FLUXDATA.ORG: A Collaborative Data Sharing Portal For Global Carbon-Flux Analysis

Catharine van Ingen[1], Deb Agarwal[2], Marty Humphrey[3], Monte Goode[2], Norm Beekwilder[3], Dario Papale[4], Markus Reichstein[5]

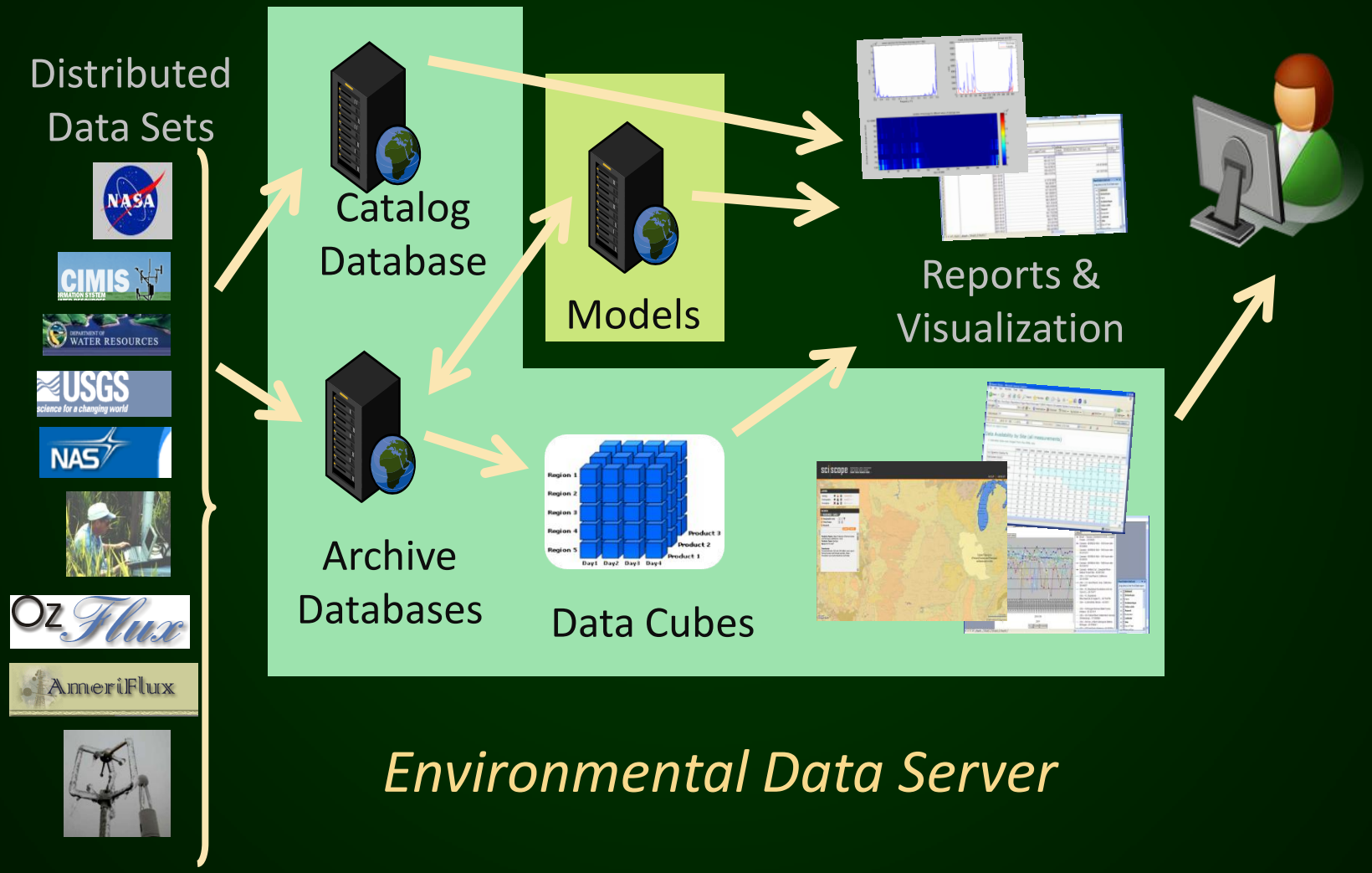
[1]Microsoft Research; [2]Lawrence Berkeley Laboratory; [3]University of Virginia; [4]University of Tuscia; [5]Max Planck Institute for Biogeochemistry

