Data Diversity in the Research Data Archive

Bob Dattore

Steven Worley

Where we are now

Preface: mainly take what the data providers give us

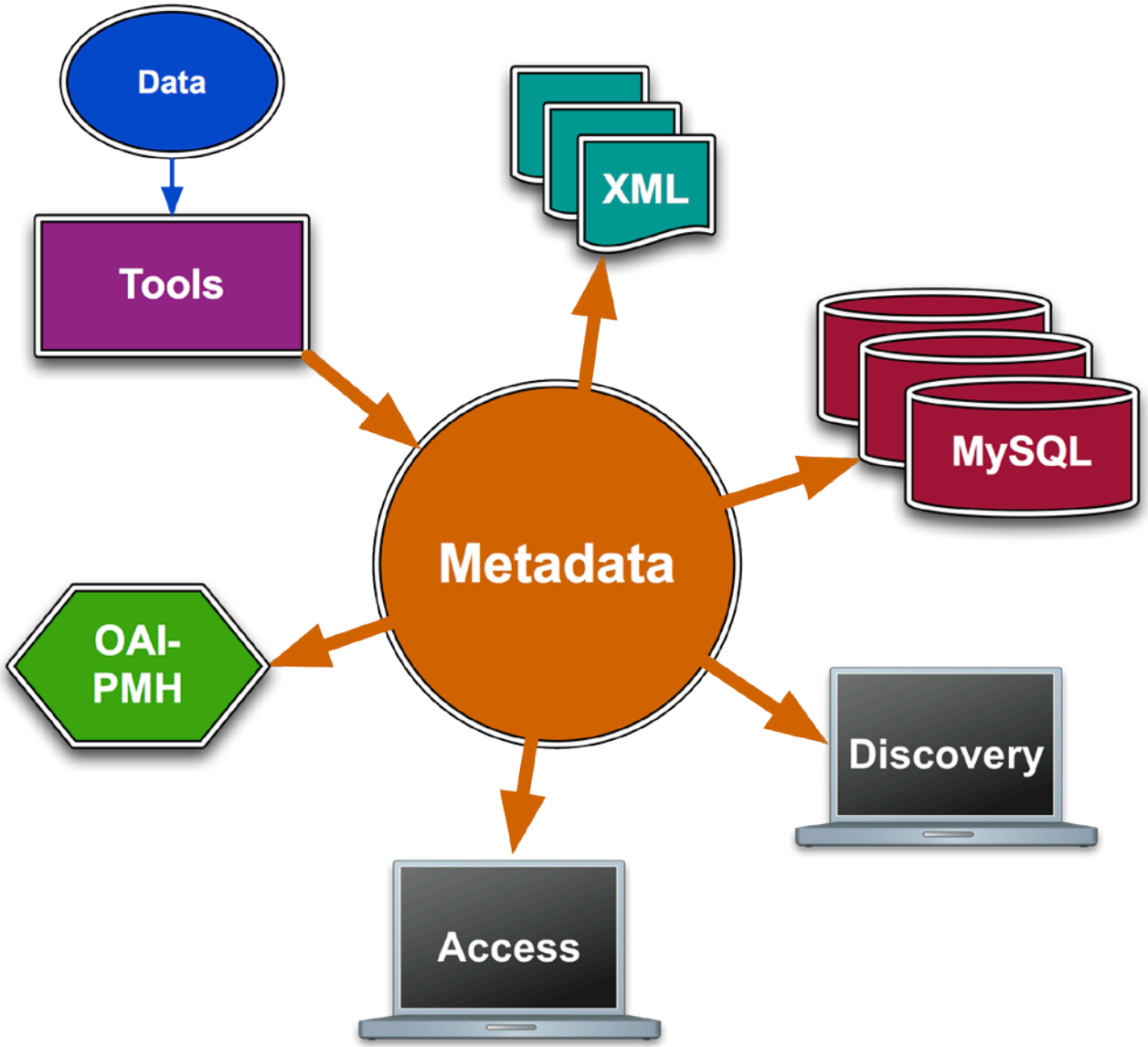
Data types/organization:

- Regularly-spaced grids
- In-situ observations
- Some remotely-sensed (e.g. - satellite, radar)
- Synoptic, time-series
- Data packaging

Where we are now

Data formats:

Where we are now



Where we plan to be near-term

Grids: currently have a good handle on these

Observations:

- A lot of diversity
  - Hourly, daily, monthly, data gaps, varying record lengths
  - Multiple station networks, different observing practices
  - Real-time streams vs. post-processed (qc'd)
- Subsetting - spatial, temporal, parameter, etc.

Format conversions:

- what we receive --> what users want

Challenges for the long-term

Bookkeeping:

- Managing the metadata (5000 tables, 1.1TB+)
- Table-driven data formats (e.g. - GRIB)
- Maintain expertise and/or migrate data
  - Parallel archives? Software solutions?

Seamless data access:

- Shield user from “datasets”
- Data providers use “standard formats” differently

Data Citation/Reproducibility:

- Ability to accurately reproduce cited data (bookkeeping)
- What about data provided by seamless access? (underlying datasets might each have DOIs)