Developing a Scientific Collaboration Data Server for Ameriflux

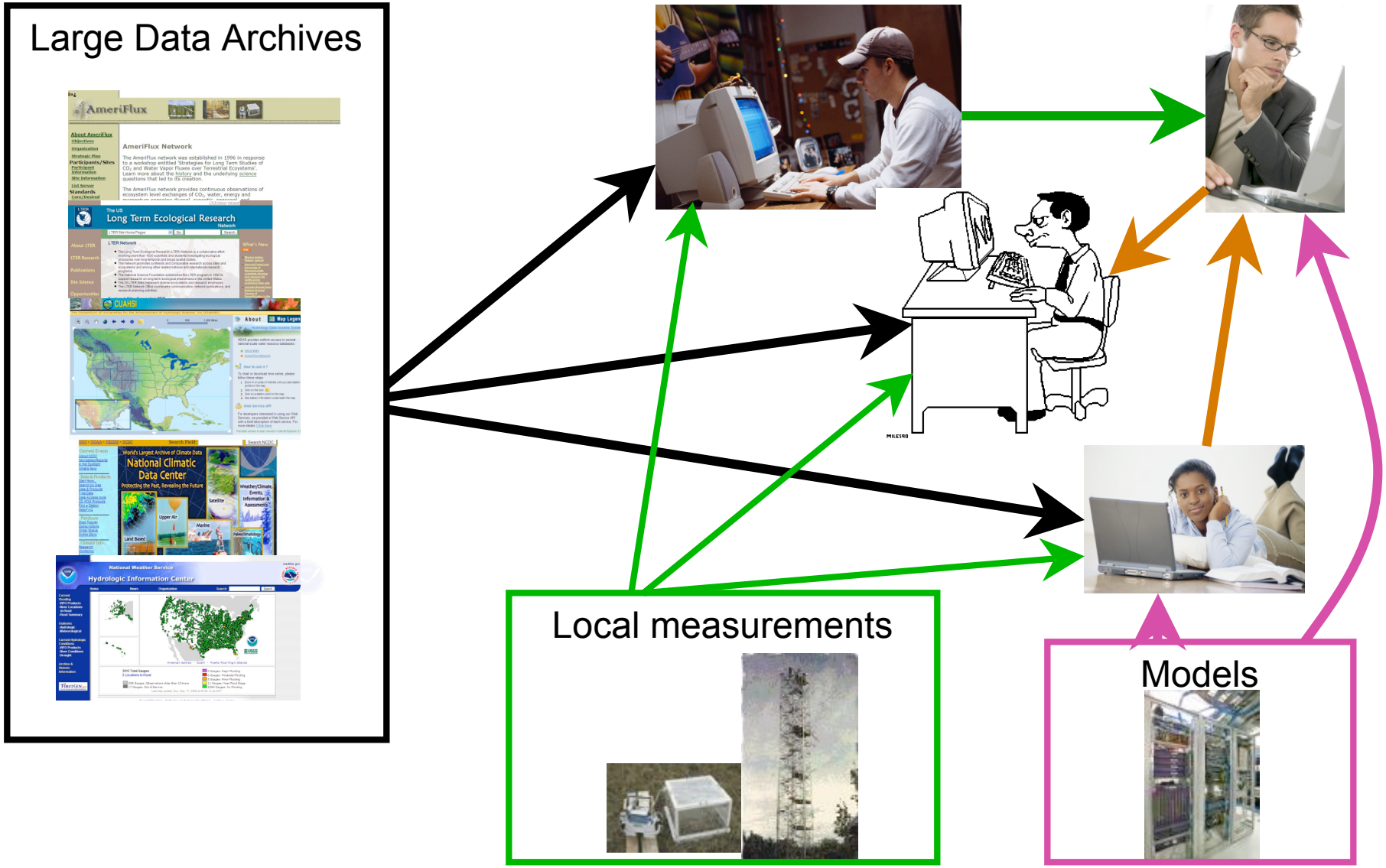
Deb Agarwal, LBL (daagarwal@lbl.gov)
Catharine van Ingen, MSFT (vaningen@microsoft.com)
Matt Rodriguez (mkrodriguez@lbl.gov)
Monte Goode (mgoode@lbl.gov)
Ameriflux Meeting – Boulder, CO -17 October 2006

Outline

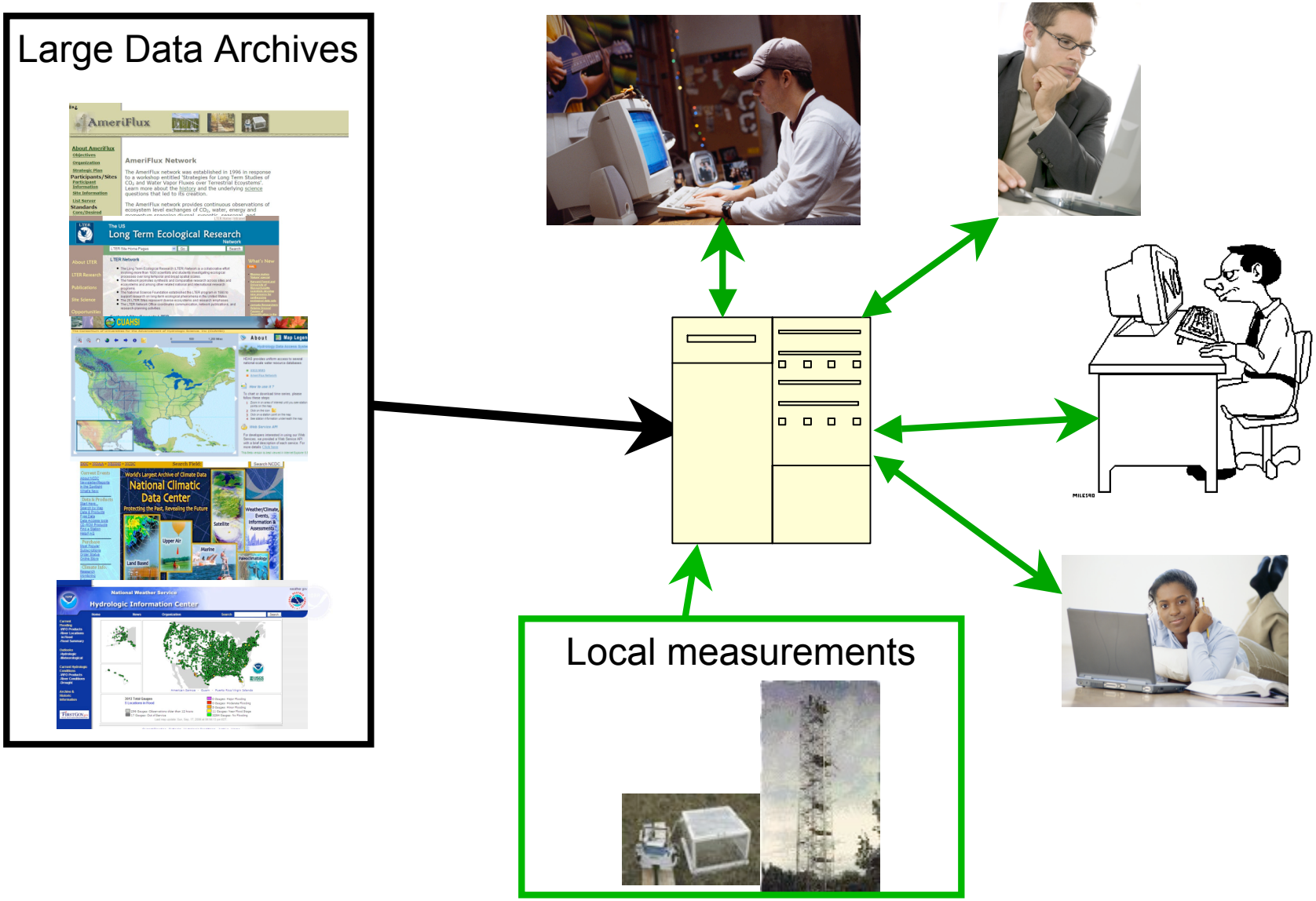
- Motivation
  - Data archives and other sources
  - Typical small group collaboration needs
  - Examples using “Ameriflux”
- Science Data Server
  - Goals and ideal capabilities
  - Approach
  - Lessons learned
- URLs
- Brief tutorial