AGU Meeting of the Americas, August 2010 IN23A-06

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
- Currently 85 paper teams with over 400 scientists

Initial Challenge: Connecting People and Data



Database Schema Abstractions

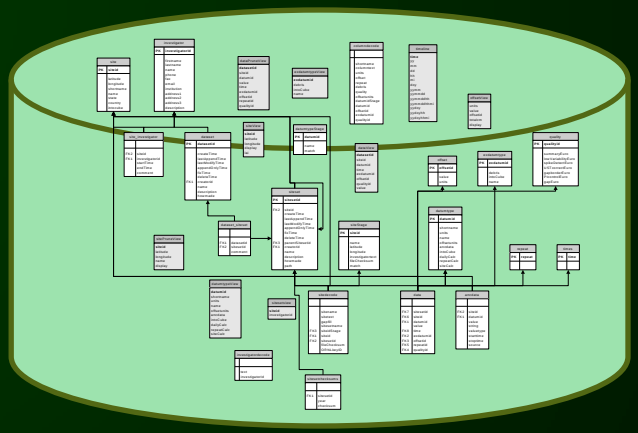
- Sensor data, ancillary data and metadata
  - Analyses often require combining time series data with fixed, or nearly fixed ancillary data such as river mile, vegetative cover, sediment grain size
  - Ancillary data used as fixed property, time series, or event time window
  - Metadata describing algorithms, measurement techniques, etc.
  - Normalized table structure simplifies adding variables and cube building
- Versioning and folder-like collections
  - Accommodate algorithm changes
  - Track derivations through processing pipeline
  - Define and track analysis "working set"
- Namespace translations
  - Data assembly traverses different repositories each with own (useful?) name space
  - Some repositories encode metadata in variable name space
  - Units conversions when aggregating temporally or spatially

Analysis Tools Excel, MatLab, R, ArcGIS

Map Mashups

Spreadsheet data summaries

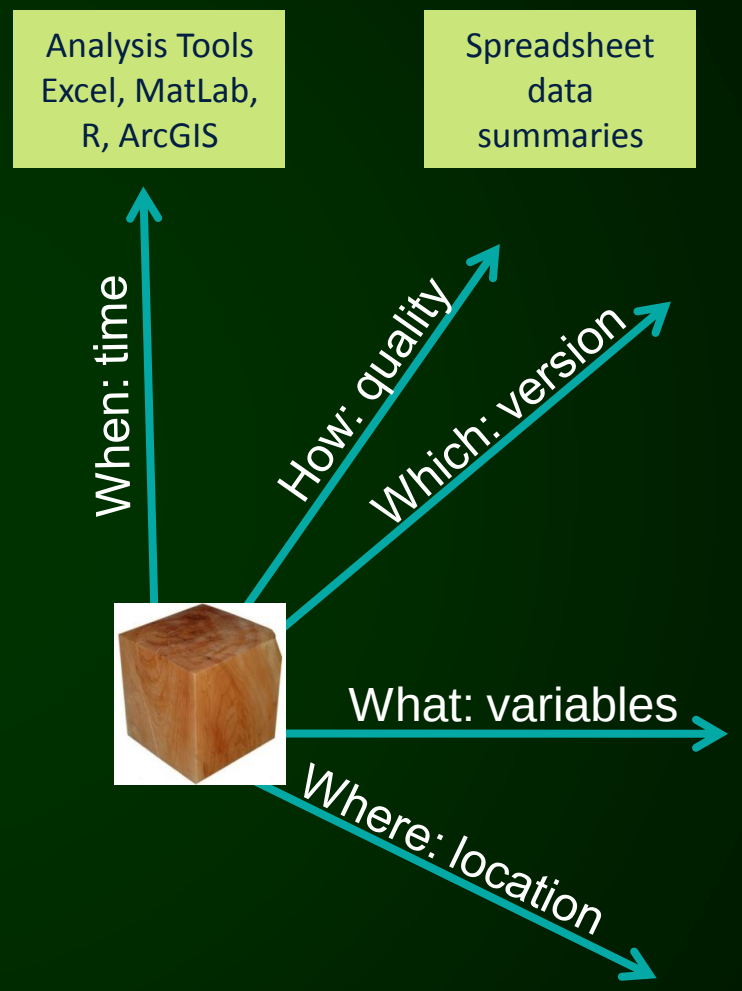
CSV, MatLab and Excel file import/export



