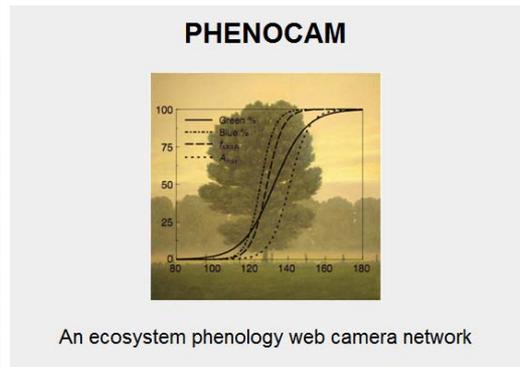




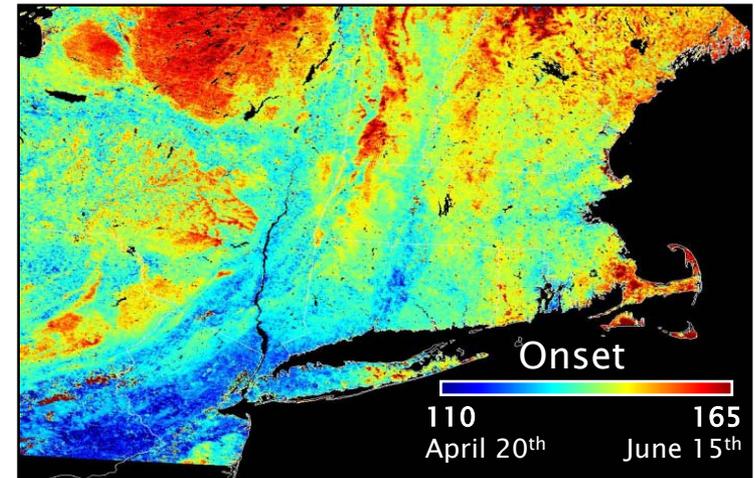
Beyond Sensors: Curating Ancillary Data for Carbon Climate Science

Catharine van Ingen and Deb Agarwal
Microsoft eScience Workshop
October 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 the FLUXNET ancillary data



Home - Fluxnet-LaThuile

Fluxnet-LaThuile

Documents

Links

Discussions

Sites

People and Groups

Welcome to the Web Site for the FLUXNET Synthesis Data set.

Announcements

Fluxnet synthesis dataset highlighted on the Microsoft Science website

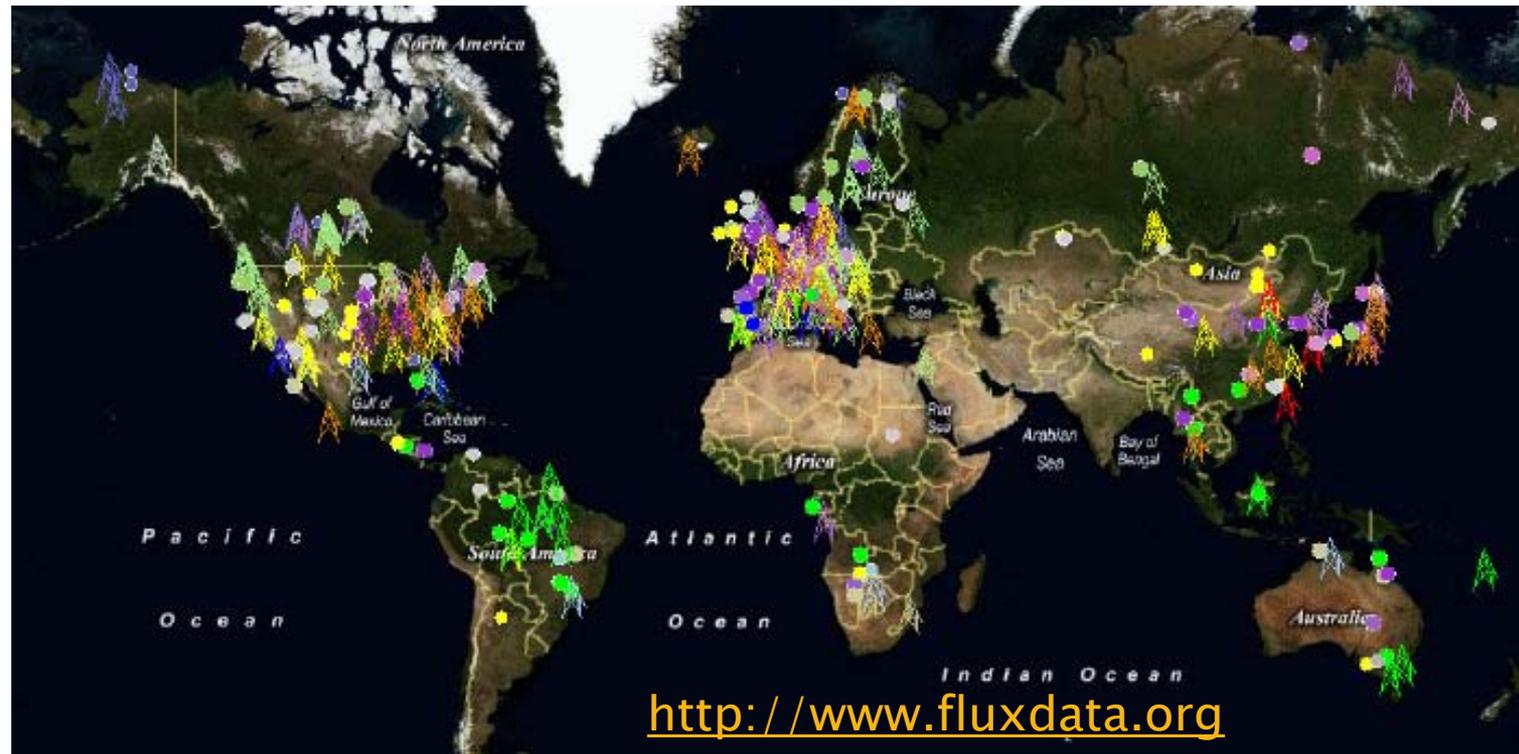
Fluxnet synthesis dataset and the server that supports it are currently highlighted on the Microsoft Science website.

