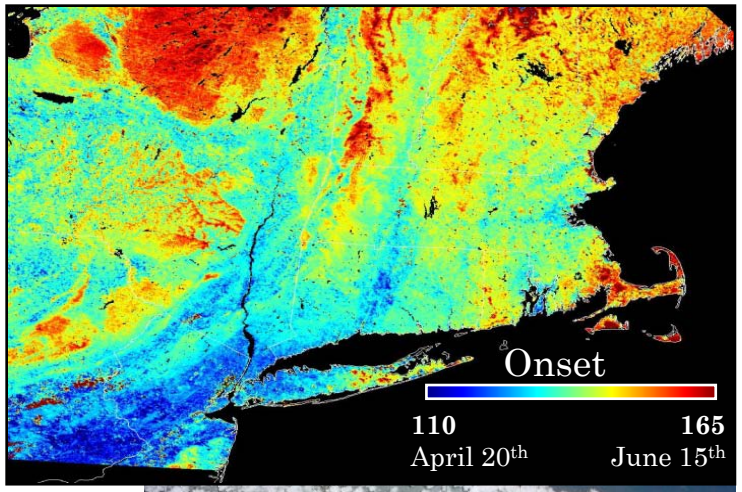
Beyond Sensors: Curating Ancillary Data forWater Science Synthesis

Catharine van Ingen
Microsoft Research
13 November 2009

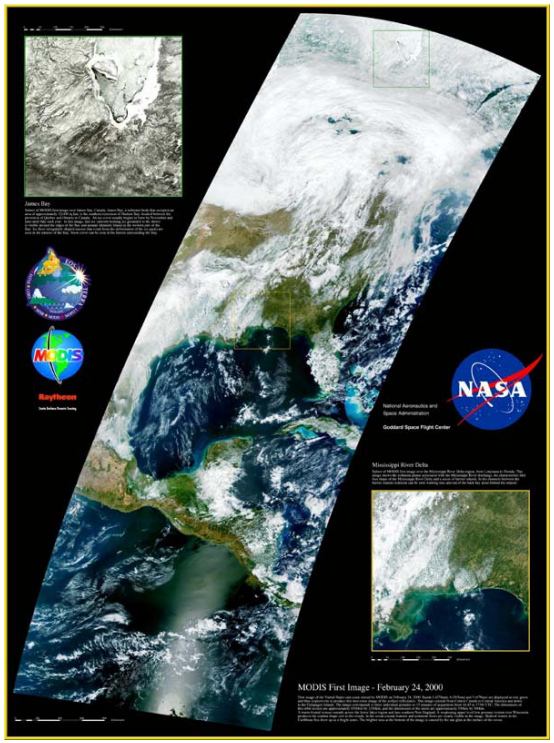
Introduction

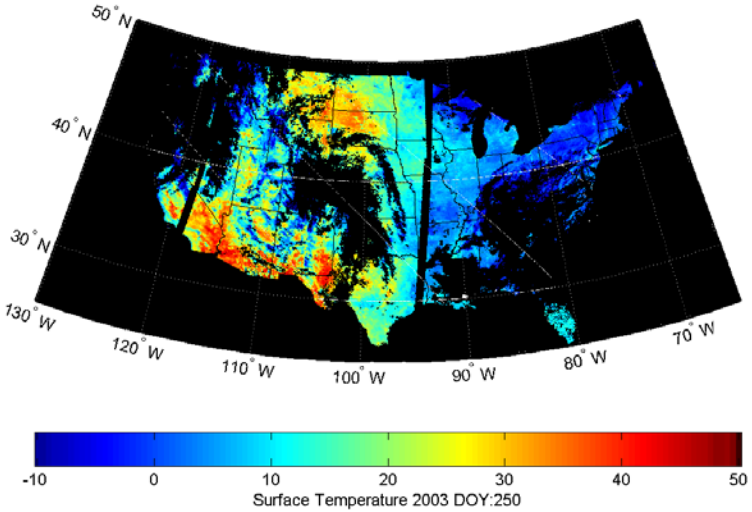
- The era of remote sensing, cheap ground-based sensors and web service access is here
- Turning sensors into science often requires additional "ancillary data"
- This talk relates our experiences since early 2007 curating hydrologic and other environmental data sets



Environmental Remote Sensing Data

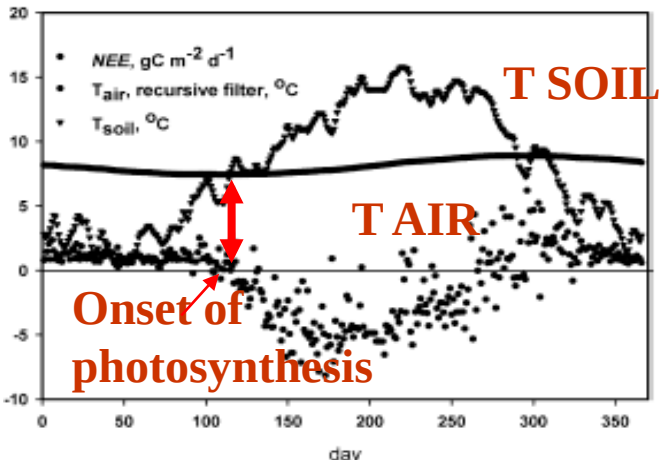
- Time series raster data
  - Over some period of time at some time frequency at some spatial granularity over some spatial area
  - Conversion from L0 data to L2 and beyond as well as reprojections still a specialized skill
- Can be “cut out” to creat virtual sensors
- Today: PBs (L0) to TBs (L2+)

