Discovering OPeNDAP Datasets

Greg Janée
gjanee@alexandria.ucsb.edu

What data exists?

- Multiple approaches
  - registries (GCMD, etc.)
  - co-opting search engines (Google, etc.)
  - crawling
    - custom
    - reuse others' (e.g., Stanford WebBase)
- Ours (so far): combination of registries and Google

Where is the data?

- Outline
  - theoretical metadata model
  - queries
  - a survey of reality
  - approach

Metadata we care about

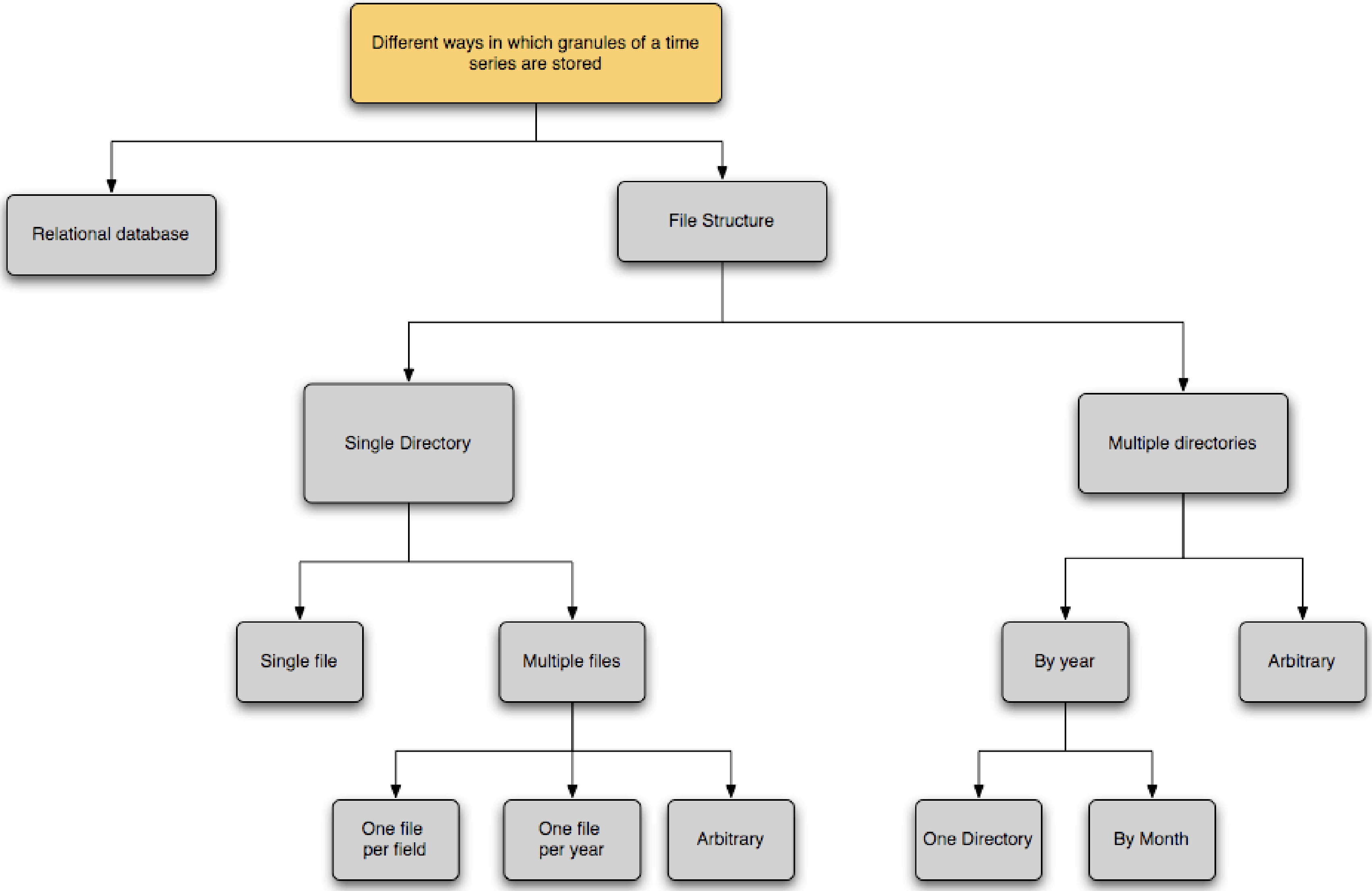
- “Text”
  - title, description, etc.