Fluxletter Volume 1, Issue 1 now available

Fluxnet related papers posted

How to Request an Account

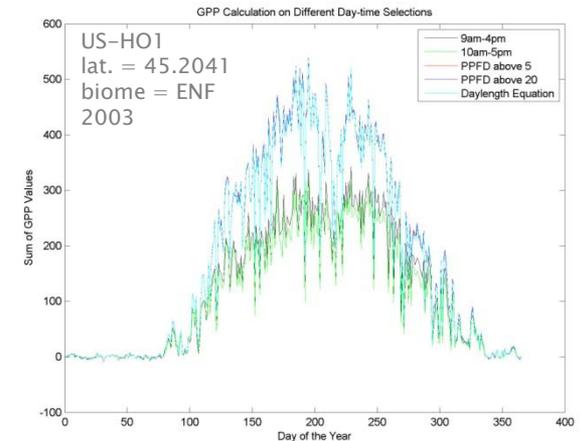
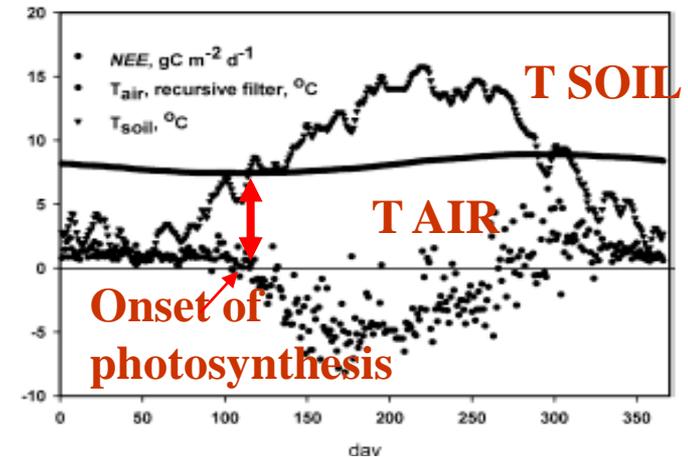
FLUXNET : A Network of Networks



