Toward Rich, User-Defined Aggregation & Subset-Selection Services

Dave Fulker, President, OPeNDAP, Inc

ESIP Summer Meeting on 9-12 July 2013
Thursday 3:30 PM Session

Aggregation / Subsetting — What's It To You

A Thin Slice of

Subset-Selection History

- ‘70s — Relational data bases demonstrated
  Data access ⟺ [data model + operations]
  - Ops included select-by-value subsetting
  - But without array types (generally)
- ‘80s-‘90s — CDF, NetCDF, HDF
  - Non-tabular data models (arrays / select-by-index)
  - But without select-by-value operations
- Mid ‘90s — DAP2 *(DODS ➔ OPeNDAP)*
  - Remote access (a Web service)
  - Most often employed via NetCDF API

Why Arrays Merited Sacrifice of Select-by-Value

- N-dim arrays are natural in key cases
  - Imagery (satellites, radars, telescopes...)
  - Simulations on rectangular meshes
- Arrays often yield great efficiencies
  - Store/retrieve/subset without searching/testing
  - Ideal for derivative & image-processing ops
- Furthermore...
  - What does select-by-value mean?

Note: OPeNDAP Actually Offers Both Forms of Subset Selection

- DAP2 (1993) embraced "sequences"
  - With select-by-value subset creation
- DAP2 *distinguished* these from arrays, where selection occurs by index
  - Exception: bounding-box constraints may be applied to "coordinate-variable" arrays
- DAP sequences are relatively rare
  - Partly because netCDF API ignores them
  - Notable exceptions: ERDAP & JGOFS

A Thin Slice of

(NetCDF) Aggregation History

- late ‘80s — Unidata discussions (incl ideas about Unix-style data filters) → NetCDF
- late ‘90‘s — Zender’s NetCDF Operators included “concatenation”
- ‘00s — THREDDS virtual aggregation
  - Defined to be a server configuration (employing NCML), akin to concatenation
  - Adopted in other DAP-based servers
  - Configured by providers, not users

NCML-Style (Virtual) Aggregation Is Ideal when

- ✦ The results are "natural"
  - E.g., time is a coordinate rather than something encoded in (obtuse) granule names
- ✦ Multi-granule access enhances usability
  - Easy to subset in desired ways
  - Does not lead to excessive delays, etc.
- ✦ Providers can't go wrong...

...but NCML-Based Aggregation by Providers Is Less than Ideal...

- ✦ If granules don't "line up" naturally
  - E.g., swath data might be aggregated in various ways, none satisfying all users
- ✦ If cross-granule access is slow/costly
  - E.g., retrieval might involve per-granule overhead such that latencies accumulate
- ✦ *Lesson:*
  abstraction is nice; concrete is hard

Idea: Empower Users for Subset Selection & Aggregation

- Even in hard cases, users may know how & whether granules should be aggregated
- Users may know how they'd like to receive results of test-by-value operations
- Service-invocation protocols could allow (beyond subset selection) —
  - *A rich set of pre-retrieval operations or even workflows*

OPeNDAP✳ Is Pursuing

✳ *with Unidata & NOAA/PMEL*

✦ Funds (NSF & NASA) to explore:

- A rich set of pre-retrieval server functions (beyond index-based subset selection)
- A protocol/language for invoking these

✦ Success would augment current efforts (NOAA-funded) to complete DAP4

Final Thoughts

- Pre-retrieval ops will be of increasing value (data-proximate computation...)
- Invocation language is key to success
  - Must enable community-driven extension
  - Aggregation expressions must refer to multiple granules
- There are significant challenges
  - Granule/subset selection often requires iteration
  - Operations affect provenance & other metadata
  - Should avoid ops that increase xfer volumes
  - Async/cached responses may be required...

Thank You

✦ OPeNDAP site — www.opendap.org

✦ Dave Fulker — dfulker@opendap.org