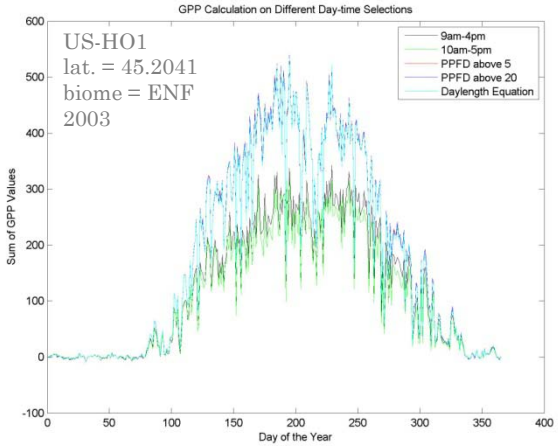




Environmental Sensor Data

- Time series data
  - Over some period of time at some time frequency at some spatial location.
  - May be actual measurement (L0) or derived quantities (L1+)
- (Re)calibrations, gaps and errors are a way of life.
  - Birds poop, batteries die, sensors fail.
  - Various quality assessment and signal correction algorithms.
  - Gap filling algorithms key as regular time series enable more analyzes
- Today: GBs to TBs

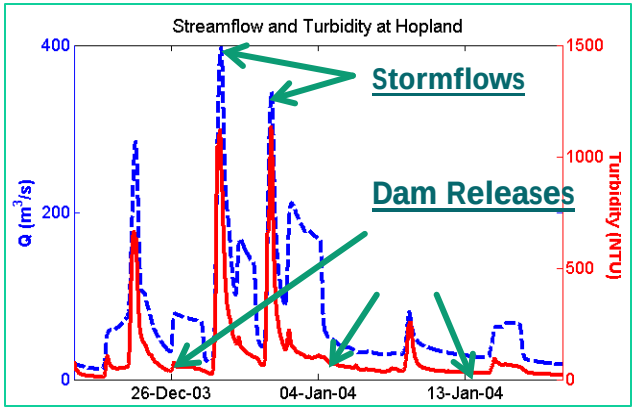


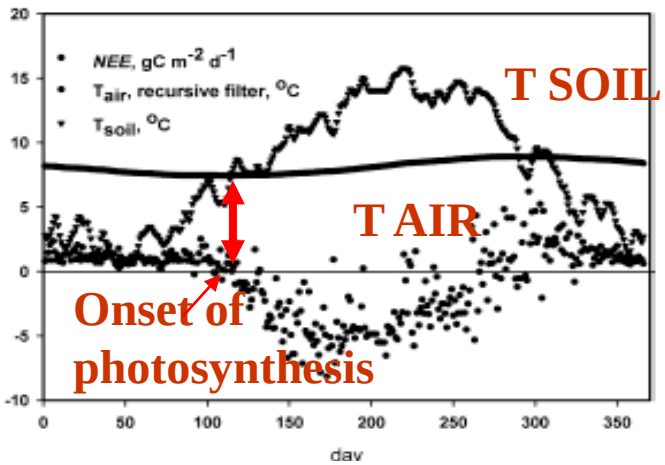


"Time is not just another axis"

Environmental Sensor Data

- Time series data
  - Over some period of time at some time frequency at some spatial location.
  - May be actual measurement (L0) or derived quantities (L1+)
- (Re)calibrations, gaps and errors are a way of life.
  - Birds poop, batteries die, sensors fail.
  - Various quality assessment and signal correction algorithms.
  - Gap filling algorithms key as regular time series enable more analyzes
- Today: GBs to TBs

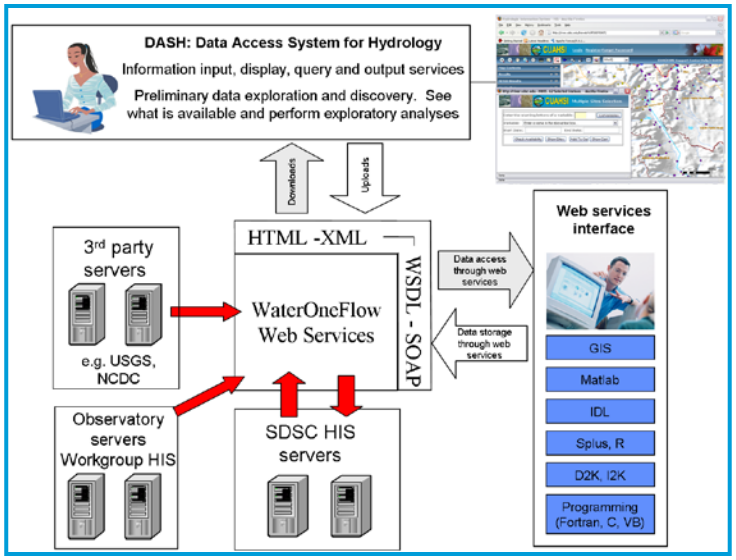


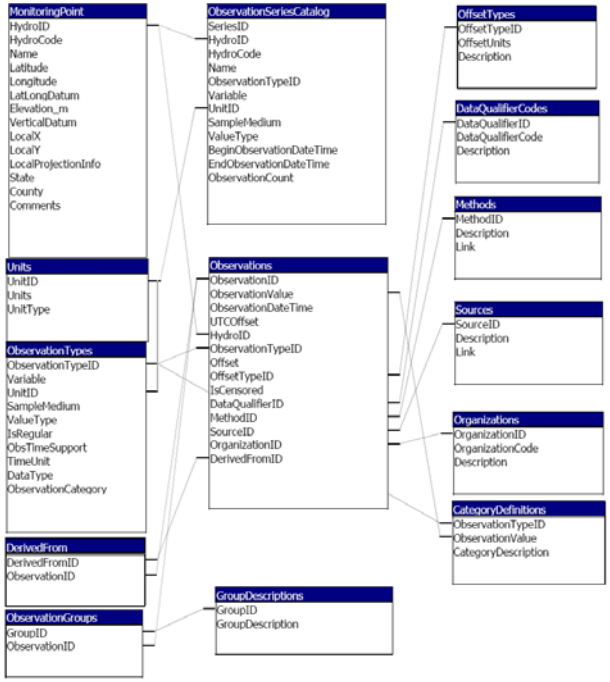


"Time is not just another axis"