- ▶ 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

Environmental Sensor Data

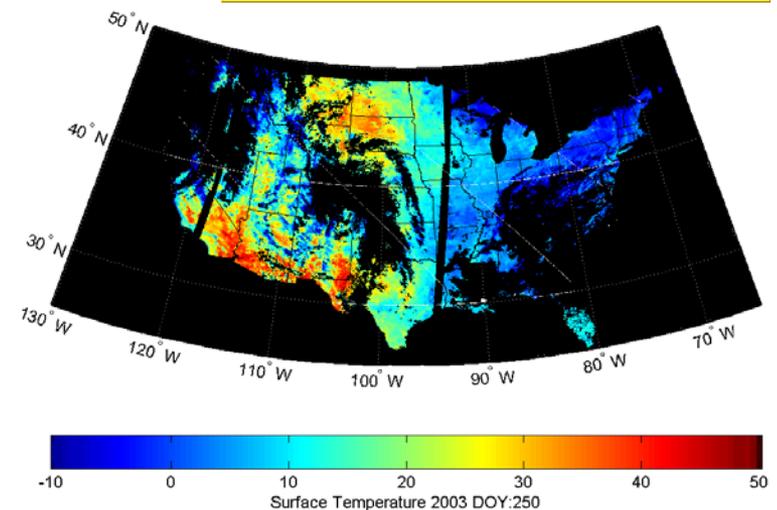
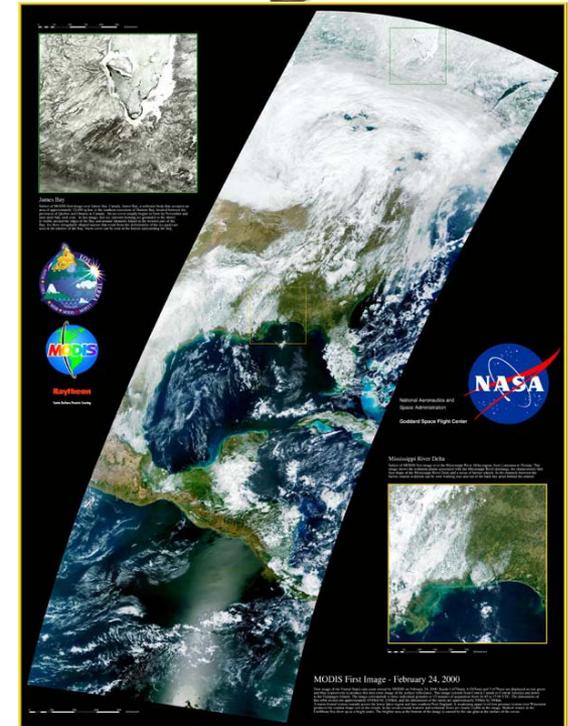
- ▶ Time series data
 - Over some period of time at some time frequency at some spatial location.
 - May be actual measurement (L0) or derived quantities (L1 +)
- ▶ (Re)calibrations, gaps and errors are a way of life.
 - Birds poop, batteries die, sensors fail.
 - Various quality assessment and signal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



“Time is not just another axis”

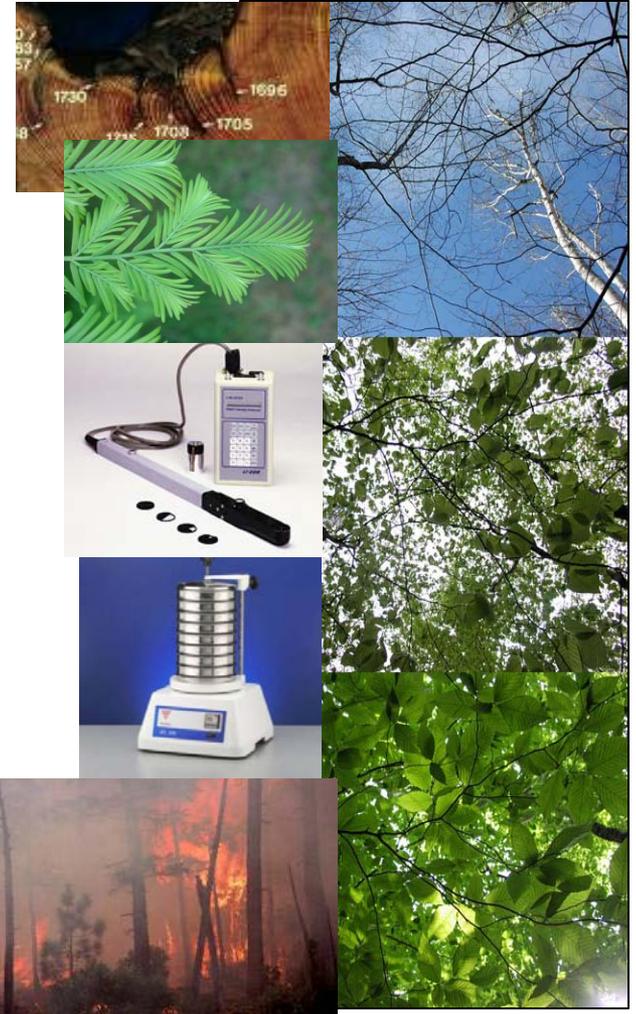
Environmental Remote Sensing Data

- ▶ 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 create virtual sensors
- ▶ Today: PBs (L0) to TBs (L2+)