Unprecedented Data Availability

Typical Data Flow Today



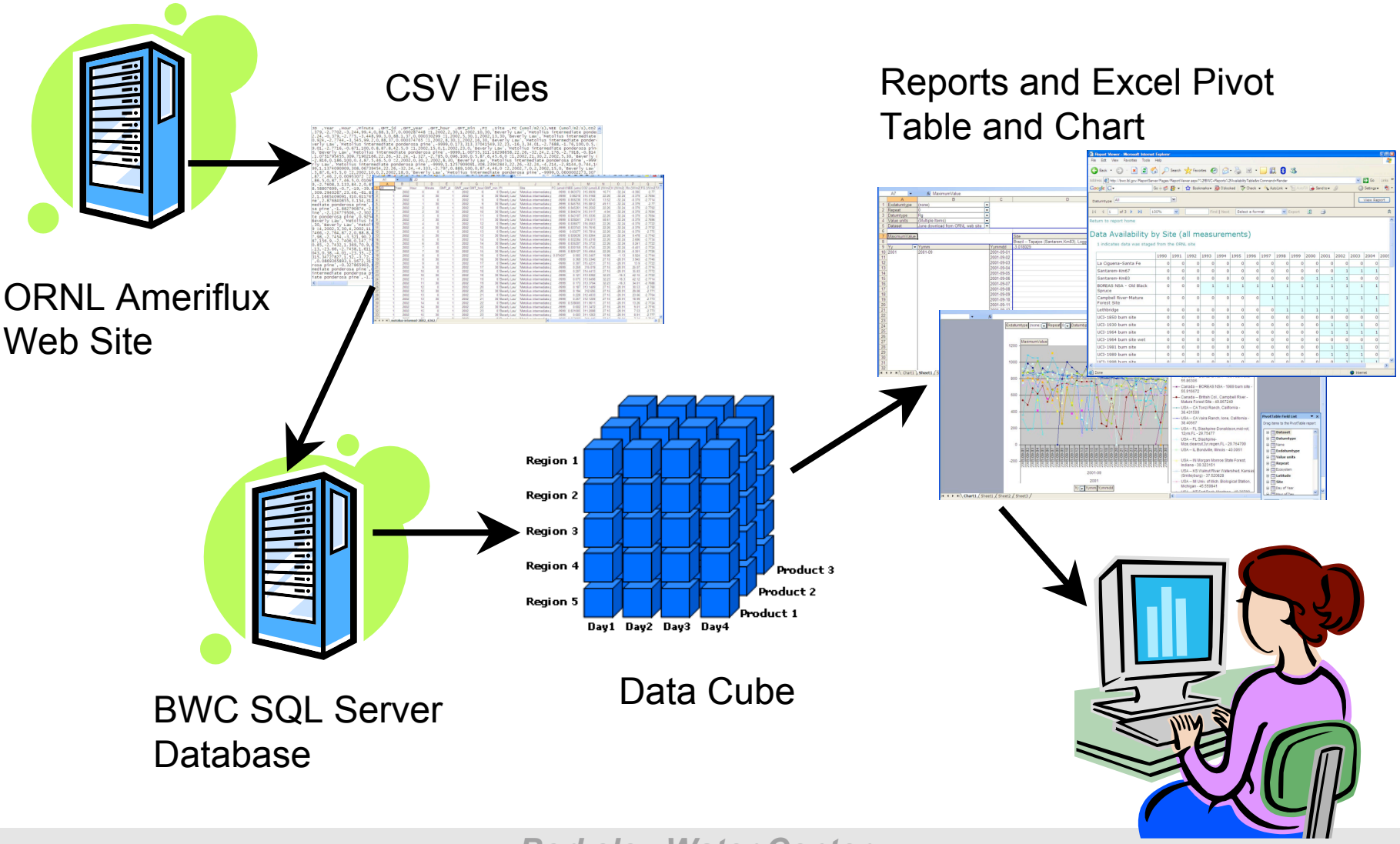
Scientific Data Server



Scientific Data Server - Goals

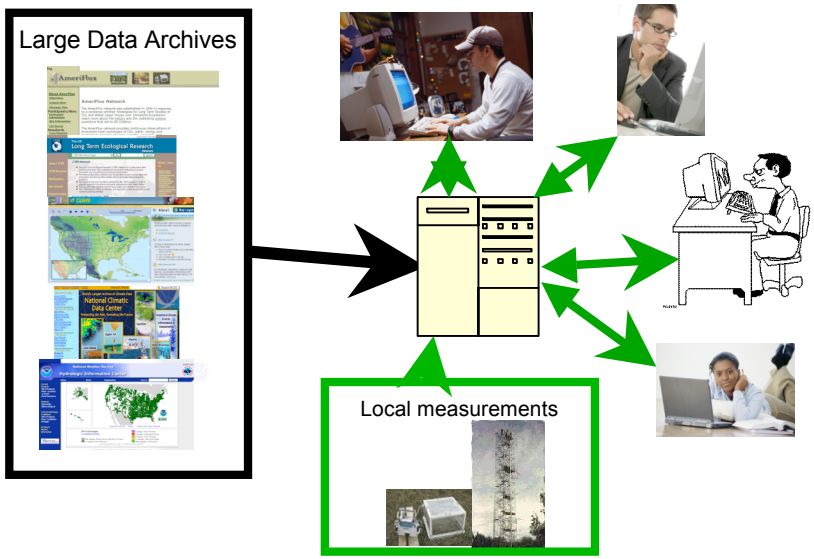
- Act as a local repository for data and metadata assembled from Ameriflux and other data sources
  - Simplify provenance by providing a common “safe deposit box” for assembled data
- Interact simply with existing and emerging Internet portals for data and metadata download, and, over time, upload
  - Simplify data assembly by adding automation
  - Simplify name space confusion by adding explicit decode translation
- Support basic analyses across the entire dataset for both data cleaning and science
  - Simplify mundane data handling tasks
  - Simplify quality checking and data selection by enabling data browsing