Sensor Databases and Web Services

- Many groups using databases and web services to access, curate, and republish sensor data
- Most use a mostly normalized schema with the data in the center
- Example is CUAHSI ODM
  - Initially targeted to solve the problem that in the US, agency data available over the internet, but too hard to find, too hard to download all the data, too hard to get "just the new data"
  - Included water quality bottle samples, a notion of data revisions
  - 11 initial tests beds growing over time



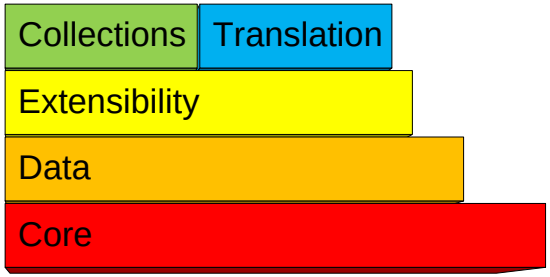


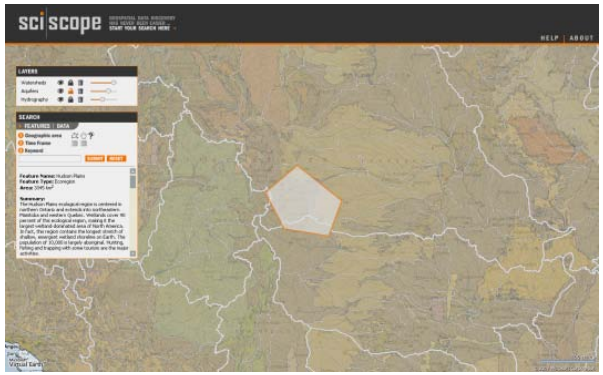
http://www.cuahsi.org

CUAHSI
universities allied for water research

Combined learnings from BWC and CUAHSI HIS

- Put the catalog, not the data at the center for both simplicity and scaling
- Add geospatial support for features such as watershed or ferry path
- Improved “folder” concept for versions and groups
- Assume that no single ontology or controlled vocabulary exists – build translation tables to store source-specific original and “renames”
- Subset for simple bites (it’s complex!)
- Generalize site properties and store from the sensor data





http://www.sciscope.org

Series Catalog
SeriesID
SiteID
SiteCode
SiteName
VariableID
VariableCode
VariableName
VariableUnitsID
VariableUnitName
SampleMedium
ValueType
TimeSupport
TimeUnitsID
TimeUnitsName
DataType
GeneralCategory
MethodID
MethodDescription
SourceId
Organization
SourceDescription
QualityControlLevelID
BeginDateTime
BeginDateTimeUTC
EndDateTimeUTC
ValueCount

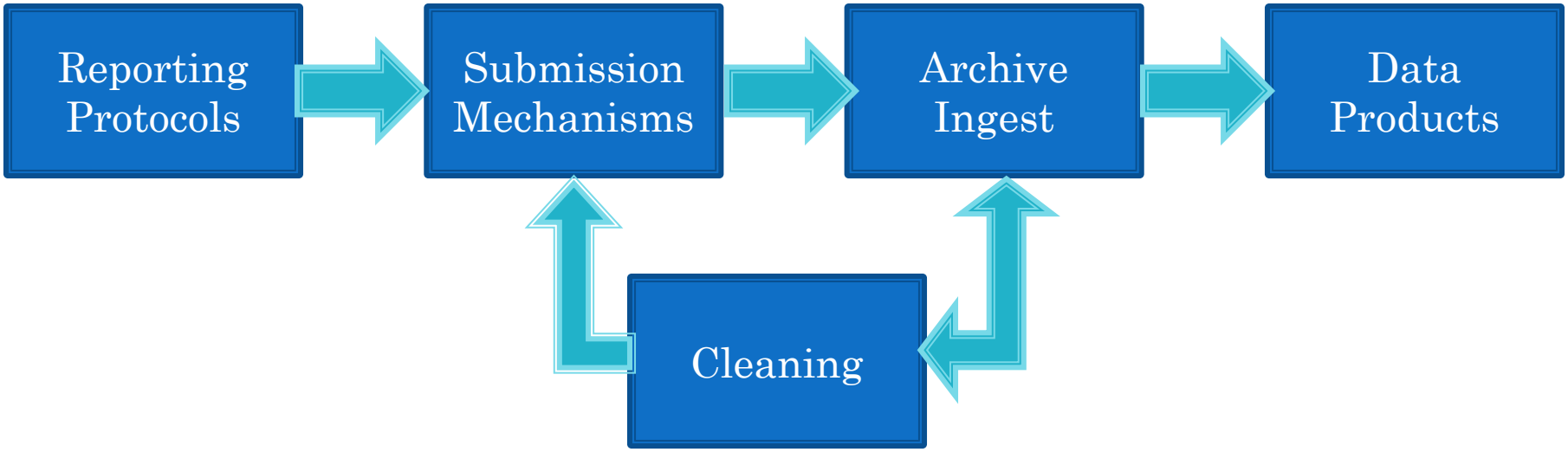
Environmental Ancillary Data

- Almost everything else!
  - ‘Constants’ such as latitude or longitude
  - Intermittent measurements such as grain size distributions or fish counts
  - Anecdotal descriptions such as “ripple” or “shaded”
  - Events such as algal blooms or leaf fall including those derived from sensor data such as “flood”
  - Disturbances such as a fire, harvest, landslide
- Not metadata such as instrument type, derivation algorithm, etc.
- Today: KBs to maybe GBs.