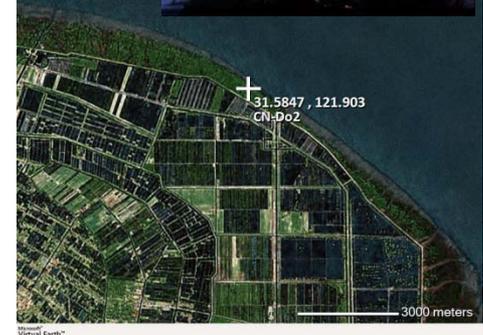
Environmental Ancillary Data

- ▶ Almost everything else!
 - ‘Constants’ such as latitude or longitude
 - Intermittent measurements such as LAI (leaf cross-sectional area) or soil grain size distribution
 - Anecdotal descriptions
 - Events such as bud break 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.

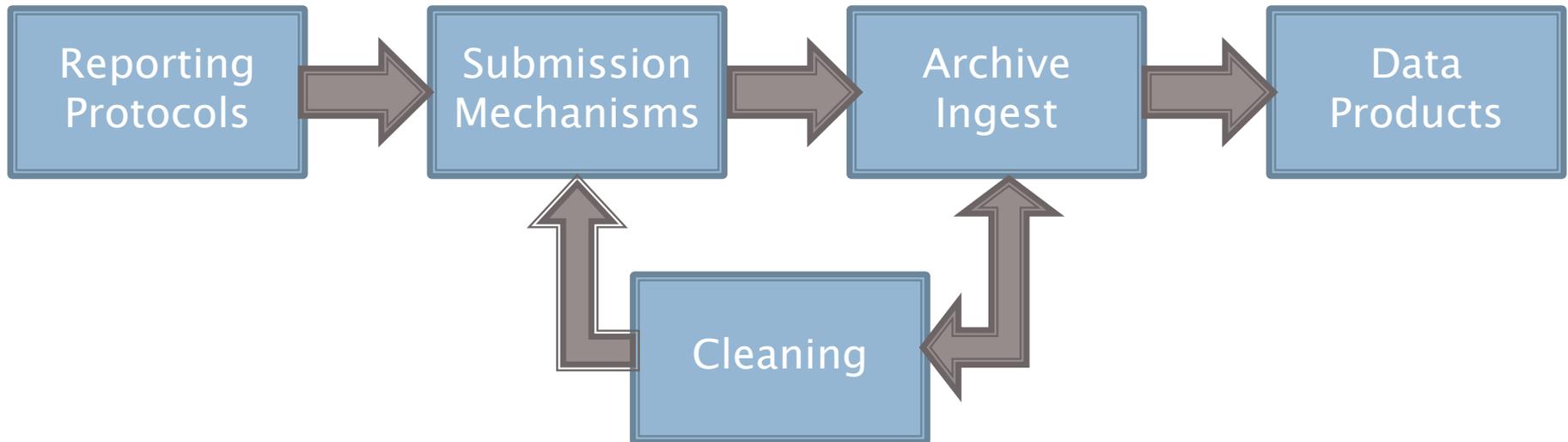


Ancillary Data is Different

- ▶ Very hard won
 - Dig a pit or shoot an air rifle to get samples
 - Lab costs can be considerable
 - Gleaning from literature (and cross checking!)
- ▶ Very small
 - FLUXNET collection is currently ~30K numbers.
 - 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 after, sites with 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



