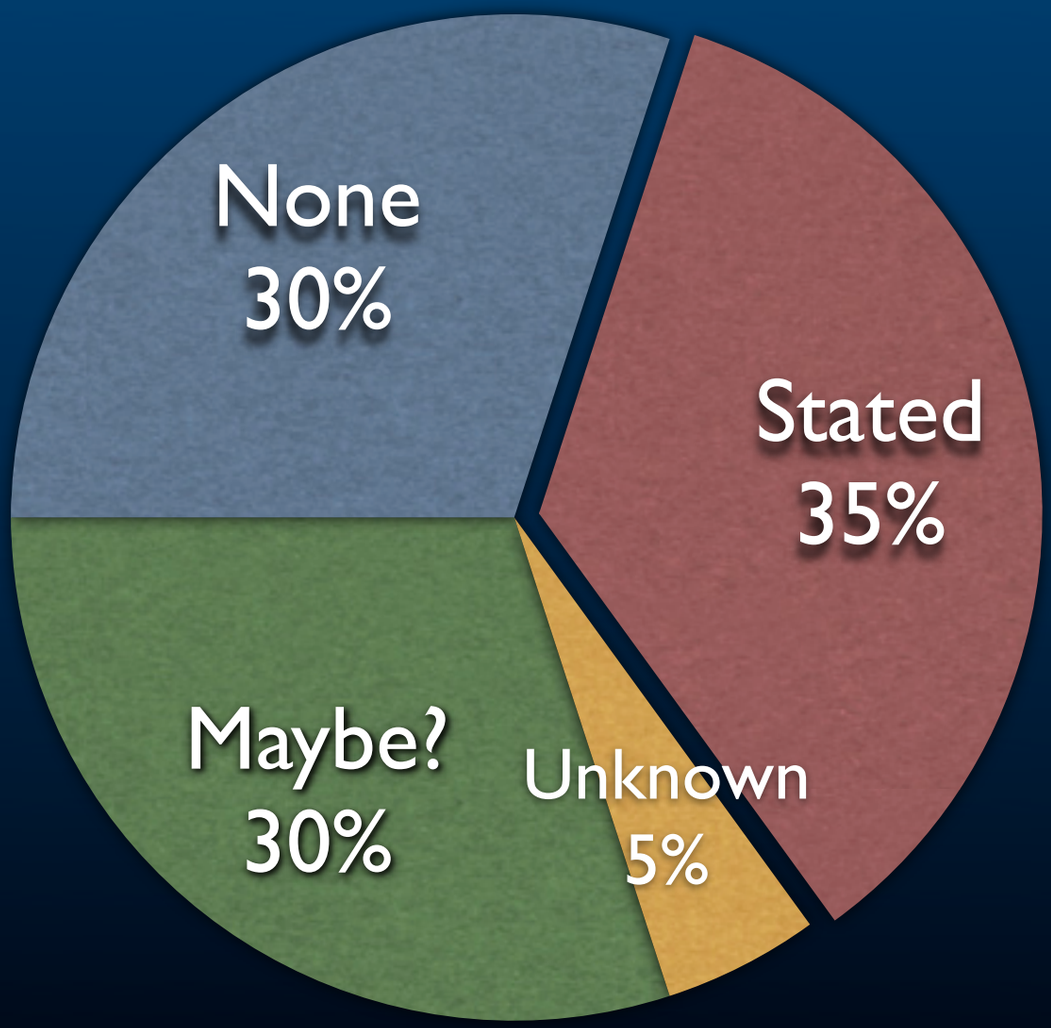


Suitability for Search

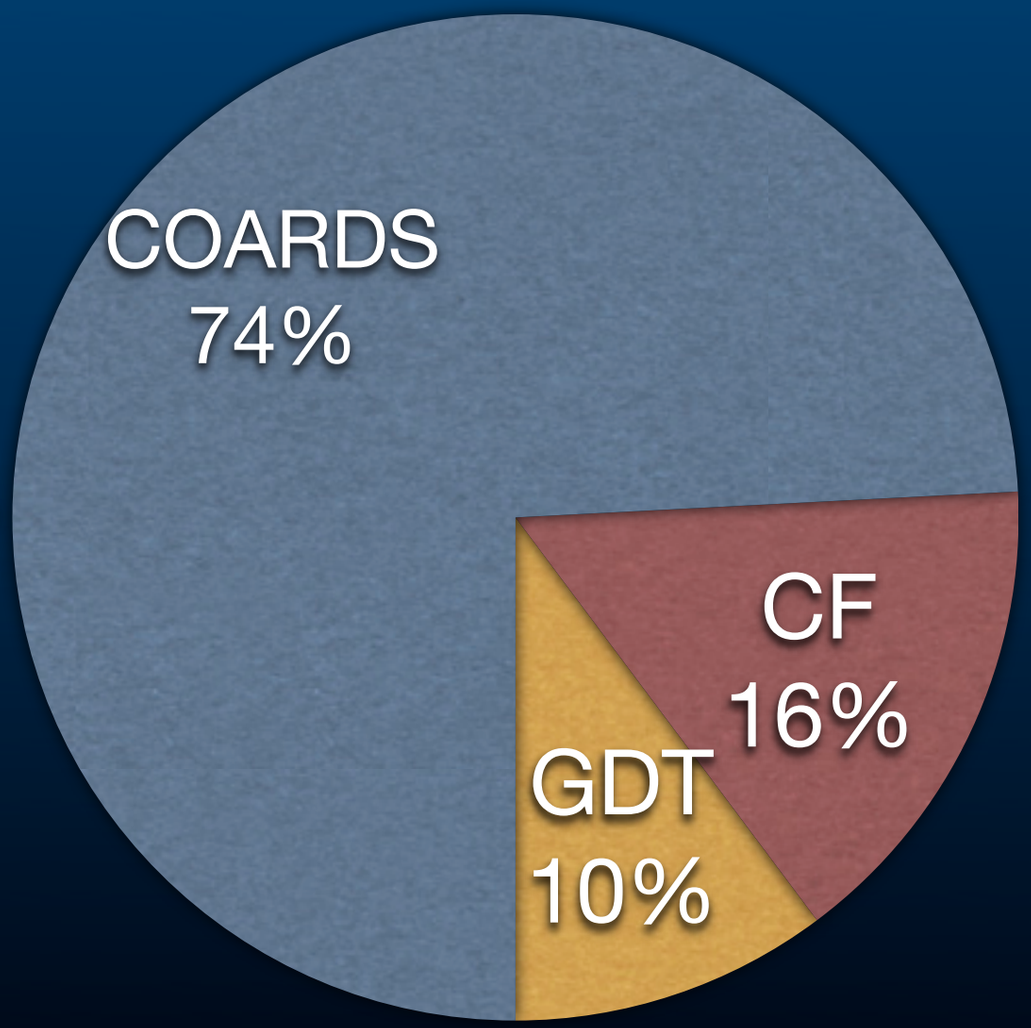
Text	**Good**	*80% have some general description (e.g., title, institution, etc.)*
Coverage	**Not Bad**	*Lat / Lon / Time identifiable 80% of the time, but reliable bounds <10%*
Resolution	**Poor**	*Spatial 40%, Time <10%, and requires data access*

Metadata Conventions

Use of Conventions
(n=396,638)



Types of Conventions
(when stated)



Some Findings

- High variability in suitability for search
  - Support for text search better than coverage or resolution
- Most metadata *appear* to use conventions (>60%)
  - Usually “COARDS-like”, strict compliance is rare
  - Even if, wouldn’t solve fundamental problems of semantic heterogeneity

Results

- Study: 150+ servers, 400k datasources
- Suitability of existing metadata for search:
  - Text? *OK but minimal.*
  - Coverage/resolution? *Maybe.*
- Metadata conventions (COARDS, CF)
  - 35% stated & *possibly* 30% more

Summary

- Unified search service for OPeNDAP
- Survey of existing metadata from 162 servers
- Key findings
  - Semantic heterogeneity is large in general
  - But, core “search” semantics has surprising good homogeneity
  - Exploit syntax & data for semantic purposes

Thank you.