Ancillary Data is Different

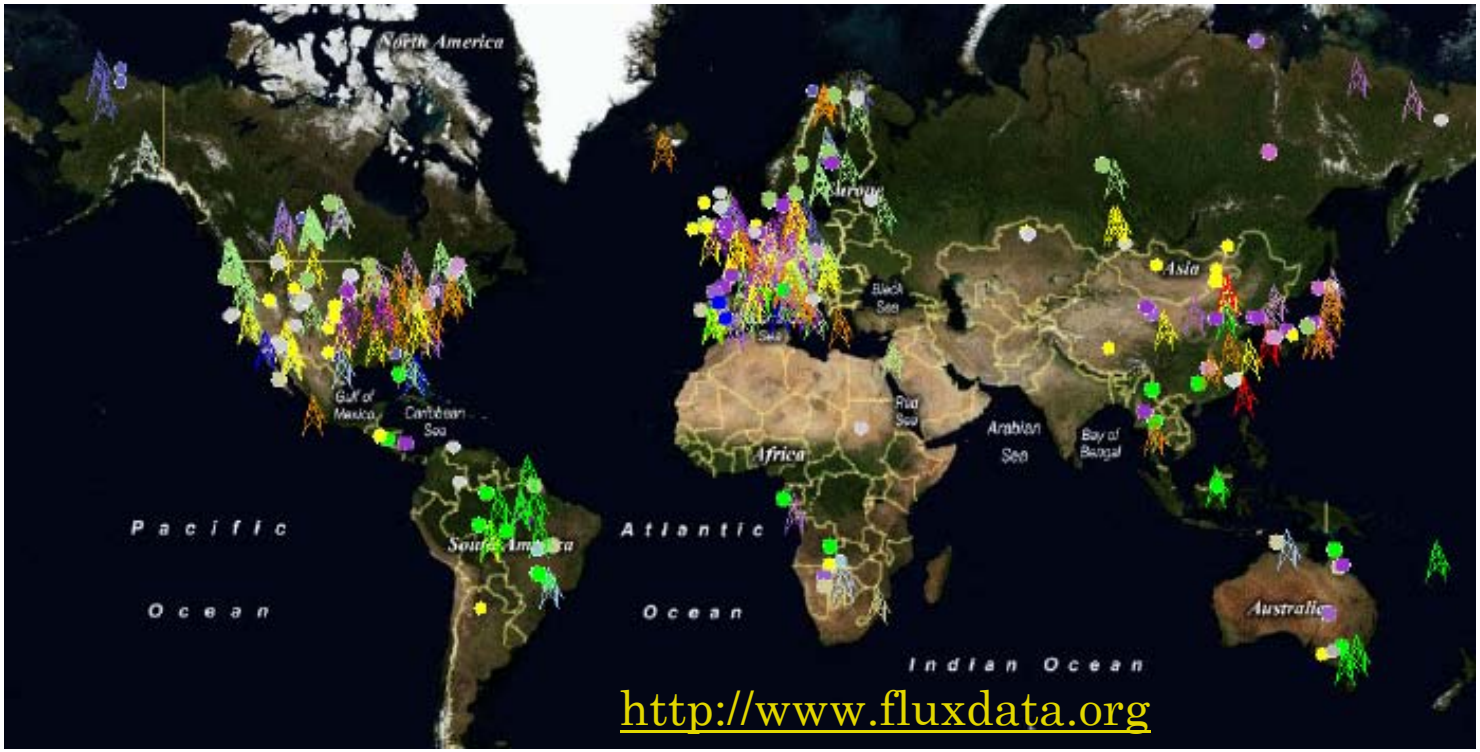
- Very hard won
  - Human time and skill
  - Lab costs can be considerable
  - Gleaning from literature (and cross checking!)
- Very small
  - Dwarfed by data from even slow sensors
  - Often passed around in email
- Very different usage patterns
  - Constant location attributes or aliases
  - Time series via splines or step functions
  - Filters for sensor data: periods before or summer LAI > x, etc
  - Time benders: "since <event>"
  - Spatial aggregation via models
- Often requires science judgment
  - Different scientists don't always agree
  - Anecdotal reporting difficult to interpret

Curating Ancillary Data



- But that's no different from curating sensor data!
- How hard can it be?
- What could possibly go wrong?

FLUXNET : A Network of Networks

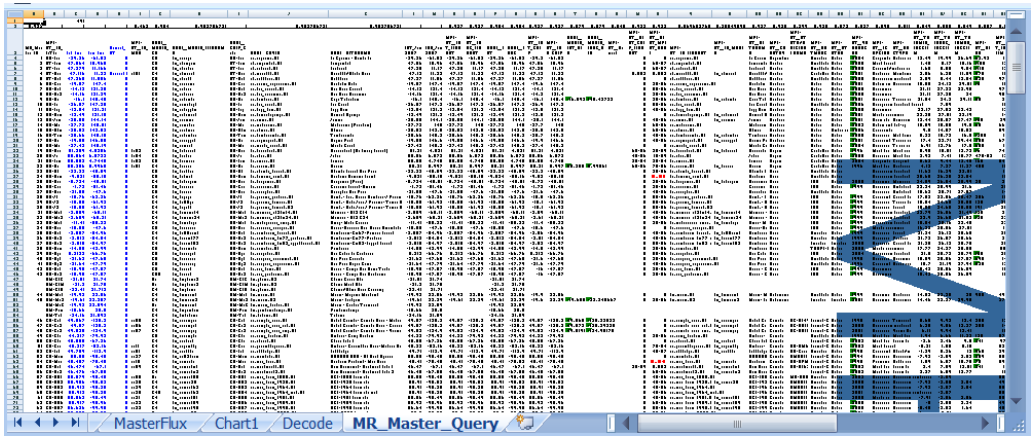


- 467 towers world wide
- 967 site-years of sensor data from 253 towers
- ~20 sensor measurements per tower; 20 derived science variables
- 145 ancillary variables

- Original data set assembled and processed in 2007
- 20x larger than previous synthesis dataset
- 75 paper teams with over 200 scientists

Ancillary Data at LaThuile (2007)

- Original collection amassed from 3 private collections and "CEIP" templates
  - One spreadsheet with 76 columns and 458 rows created by cut/paste transcription
- Major effort to determine the list of tower sites with latitude and longitude



5 clustered Zotino Sites

Ancillary Data at LaThuile

- Some "controlled" vocabularies
- Different species names for the same plants
- Profiles by depth or by soil horizon
- Disturbances free form text only
- Data reported in different units
- Initial spreadsheet passed around and updated at the conference caused immediate variants