Data Pipeline

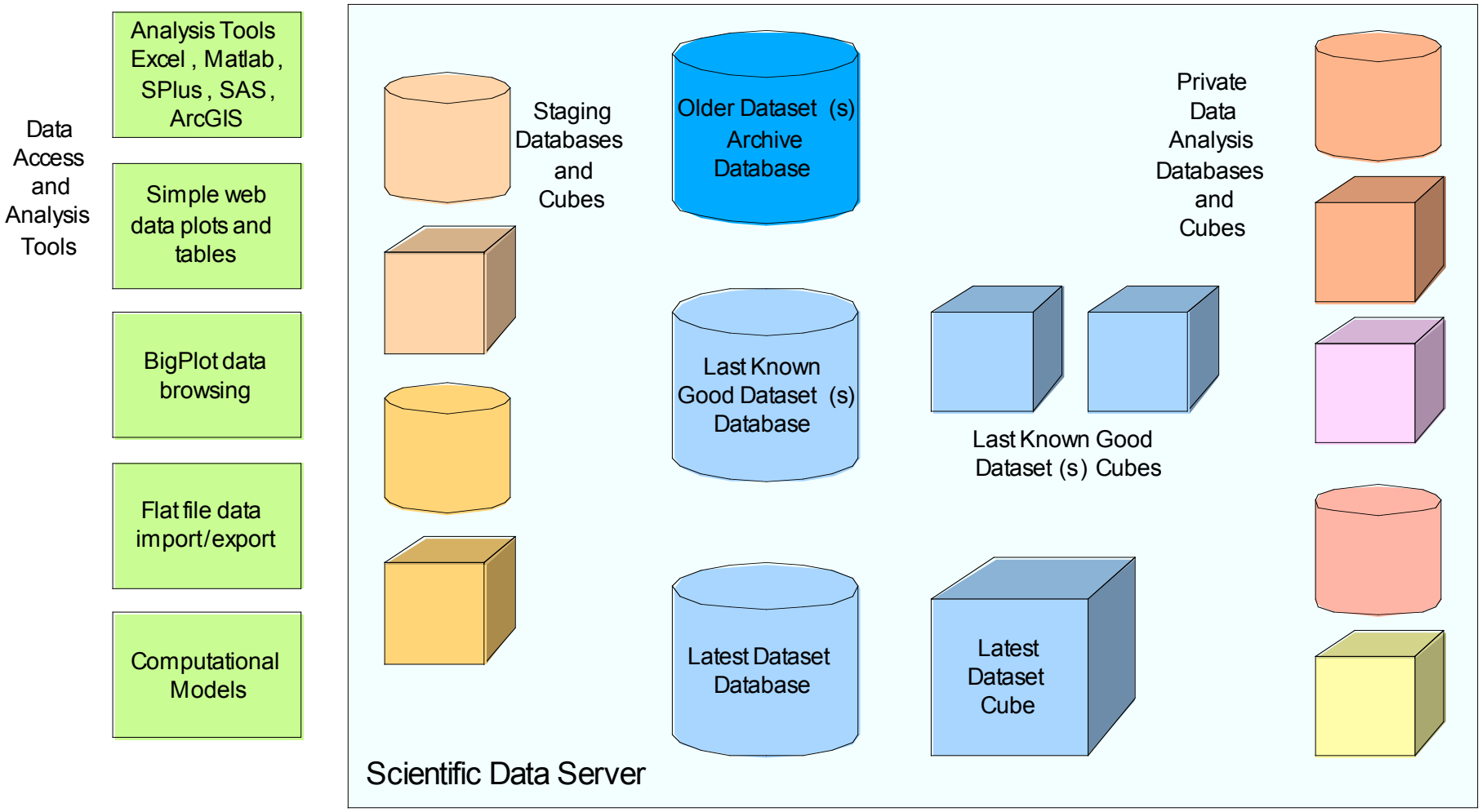


Scientific Data Server - Workflows

- Staging: adding data or metadata
  - New downloaded or field measurements added
  - New derived measurements added
- Editing: changing data or metadata
  - Existing older measurements re-calibrated or re-derived
  - Data cleaning or other algorithm changes
  - Gap filling
- Sharing: making the latest acquired data available rapidly
  - Browsing new data before more detailed analyses
- Private Analysis: Supporting individual researchers (MyDB)
  - Stable location for personal calibrations, derivations, and other data transformations
- Curating: data versioning and provenance
  - Simple parent:child versioning to track collections of data used for specific uses