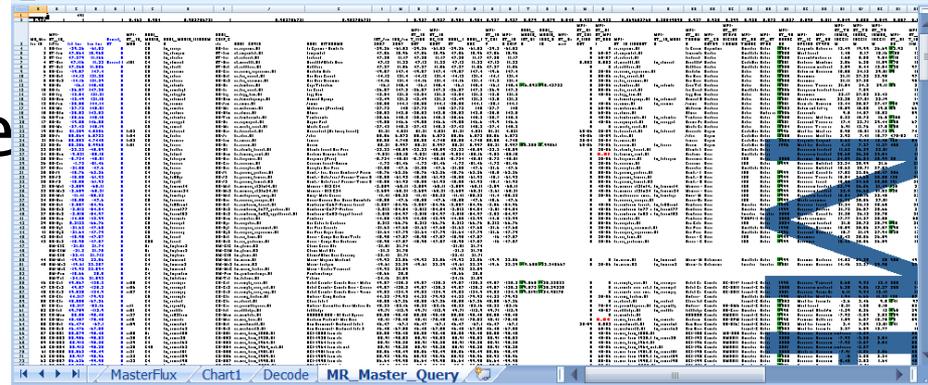
Curating Ancillary Data



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

Reporting Protocols at LaThuile

- ▶ 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



A screenshot of a Microsoft Excel spreadsheet. The spreadsheet is filled with data organized into columns and rows. The columns are labeled with various identifiers and names, and the rows contain numerical and text data. The spreadsheet is viewed from a top-down perspective, showing the grid structure and the data entries within each cell. The interface includes the standard Excel menu bar and toolbar at the top, and the status bar at the bottom.



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

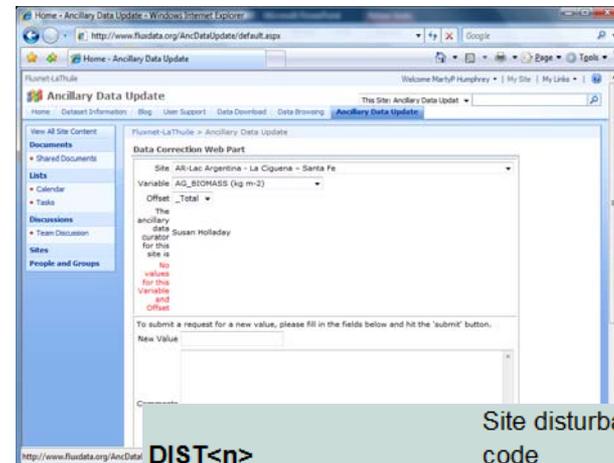
CANOPY HEIGHT (in meters)

- 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

- ▶ 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

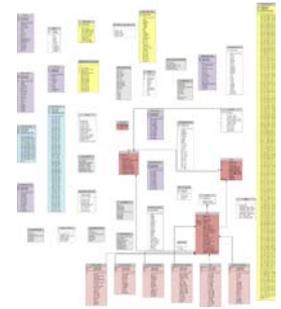
HEIGHTC	Mean canopy height	m
HEIGHTC_SIGMA	HEIGHTC error estimate	
HEIGHTC_DATE	Mean canopy height measurement date	DOY/YYYY
HEIGHTC_COMMENT	Mean canopy height comments	



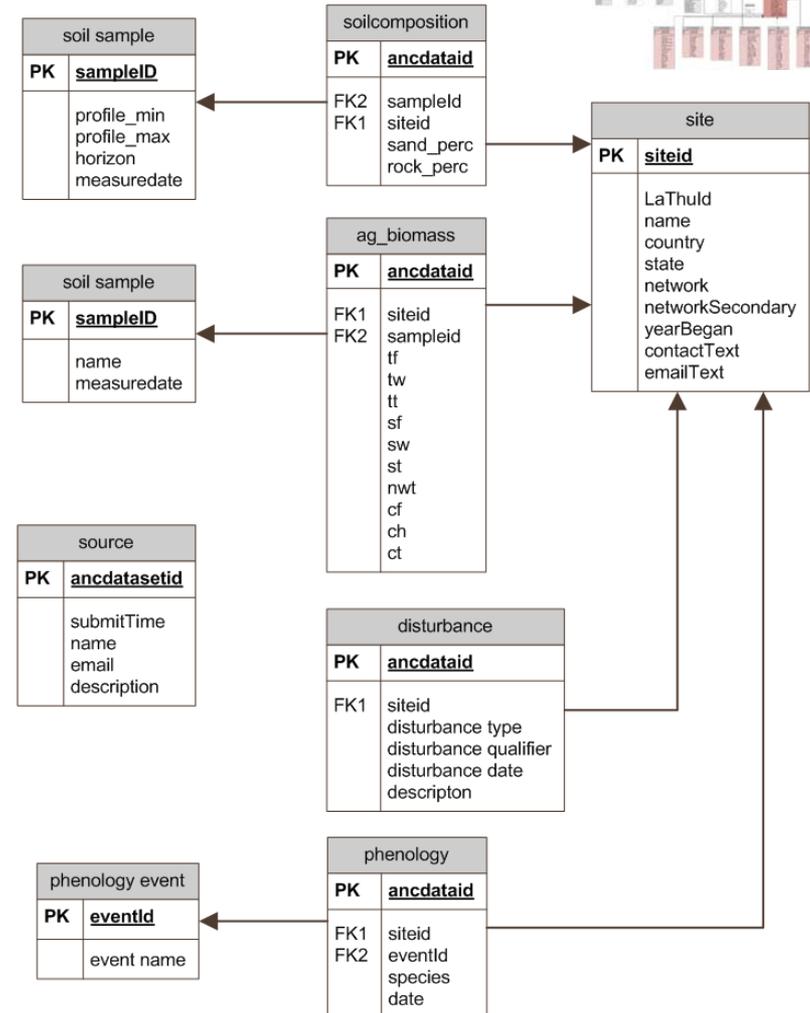
DIST<n>	Site disturbance history code
DIST<n>_QUAL	Site disturbance code qualifier
DIST<n>_DATE	Date of site disturbance in mm/dd/yyyy format
DIST<n>_DATE_QUAL	Date of site disturbance qualifier
DIST<n>_DATE_QUAL2	Date of site disturbance qualifier
DIST<n>_COMMENT	Disturbance comments