VEG_TYPE (Controlled vocabulary)

- Grassland
- The wetland characteristic of the region is an aquatic zone that appears temporarily covered with stagnant water of natural regime, of little depth, characterized by the presence of summer herbaceous vegetation.
- 30% areal coverage, mesquite 3-4 m high

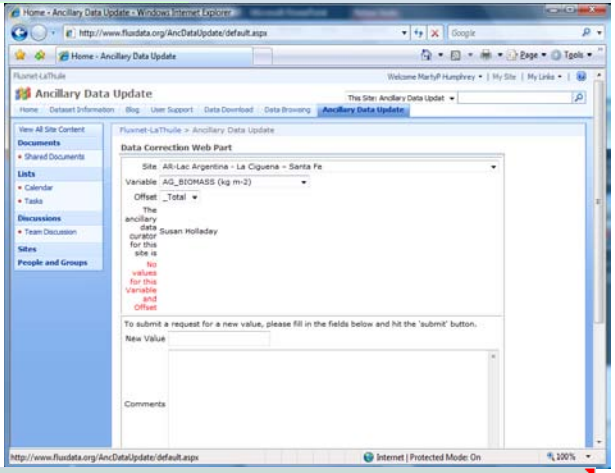
CANOPY HEIGHT (in meters)

- 15.9
- Herbaceous 0 – 1.5 m; Shrubs 0.5 – 3 m; Trees up to 10 m
- maximum grass height in the peak growth period (late April to early May) could reach up to 55 +/-12

BADM Reporting Protocols (2008)

- Developed after LaThuile reporting experience and with some specific analyses in mind
  - Joint effort by scientists from different tower types and computer scientists
- Different reporting practices for different plant types
- Specification of field methodologies, units, and reporting frequencies
- Free form comment fields in addition to data fields
- Still a work in progress

HEIGHTC	Mean canopy height	m
HEIGHTC_SIGMA	HEIGHTC error estimate	
HEIGHTC_DATE	Mean canopy height measurement date	DOY/YYYY
HEIGHTC_COMMENT	Mean canopy height comments	

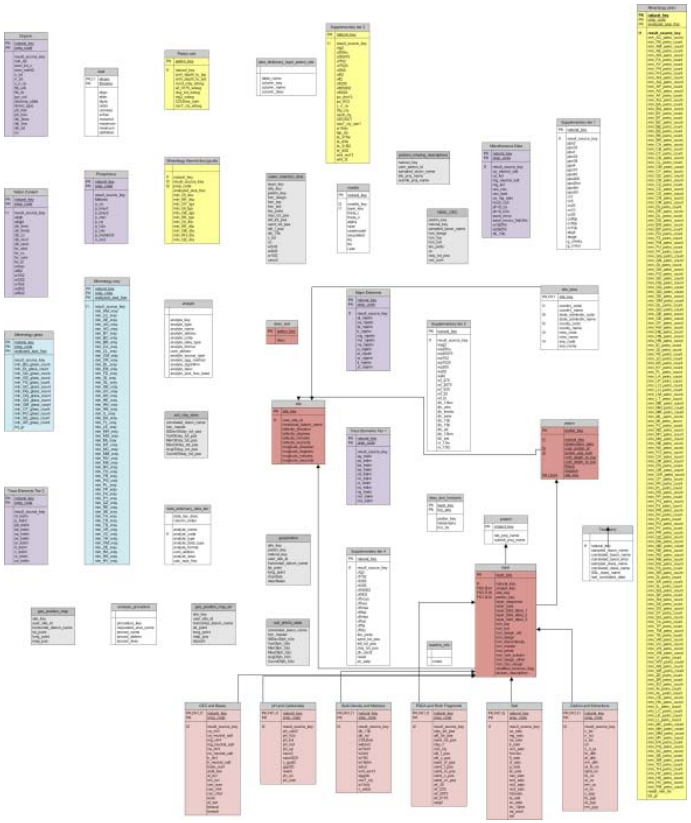


DIST<n>	Site disturbance history code
DIST<n>_QUAL	Site disturbance code qualifier
DIST<n>_DATE	Date of site disturbance in mm/dd/yyyy format
DIST<n>_DATE_QUAL	Date of site disturbance qualifier
DIST<n>_DATE_QUAL2	Date of site disturbance qualifier
DIST<n>_COMMENT	Disturbance comments

FAO protocols (Law et al 2008)