All access layers share the same abstractions

Data Cube Abstractions

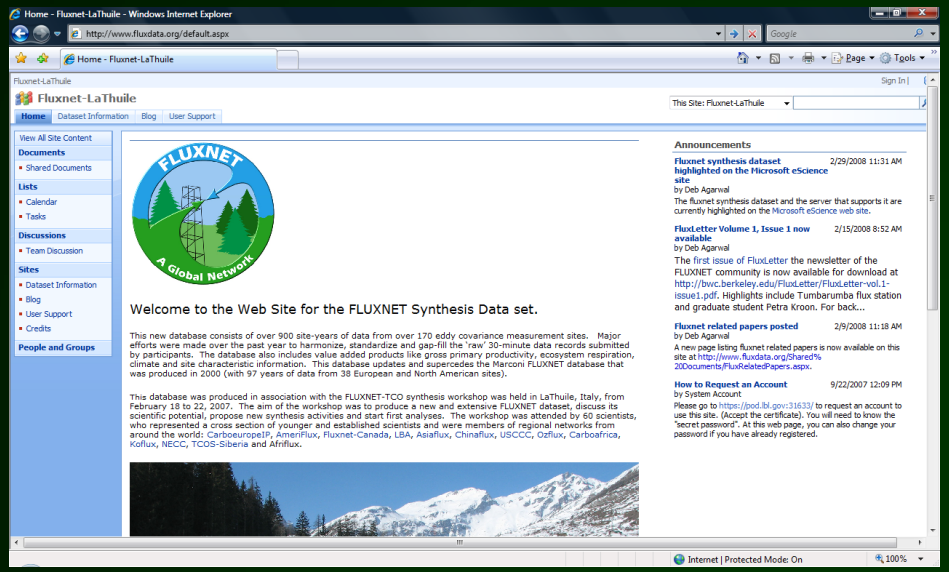
- A data cube is a database specifically for data mining (OLAP)
  - Initially developed for commercial needs such as sales of Oreos and milk
  - Simple aggregations (sum, min, or max) can be pre-computed for speed
  - Hierarchies for simple filtering with drilldown capability
  - Additional calculations (median) can be computed dynamically or pre-computed
  - All operate along dimensions such as time, site, or variable type
  - Constructed from a relational database
  - A specialized query language (MDX) is used
- Client access is key
  - Excel PivotTables allow simple data viewing
  - More powerful analysis and plotting using Matlab, R and ArcGIS



Dimensions driven by the data

space and time are fundamental drivers

Data Centric Collaboration Portal (1/2)