- Axes by which data is organized
  - common axis across datasets provides an indexing mechanism

Standard axes

- Some axes are more equal than others

- Space
- Time
- Spatial resolution
- Temporal resolution
- Text



Thanks, Peter!

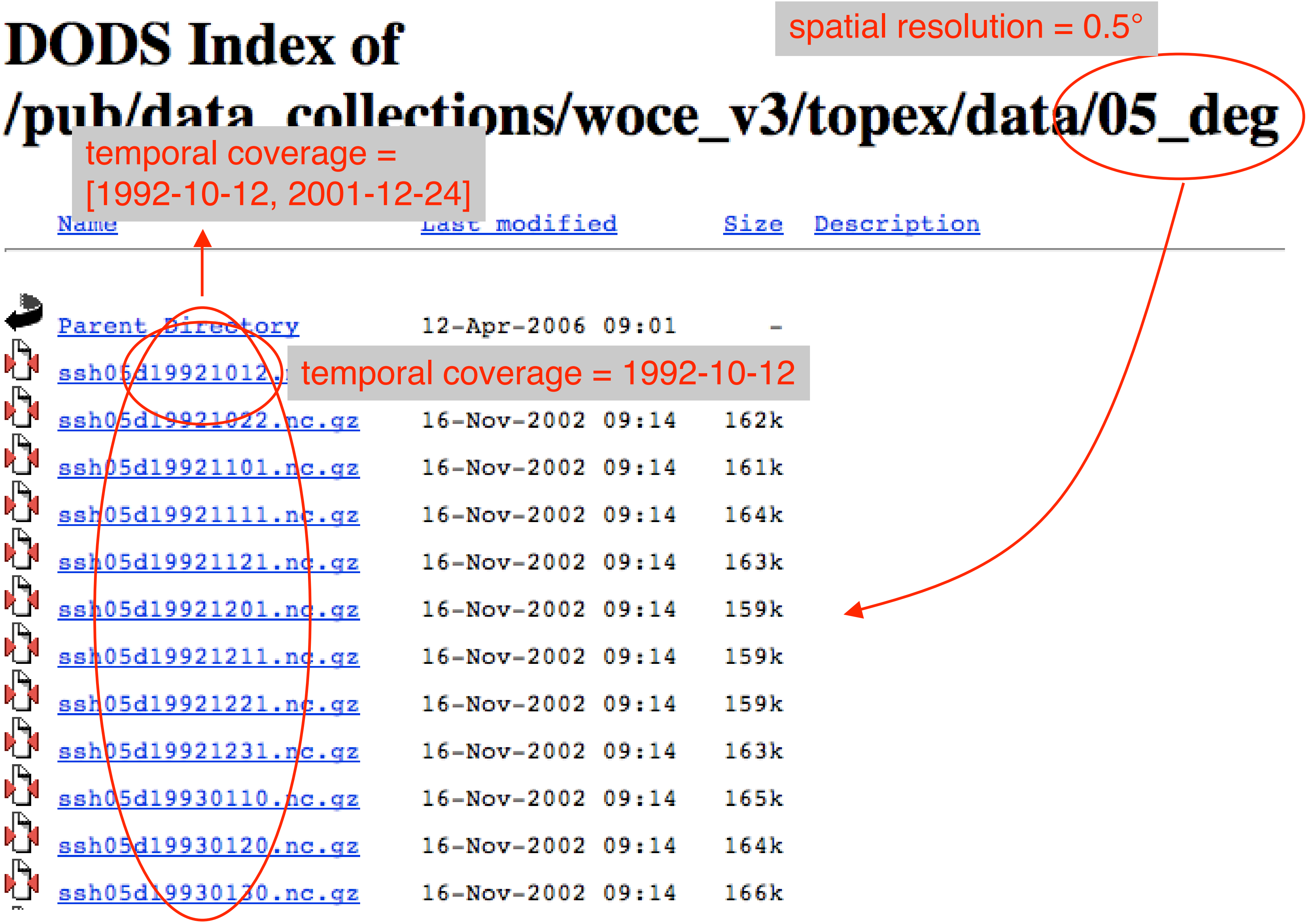
Metadata: crawling

- Start from server root directory or THREDDS catalog
- Directories/collections provide semantics
- Granules inherit, override parent metadata
- Derive metadata from data itself

- Result: contextual metadata propagated down to granule level

Metadata: aggregation

- Common axis among container members propagates up an aggregate value for the axis



DODS Index of /pub/data_collections/woce_v3/topex/data/05_deg

Name	Last modified	Size	Description
Parent Directory	12-Apr-2006 09:01	-	
ssh05d19921012.			
ssh05d19921022.nc.gz	16-Nov-2002 09:14	162k	
ssh05d19921101.nc.gz	16-Nov-2002 09:14	161k	
ssh05d19921111.nc.gz	16-Nov-2002 09:14	164k	
ssh05d19921121.nc.gz	16-Nov-2002 09:14	163k	
ssh05d19921201.nc.gz	16-Nov-2002 09:14	159k	
ssh05d19921211.nc.gz	16-Nov-2002 09:14	159k	
ssh05d19921221.nc.gz	16-Nov-2002 09:14	159k	
ssh05d19921231.nc.gz	16-Nov-2002 09:14	163k	
ssh05d19930110.nc.gz	16-Nov-2002 09:14	165k	
ssh05d19930120.nc.gz	16-Nov-2002 09:14	164k	
ssh05d19930130.nc.gz	16-Nov-2002 09:14	166k	

spatial resolution = 0.5°

temporal coverage = [1992-10-12, 2001-12-24]