“Access” Database Approach

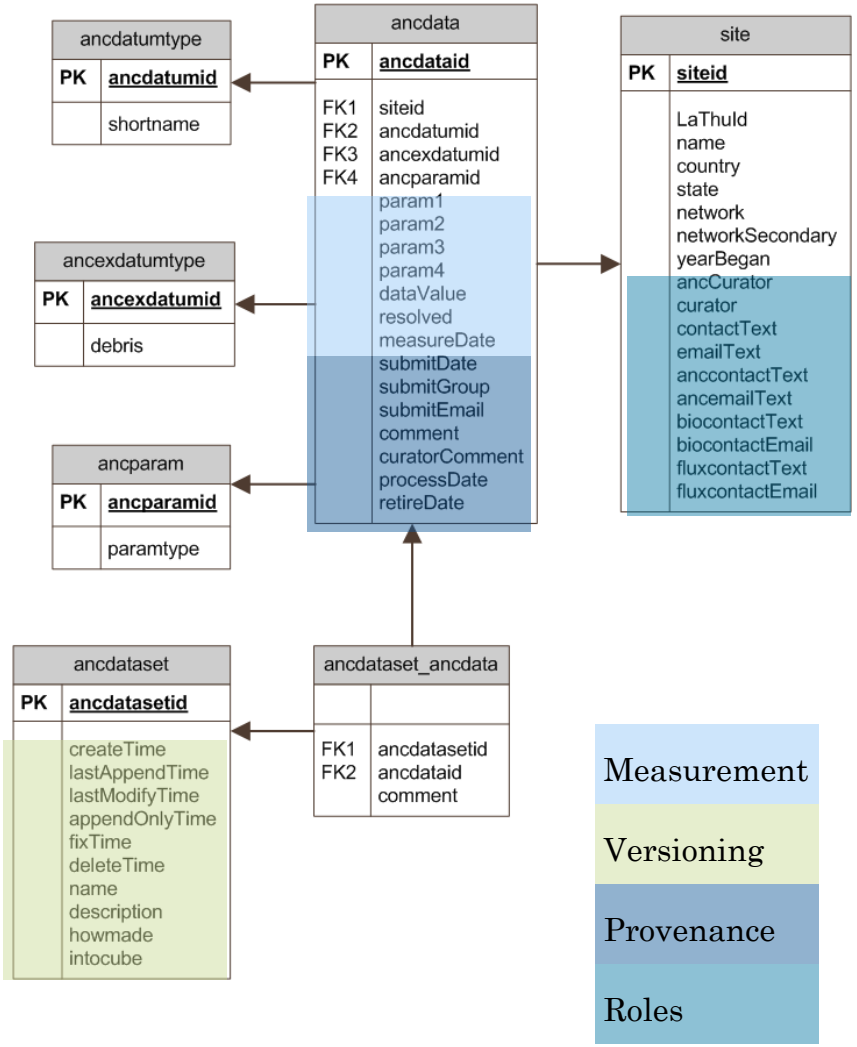
- Tables group related variables
  - Grouping often determined by the nature of the sampling or laboratory processing
  - Example: describing a fire or harvest is different from recording a hobo water temperature sample
- Data held in natural data type – real, datetime, fixed text strings
- Some provenance information, but no ability to track versions, corrections, or conflicting observations
  - Often buried in a large text string
- Conceptually simple, but difficult to view the dataset as a whole



“You are in a maze of twisty little tables, all alike”

BWC Database Approach (1/2)

- Mostly normalized table with indirect variable-specific interpretation
  - Effectively a log of submissions and corrections
  - Cleaning reprocessing that does not result in a change tracked
- Common per measurement provenance for corrections and conflicts
  - Who/what/when submitted
  - When processed
  - When superceded or deleted from current use
- Many:many data release versioning and other folder like groupings via the dataset



"One Ring to find them, One Ring to bring them all"

BWC Database Approach (2/2)

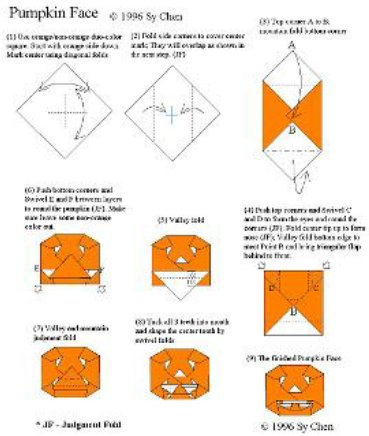
- All variables are characterized by:
  - Data value
  - 0-4 variable specific parameters
  - Variable type and extended type
  - (Optional) controlled vocabulary
  - Optional measurement date
  - Metadata including units, data type, addition/retire dates, description, etc.

ancdecode
rowtext variableType reportingType dataType units description explanation shortname debris paramtype isDate isSpecies isSurvivor hasDate hasParam1 hasParam2 hasParam3 hasParam4 hasComment hasCV docSequence docAddDate docRetireDate ancdatumid ancexdatumid ancparamid inSummaryReport

- Contained in a single translation table with one row for each parsing token
  - Documentation, web forms, data product creation, and data ingest all driven from this table
- Simple controlled vocabulary can be replaced when a richer solution needed

ancCV
vocabulary shortname description hasQual qualIsValue qualCV alias hierarchy