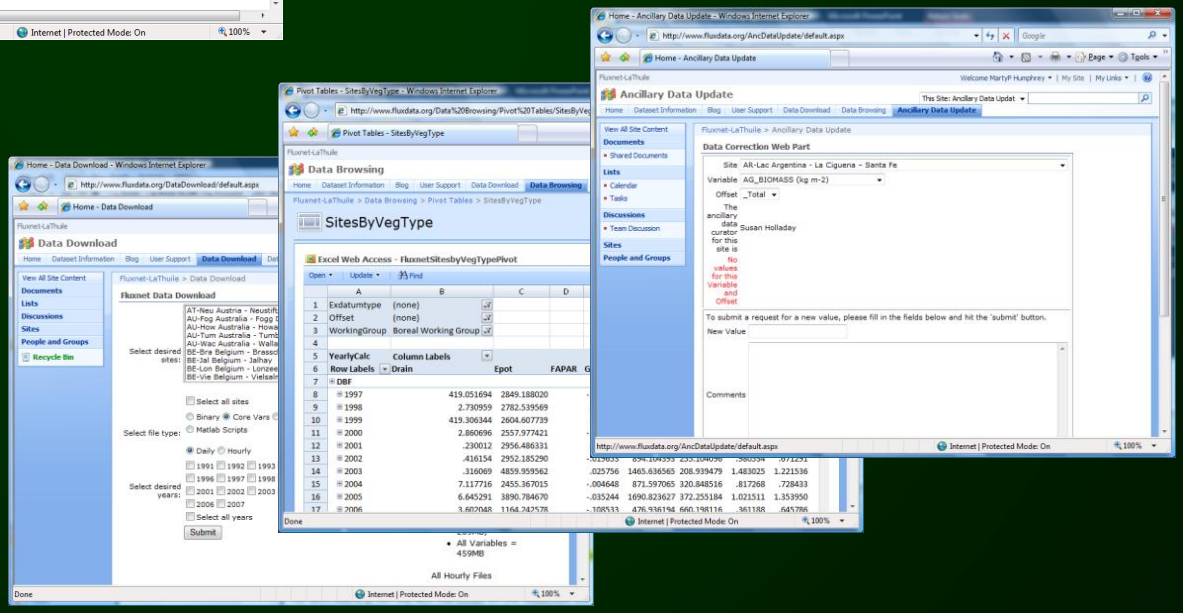


For the public:

- General dataset and
- collaboration information
- User support

For scientists analyzing the data:

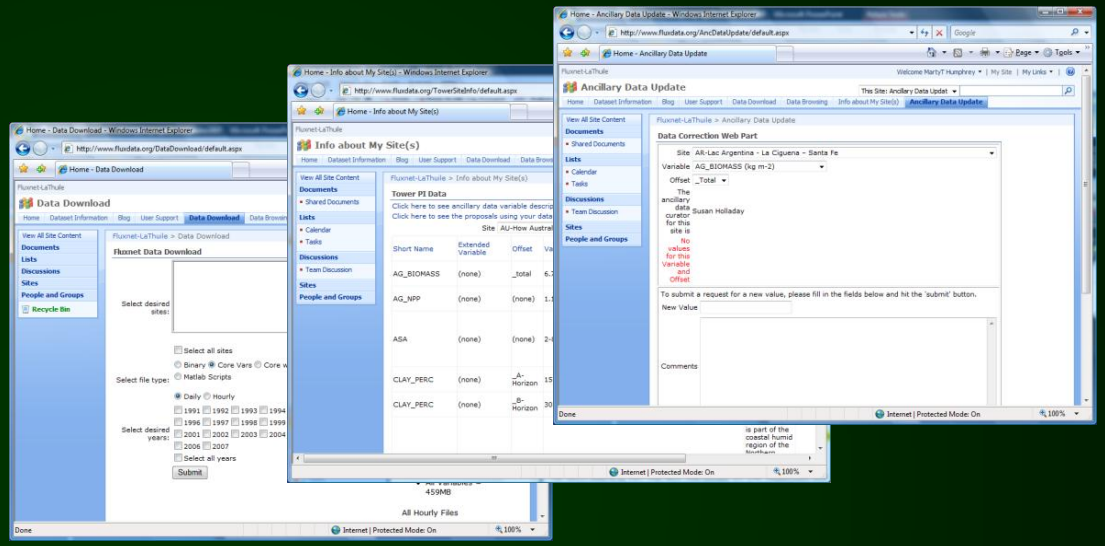
- Download data
- Interactively explore data
- Suggest ancillary data updates and corrections