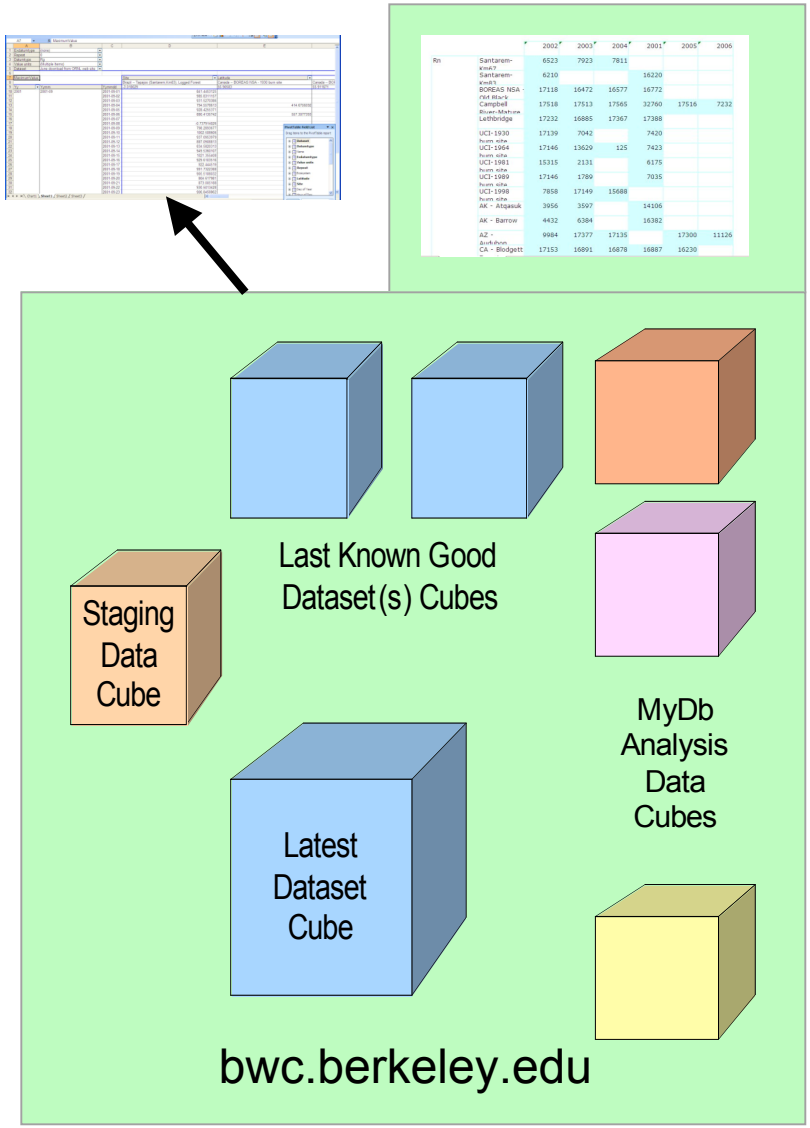


Scientific Data Server - Logical Overview



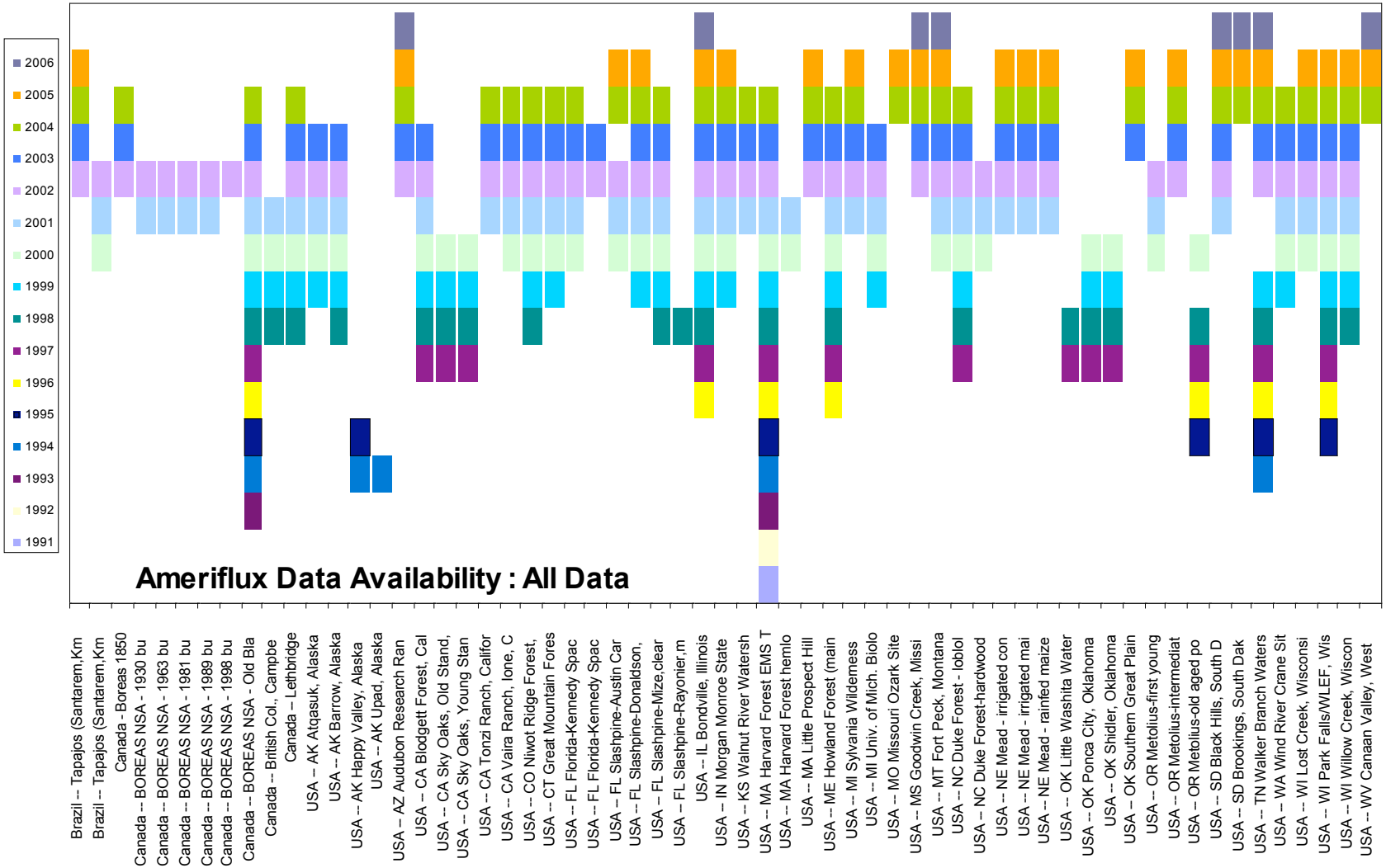
Data Cubes

- A data cube is a database specifically for data mining (OLAP)
  - Organizes data along *dimensions* such as time, site, or datumtype
  - Simple *aggregations* (sum, min, or max) can be pre-computed for speed
  - Additional calculations (median) can be computed dynamically
- Client tool integration is evolving
  - Excel PivotTables allow simple data viewing
  - Reporting Services provides web-based data browsing