[FAO protocols \(Law et al 2008\)](#)

“Access” Database Approach



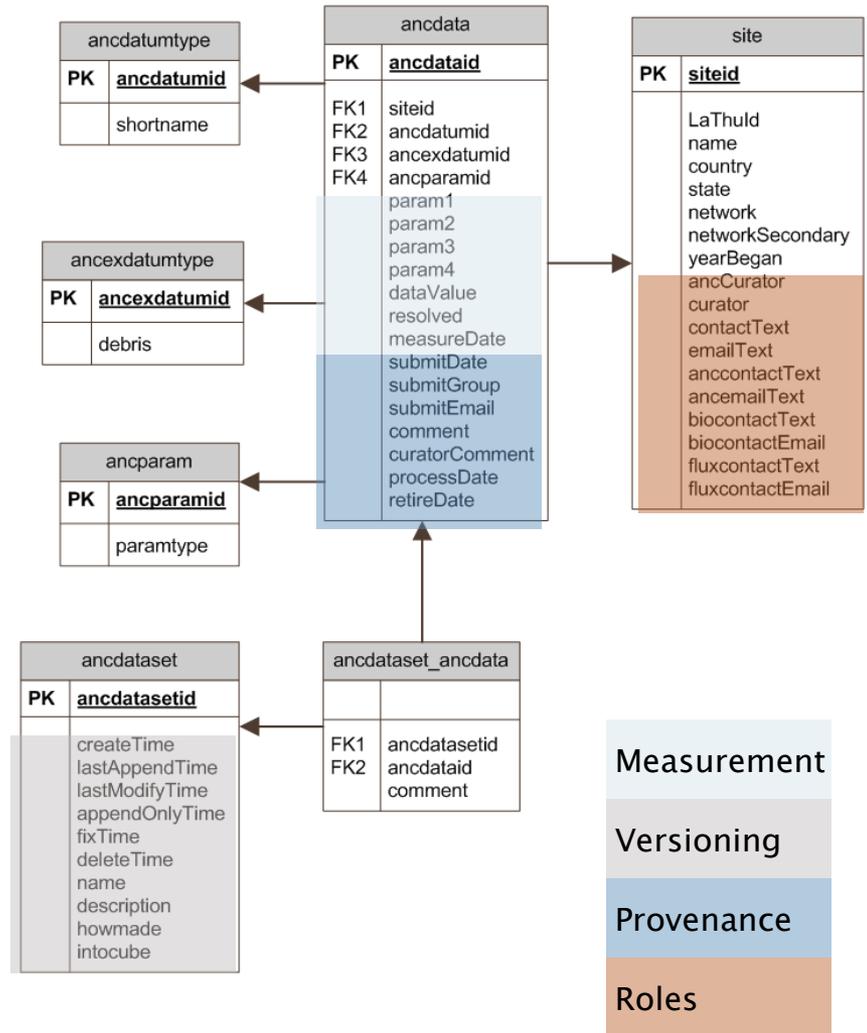
- ▶ Tables group related variables
 - Grouping often determined by the nature of the sampling or laboratory processing
- ▶ Data held in natural data type – real, datetime, fixed text strings
- ▶ Source provenance information, but no ability to track versions, corrections, or conflicting observations
- ▶ Conceptually simple, but difficult to view the dataset as a whole