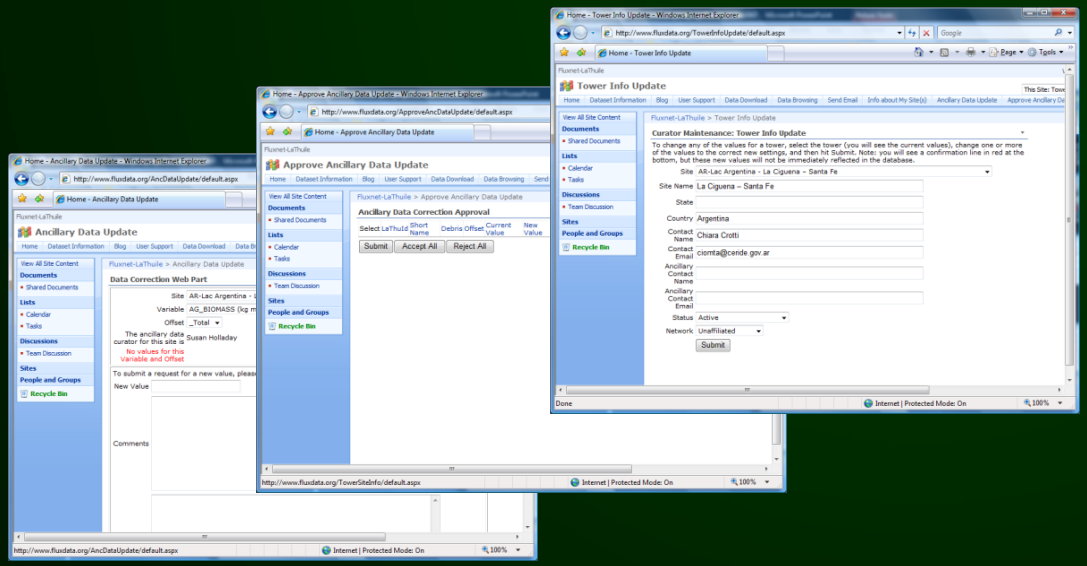


Data Centric Collaboration Portal (2/2)

For tower owners:

- Download curated tower data
- Review and update info about the tower
- Provide ancillary data updates and corrections



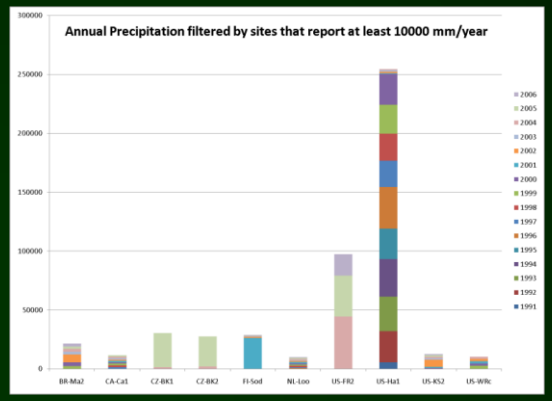


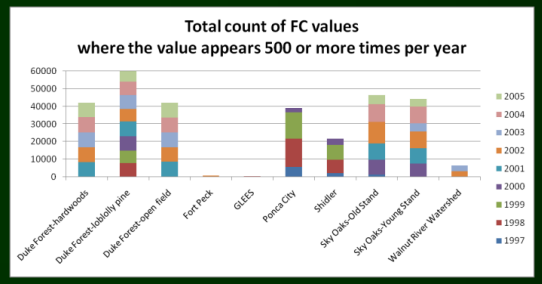
For data curators:

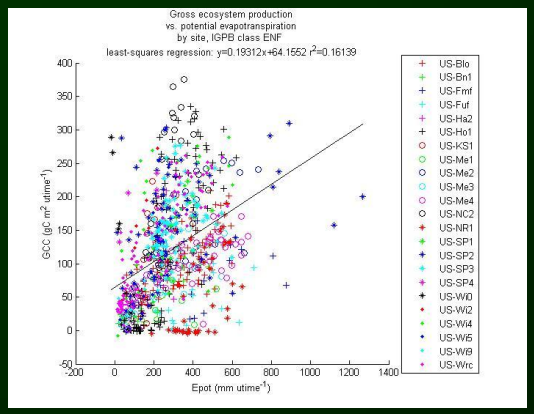
- Download data
- Provide ancillary data updates and corrections
- Reject or accept proposed data updates and corrections