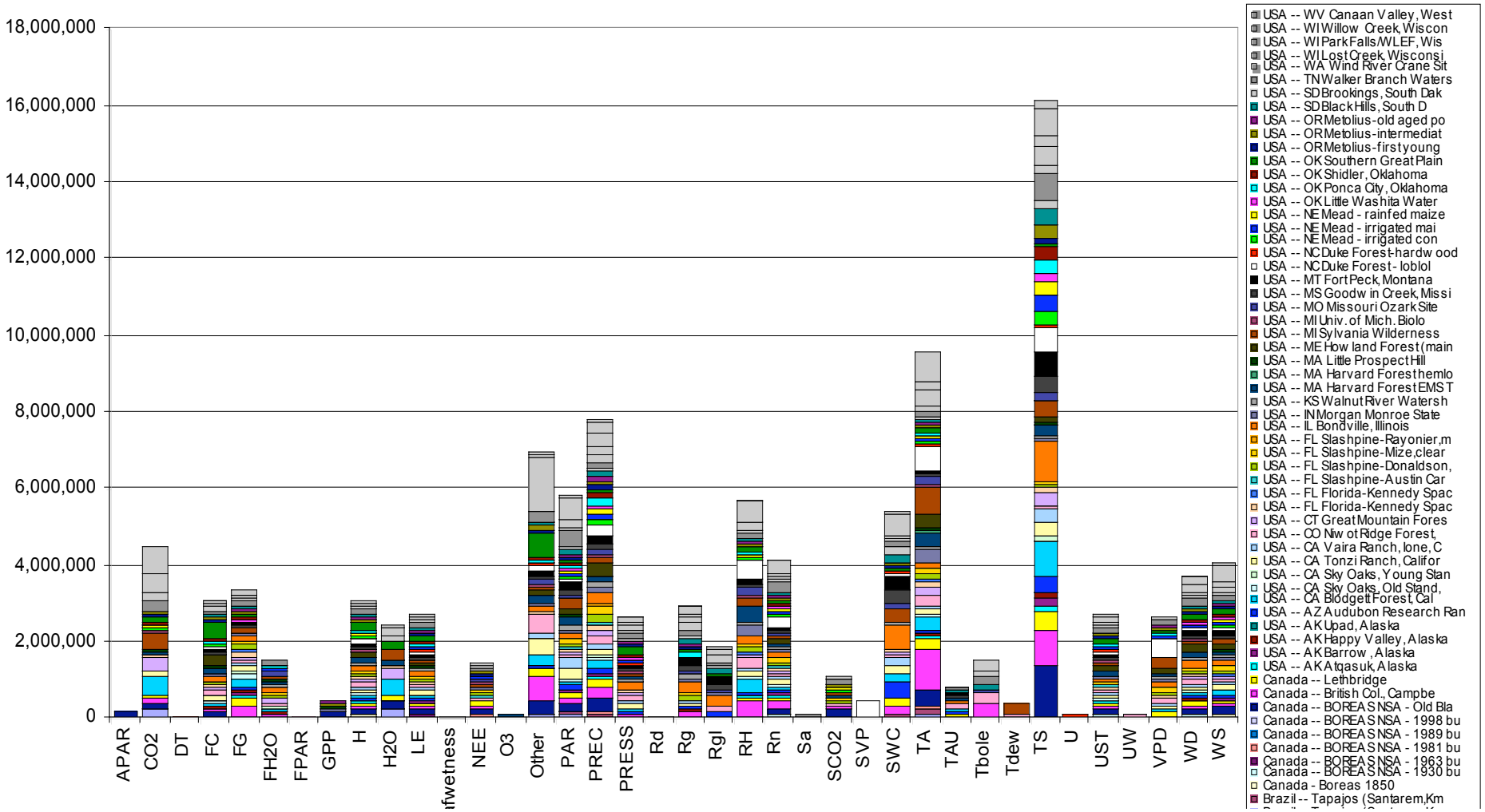
Browsing For Data Availability

Sites Reporting Data Colored by Year



Browsing For Data Availability

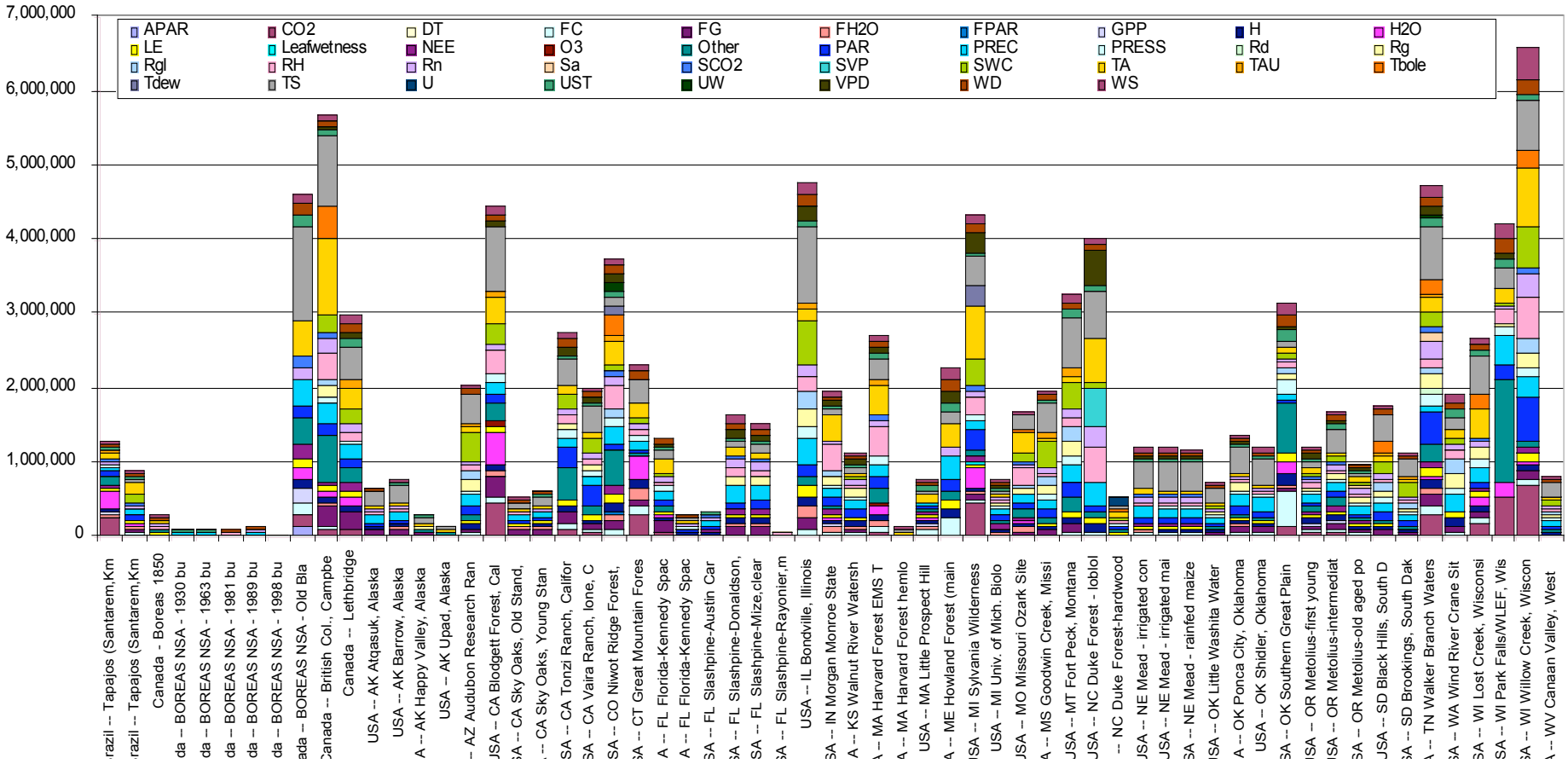
Total Data Availability by Type Colored by Site