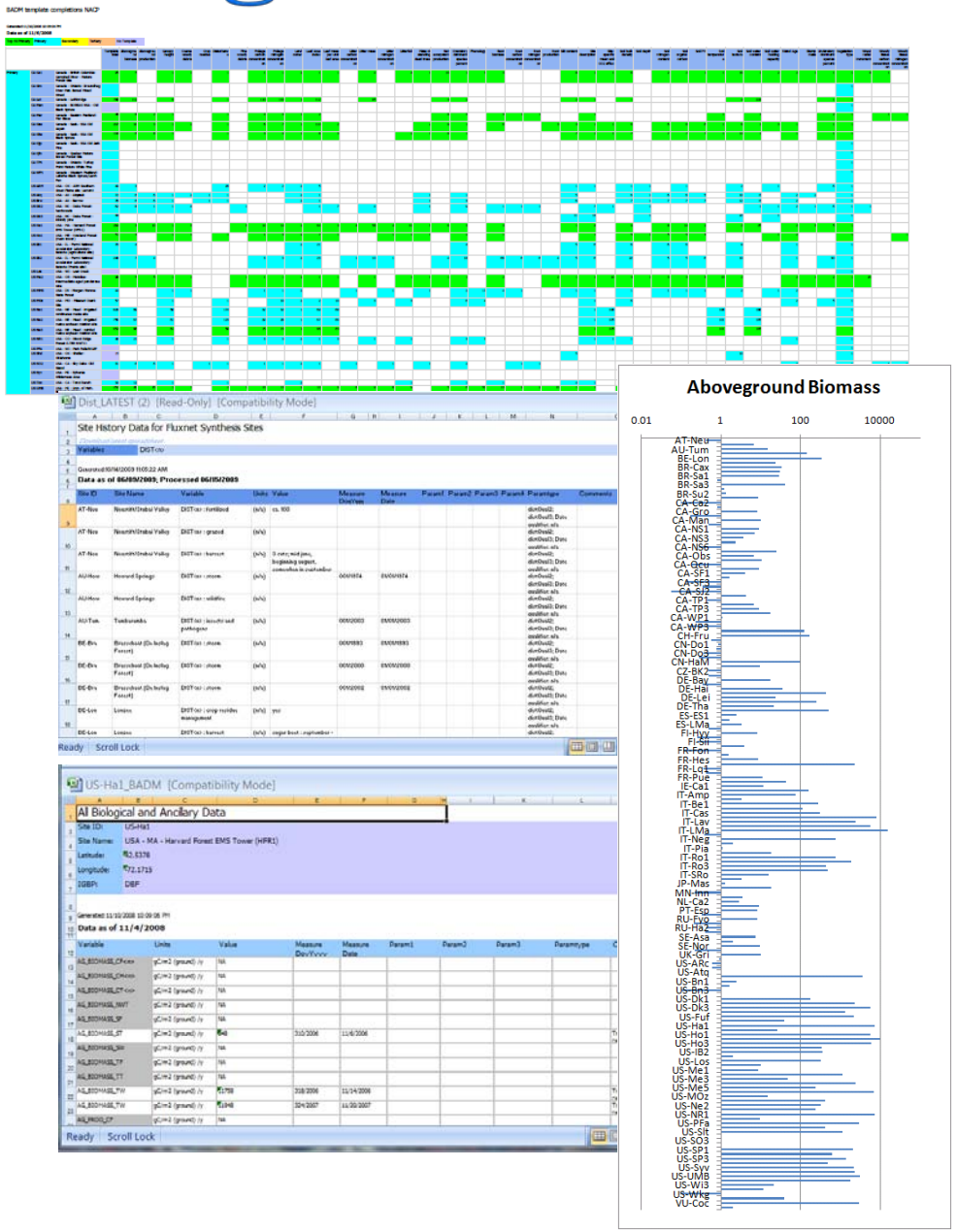
Archive Database Operations



- Templates and other spreadsheets are shredded to extract each active cell
- Related cells are then folded together often replicating cells
- Data cleaning happens both before and after folding.
  - Basic reporting protocol, format, and some value limits checked by computer science curator
  - Science data validity checked by science collaborator
- Folded data grouped into “handy” sets for exported reports
- Data cube for coverage browsing

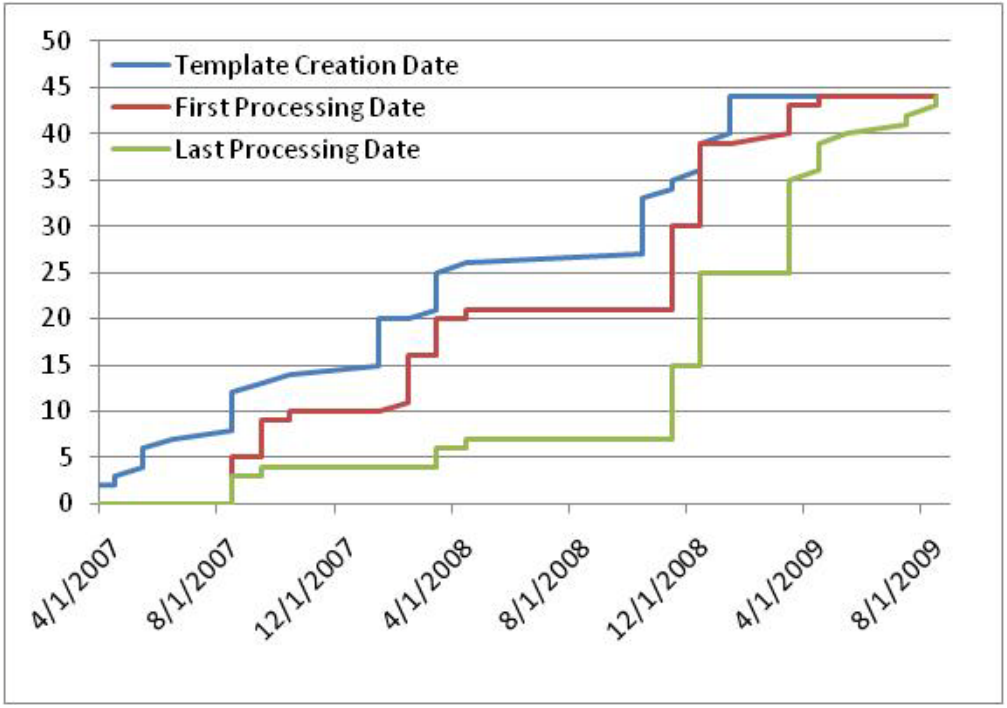
Approach Learnings

- Scientists (and some computer scientists) find the normalization counterintuitive
- More upfront work in data modeling
- Longer term maintenance much easier
  - We recently added param4 to hold site error estimate
- Producing data products and manual checks are simpler because data set can be viewed as a whole



BADM submissions (AmeriFlux)