temporal coverage = 1992-10-12

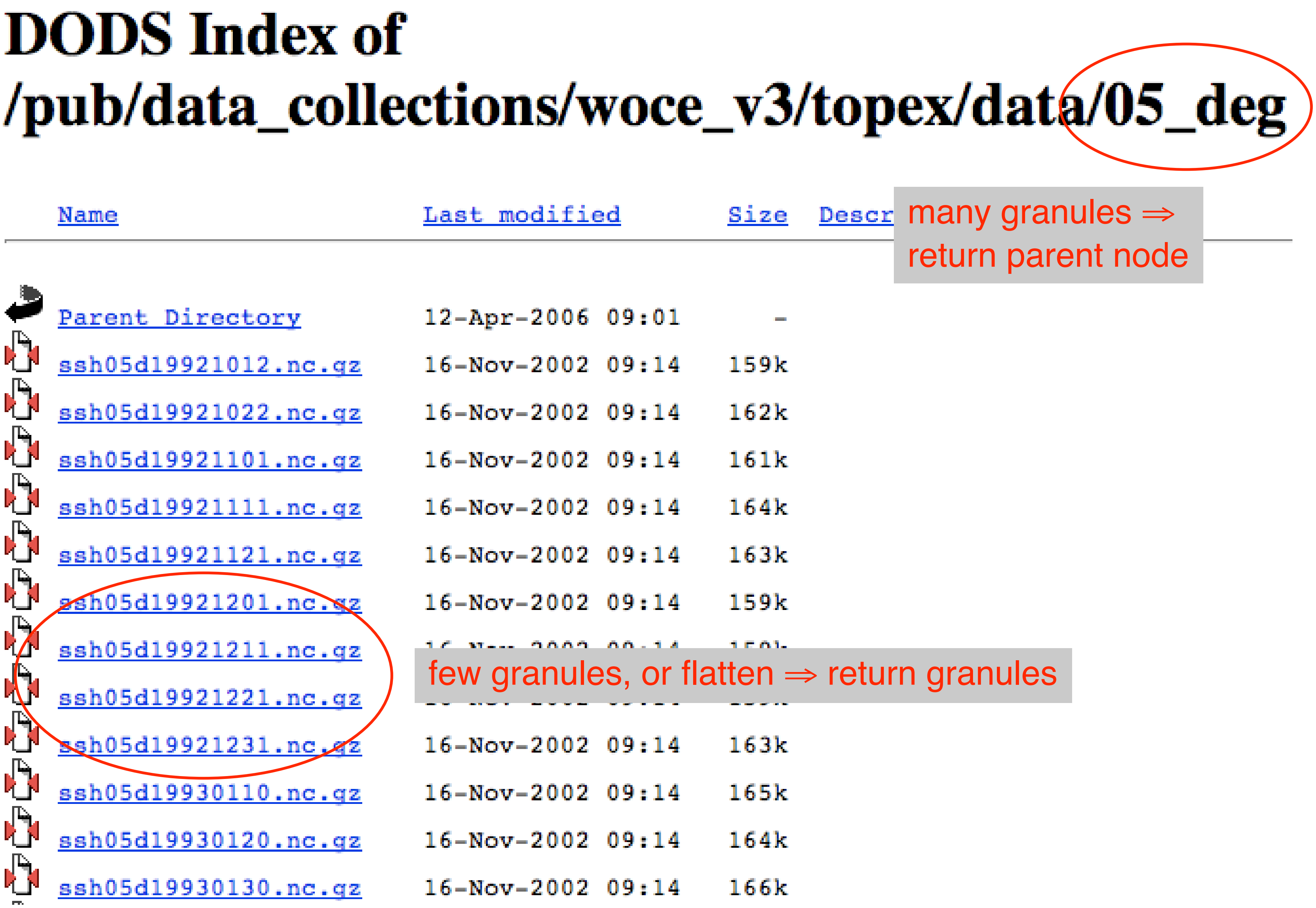
Metadata: aggregation

- Common axis among container members propagates up an aggregate value for the axis

- Result: hierarchy of nodes
  - containers: recursively searchable
  - atomic: OPeNDAP access points
  - both uniformly described by axis metadata

Queries

- Place constraints on axes
  - start with standard axes
- Return matching nodes, ranked by fit
- Several query modes
  - manual drill down: return next level
  - flatten: return granules
  - adaptive



DODS Index of /pub/data_collections/woce_v3/topex/data/05_deg

```
    Name                           Last modified       Size  Descr

    Parent Directory               12-Apr-2006 09:01      -
    ssh05d19921012.nc.gz           16-Nov-2002 09:14   159k
    ssh05d19921022.nc.gz           16-Nov-2002 09:14   162k
    ssh05d19921101.nc.gz           16-Nov-2002 09:14   161k
    ssh05d19921111.nc.gz           16-Nov-2002 09:14   164k
    ssh05d19921121.nc.gz           16-Nov-2002 09:14   163k
    ssh05d19921201.nc.gz           16-Nov-2002 09:14   159k
    ssh05d19921211.nc.gz           16-Nov-2002 09:14   159k
    ssh05d19921221.nc.gz           16-Nov-2002 09:14   [illegible]
    ssh05d19921231.nc.gz           16-Nov-2002 09:14   163k
    ssh05d19930110.nc.gz           16-Nov-2002 09:14   165k
    ssh05d19930120.nc.gz           16-Nov-2002 09:14   164k
    ssh05d19930130.nc.gz           16-Nov-2002 09:14   166k
```

many granules ⇒ return parent node

few granules, or flatten ⇒ return granules

Challenges

- Darren’s survey

Approach

- Modular framework
  - pluggable heuristics
  - manually overridable

- Model: previous ADL metadata mapping work
  - mapping language embedded in Python
  - mapping inheritance