Data type reporting is far from uniform across type

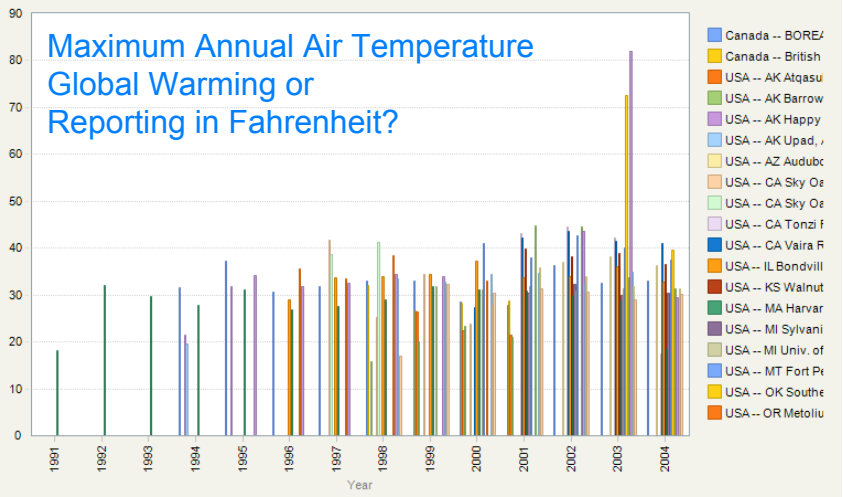
Browsing for Data Availability

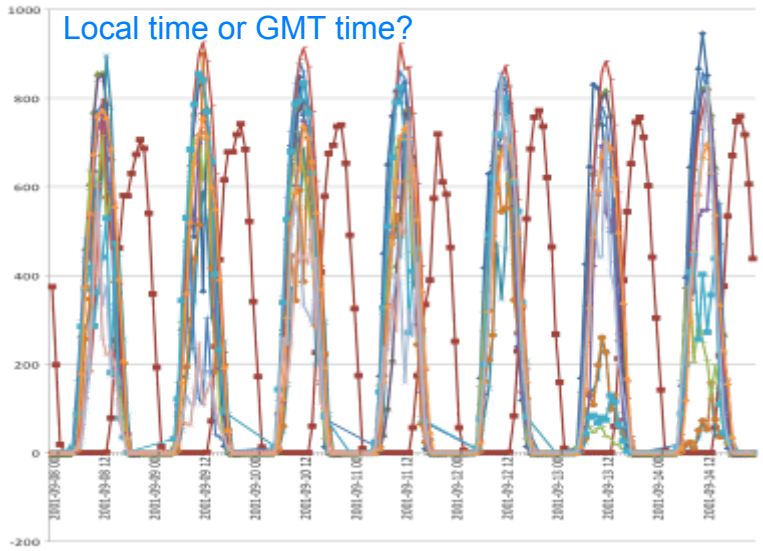
Total Data Availability by Site Colored by Type