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

Fluxdata Database Approach

- ▶ 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”

Fluxdata Translation Structures

- ▶ 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

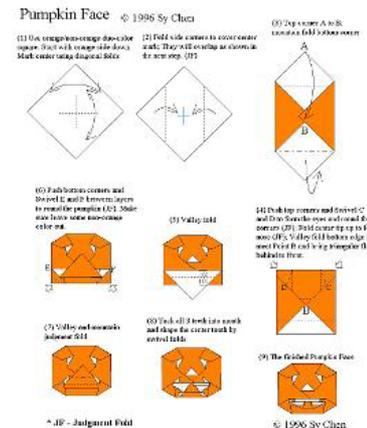
- ▶ Contained in a single translation table with one row for each parsing token
 - Documentation, web forms, data product creation, and data ingest all driven from this table

- ▶ Simple controlled vocabulary can be replaced when a richer solution needed

ancCV
vocabulary
shortname
description
hasQual
quallsValue
qualCV
alias
hierarchy

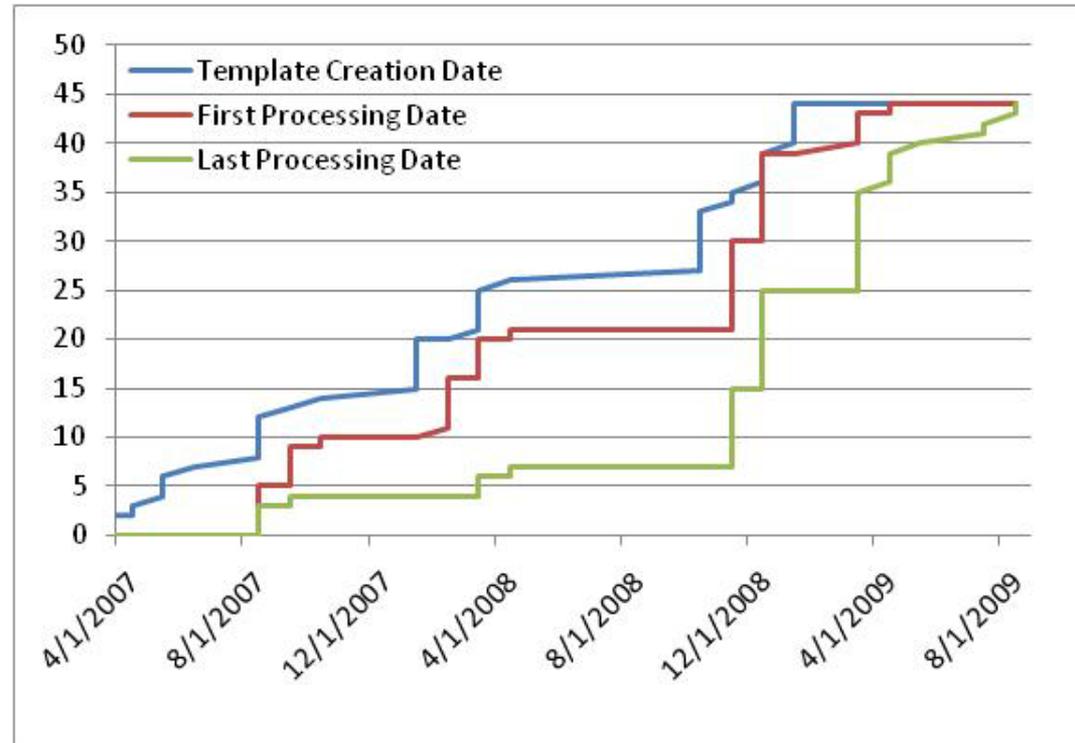
Fluxdata Archive Database Operations

- ▶ 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



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 is 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