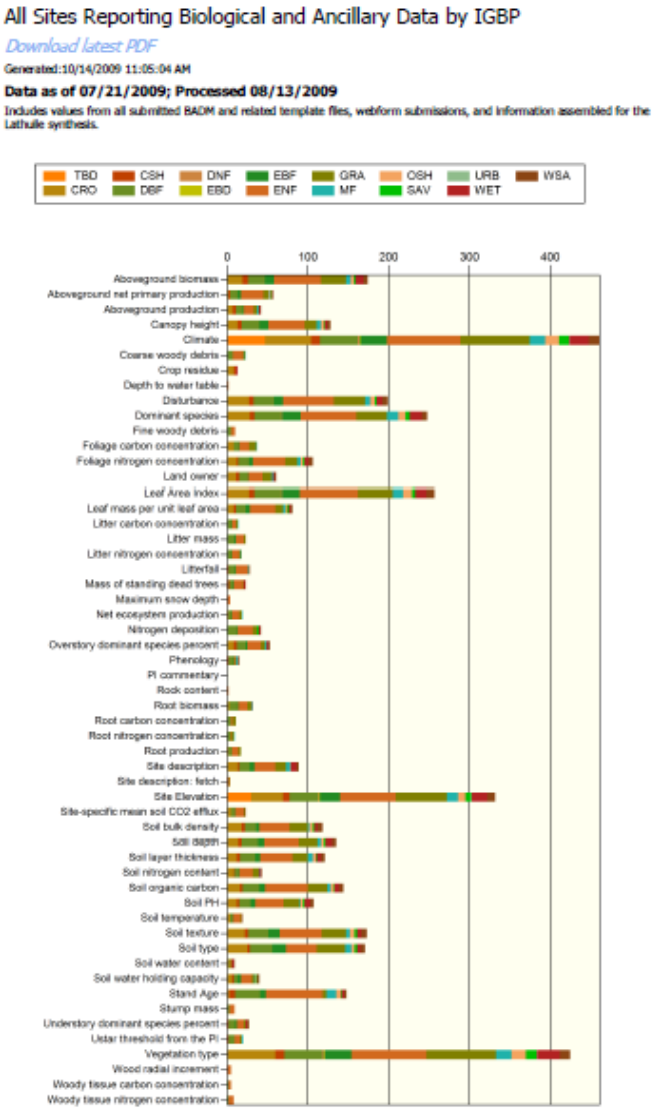
- 44 of 120 sites have submitted templates
- The differences between dates are due to time lags for updated/ corrected templates and web corrections.
- Each template has been processed at least twice (even the good ones).
- The NACP BADM campaign are the Nov 2008 submissions March 2009 processing bump which included many new templates, corrections, and updates.



- Other bumps are caused by tower teams reporting all sites at the same time (e.g. US-Me* or US-SO*).

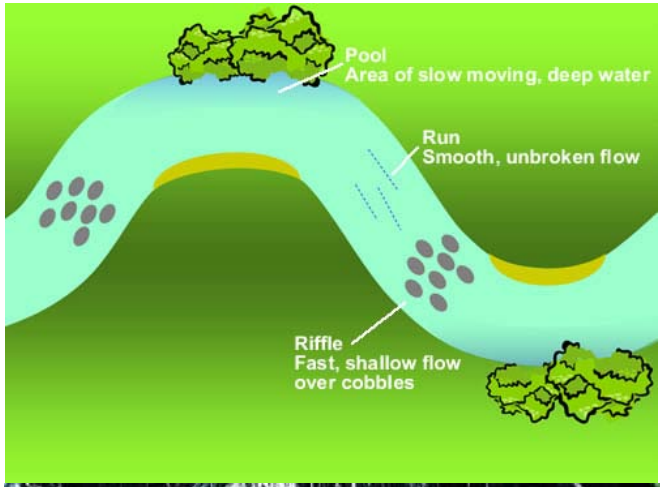
Learnings To Date

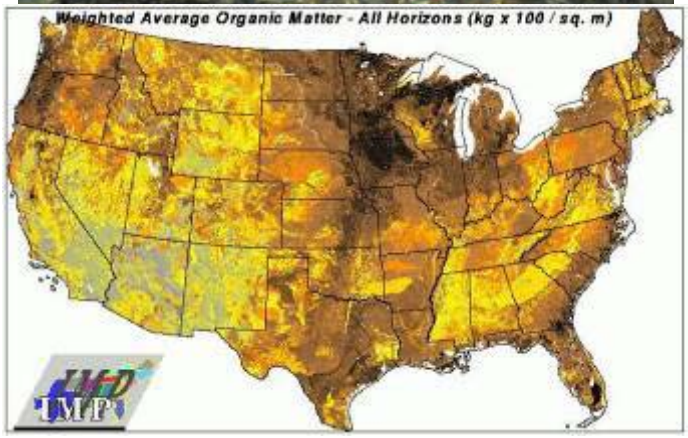
- FLUXNET is not one "collaboration" but several different sub-groups with different cultures.
  - "Color outside the lines"
  - Self-coordination
  - Publish data vs provide data access
- Partial normalization and logging are goodness
- For many sites, the template becomes THE repository
  - Plan on reprocessing with "is that an update or an error?"
- Science curators at best sanity check (and sometimes rubber stamp) yet the science users expect clean science ready data!
- Campaigns driven by specific papers or analyses are the best motivation



Futures

- Currently using the same approach with some additions
  - National Marine Fisheries "Hab8" stream habitat data
  - National Soil Carbon Network
  - National Phenology Network (citizen science)
- How do we structure the next FLUXNET ancillary data campaign?
- Find a technology to enable a scientist to do the "access database" mapping to normalized database to lower barrier to entry





Acknowledgements

- Berkeley Water Center, University of California, Berkeley, Lawrence Berkeley Laboratory
  - Jim Hunt
  - Dennis Baldocchi
  - Deb Agarwal
  - Monte Goode
  - Keith Jackson
  - Rebecca Leonardson (student)
  - Carolyn Remick
  - Susan Hubbard
- University of Virginia
  - Marty Humphrey
  - Norm Beekwilder
  - Jie Li
- Microsoft Research
  - Bora Beran
  - Yogesh Simmhan
  - Andy Sterland
  - Scott Counts
  - Tony Hey
  - Dan Fay
- San Diego Supercomputing Center
  - Ilya Zavlavsy
  - David Valentine
- University of Queensland
  - Jane Hunter
- Ameriflux Collaboration
  - Beverly Law
  - Tara Hudiburg (student)
  - Gretchen Miller (student)
  - Andrea Scheutz (student)
  - Christoph Thomas
  - Hongyan Luo (postdoc)
  - Lucie Ploude (student)
  - Andrew Richardson
  - Mattias Falk
  - Tom Boden
- Fluxnet Collaboration
  - Dennis Baldocchi
  - Rodrigo Vargas (postdoc)
  - Youngryel Ryu (student)
  - Dario Papale (CarboEurope)
  - Markus Reichstein (CarboEurope)
  - Hank Margolis (Fluxnet-Canada)
  - Alan Barr (Fluxnet-Canada)
  - Bob Cook
  - Susan Holladay
  - Dorothea Frank
- North American Carbon Program
  - Kevin Schaefer
  - Peter Thorton