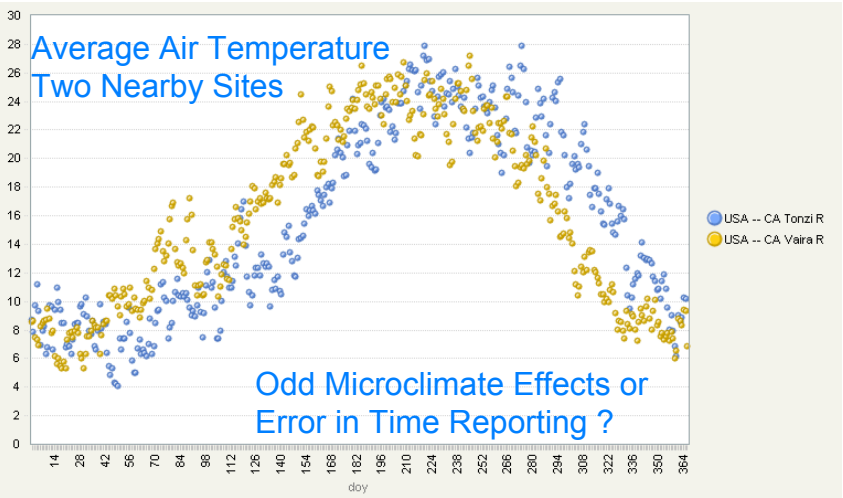


Browsing for Data Inconsistencies

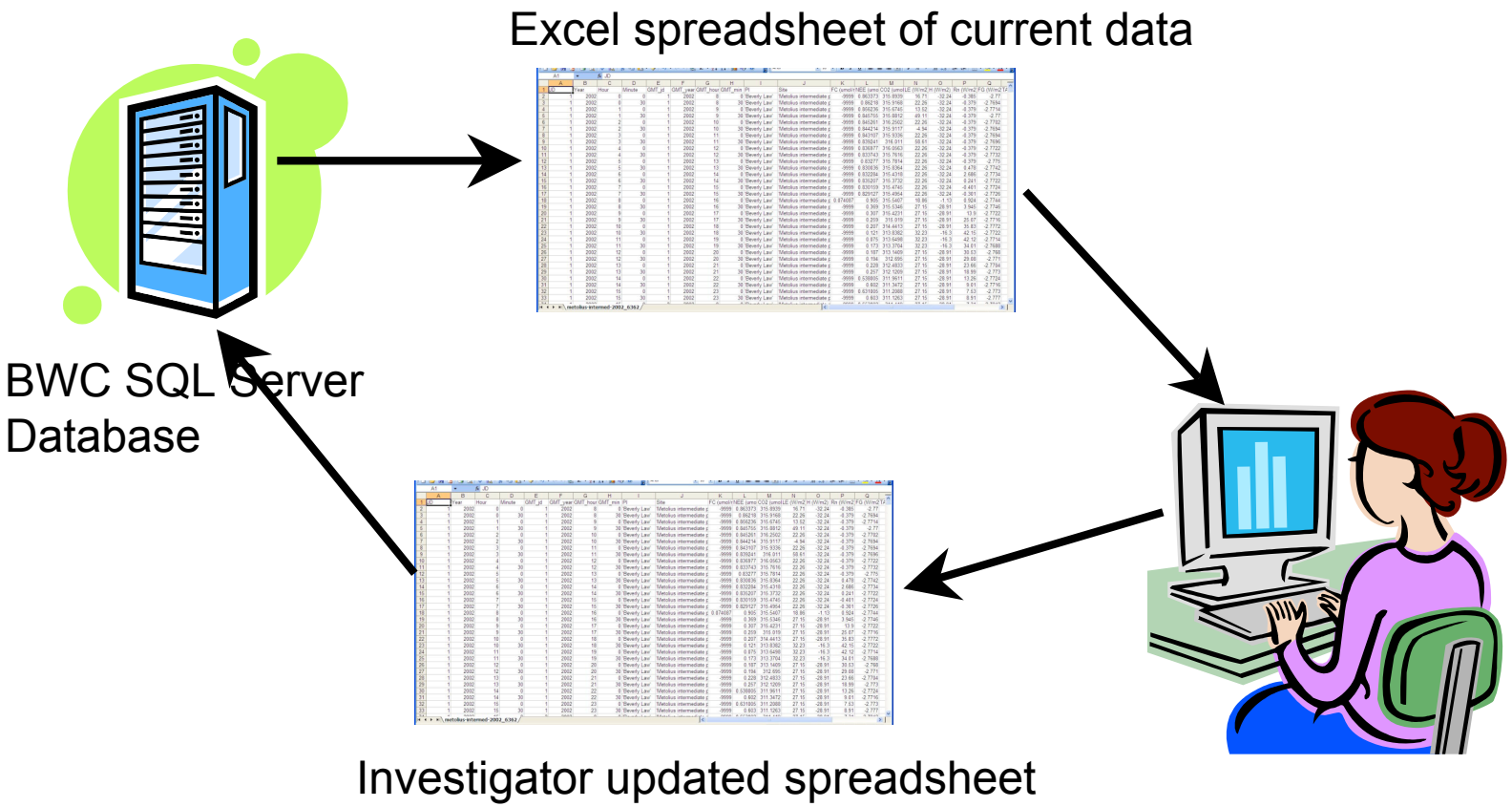
- Potential data problems
  - Sometimes easier to spot when comparing to other data







Data Cleaning and Versioning



Lessons Learned To Date

- Controlled vocabularies are important
  - Understanding how to correlate columns across towers is important
  - Descriptive names sometimes give too much information
  - *Assume a decode step in the staging pipeline*
- Metadata is as important as the data
  - Comparing sites of like vegetation, climate, and disturbances is as important as latitude or other physical quantities
  - *Curate the metadata with the data*
- *Automation is essential*
  - *Automate the data ingest, maintenance, cube, and report generation*
- Data analysis and data cleaning are intertwined
  - Data cleaning is on-going
  - *Share the simple tools and visualizations*
- *Regional data analysis methodology is evolving*

The saga continues at http://dsd.lbl.gov/BWC/amfluxblog/ and http://research.microsoft.com/~vaningen/BWC/BWC.htm

"Another decade or so, and it'll be warm enough for us."

"We have lots of information technology. We just don't have any information."

Futures

- Improve current capabilities
  - Assemble gap-filled and non-gap filled data sets
  - Implement incremental data staging to enable data editing by the scientists (version tracking)
  - Implement expanded metadata handling to enable scientist to add site characteristics and sort sites on those expanded definitions
- Handle imagery and other remote sensing information
  - Curating images is different from curating time series data
  - Graphical selection and display of data
- Support for user specified calculations within the database
- Support for direct connections to analysis and statistical packages
- Linkage with models
  - Additional (emerging) data standards such as NetCDF
  - Handling “just in time” data delivery and model result curation
- Data mining subscription services