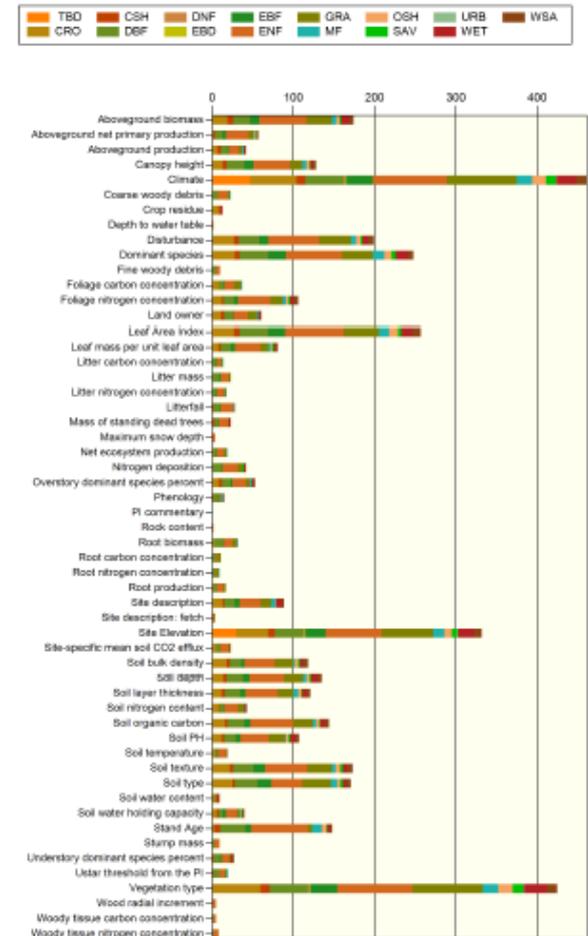
All Sites Reporting Biological and Ancillary Data by IGBP

[Download latest PDF](#)

Generated: 10/14/2009 11:05:04 AM

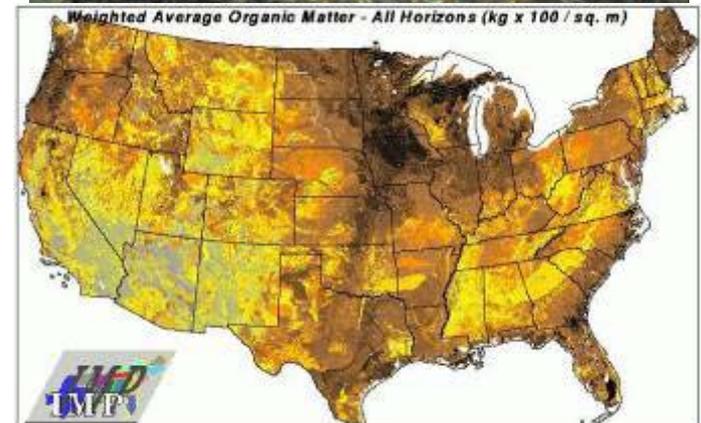
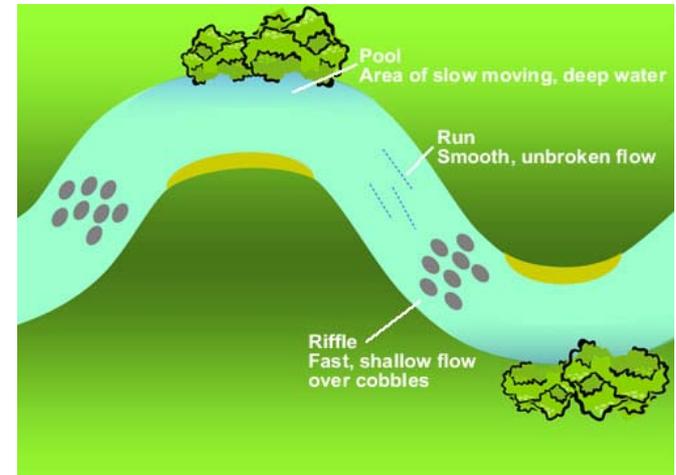
Data as of 07/21/2009; Processed 08/13/2009

Includes values from all submitted BDM and related template files, webform submissions, and information assembled for the Lathale synthesis.



Futures

- ▶ Currently using the same approach with some additions
 - National Marine Fisheries “Hab8” data
 - National Soil Carbon Network
- ▶ 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



Acknowledgements

Berkeley Water Center, University
of California, Berkeley, Lawrence
Berkeley Laboratory

Jim Hunt
Dennis Baldocchi
Deb Agarwal
Monte Good
Keith Jackson
Robin Weber
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
Jim Gray

Ameriflux Collaboration

Beverly Law
Tara Hudiburg (student)
Gretchen Miller (student)
Andrea Scheutz (student)
Christoph Thomas (postdoc)
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

Peter Thornton
Kevin Schaefer