URLs

- **BWC Ameriflux Data Server**

  http://bwc.berkeley.edu/Amflux
- **BWC Ameriflux Project**

  http://dsd.lbl.gov/BWC/amfluxblog/

  http://esd.lbl.gov/BWC/thrust_areas/ameriflux.html
- Berkeley Water Center (BWC)

  http://esd.lbl.gov/BWC/

Funding – This project is funded by the Microsoft Corporation

Demo

Excel Pivot Table

A7 | MaximumValue

	A	B	C	D	E	
1	Exdatumtype	(none)				
2	Repeat	0				
3	Datumtype	Rg				
4	Value units	(Multiple Items)				
5	Dataset	June download from ORNL web site				
6						
7	MaximumValue			Site	Latitude	
8				Brazil -- Tapajos (Santarem,Km83), Logged Forest	Canada -- BOREAS NSA - 1930 burn site	Canada -- BOR
9	Yy	Yymm	Yymmdd	-3.018029	55.90583	55.911671
10	2001	2001-09	2001-09-01	841.4453125		
11			2001-09-02	985.8311157		
12			2001-09-03	931.5270386		
13			2001-09-04	794.5578613	414.6756592	
14			2001-09-05	928.4255371		
15			2001-09-06	880.4135742	557.3077393	
16			2001-09-07			
17			2001-09-08	-0.737914026		
18			2001-09-09	798.2893677		
19			2001-09-10	1002.508606		
20			2001-09-11	937.0953979		
21			2001-09-12	887.0908813		
22			2001-09-13	834.5820313		
23			2001-09-14	949.5360107		
24			2001-09-15	1021.355408		
25			2001-09-16	929.6103516		
26			2001-09-17	922.444519		
27			2001-09-18	991.7322388		
28			2001-09-19	905.5108032		
29			2001-09-20	804.617981		
30			2001-09-21	873.005188		
31			2001-09-22	930.5013428		
32			2001-09-23	906.8458862		

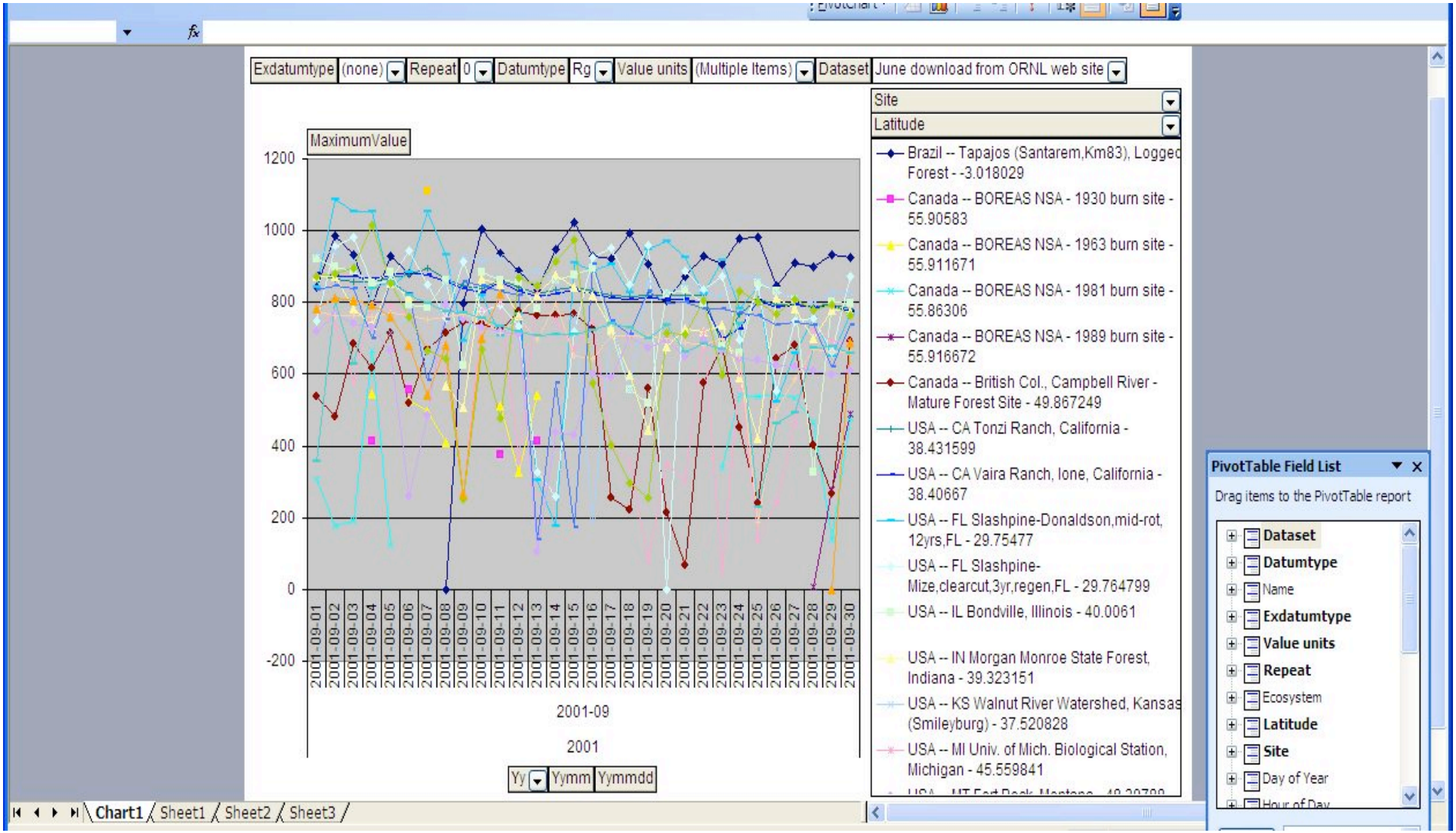
Chart1 / Sheet1 / Sheet2 / Sheet3

PivotTable Field List

Drag items to the PivotTable report

- Dataset
- Datumtype
- Name
- Exdatumtype
- Value units
- Repeat
- Ecosystem
- Latitude
- Site
- Day of Year
- Hour of Day

Excel Pivot Chart



Report Server

Report Viewer - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://bwc.lbl.gov/ReportServer/Pages/ReportViewer.aspx?%2fBWC+Reports%2fCountBrowserTable&rs%3aCommand=Render

Site: CA - Tonzi Ranch, CA - Vaira R

Datumtype: TA

Year: 2003, 2004, 2005, 2006

Offset: All

Exdatumtype: All

View Report

1 of 1 100% Find | Next Select a format Export

Return to report home

Data Count Browsing by Measurement, Site and Year

Selected offsets: All

Selected exdatumtypes: All

		2003	2004	2005
TA	CA - Tonzi Ranch	35040	34695	34306
	CA - Vaira Ranch	17520	16499	15347
	OR - Metolius-first young aged pine		17568	17520
	OR - Metolius-intermediate aged ponderosa pine	17520	17568	17520

Done Internet