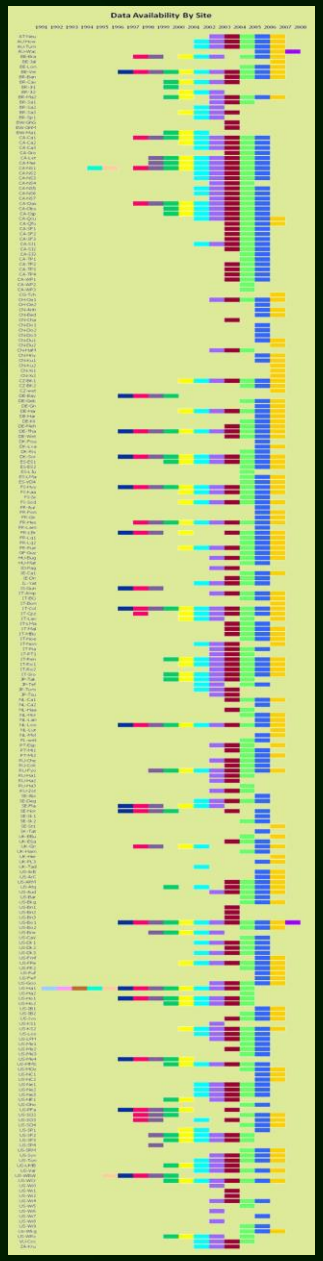
Browsing Sensor Data for Availability, Quality and Early Analysis

- Verifying that the data needed for an analysis exists at good enough quality and quantity always a first step toward that analysis
- Data cleaning never ends
  - Only when data are used for analysis can you be sure
  - Existing practice of running scripts on specific site years often misses the big picture



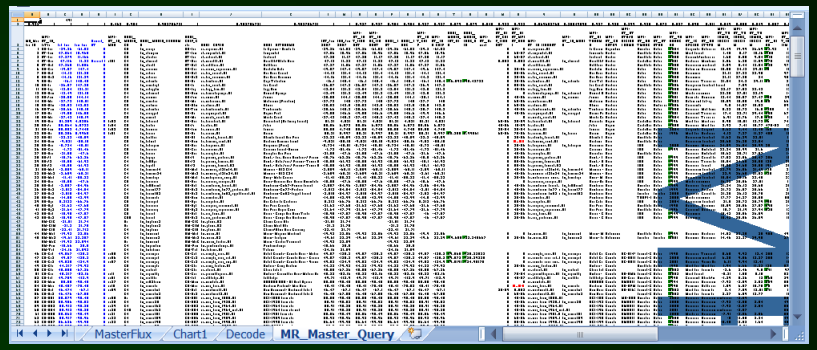






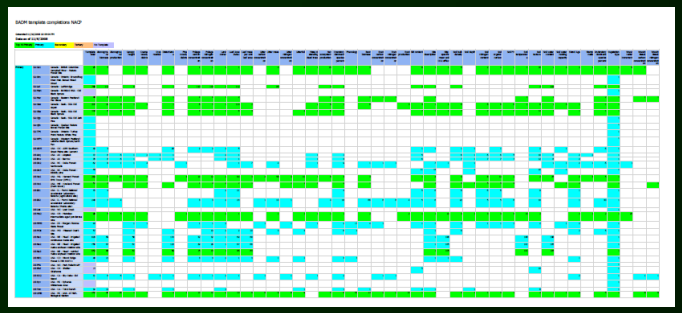
Curating Ancillary Data

- Original collection amassed from 3 private collections and “CEIP” templates
  - One spreadsheet with 76 columns and 458 rows created by cut/paste transcription
  - ‘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 and disturbances such as a fire, harvest, landslide
- Major effort to determine the list of tower sites with latitude and longitude
  - A proof point for why a living data set is so important for synthesis science
- “BADM” protocols developed by scientists from different tower types and computer scientists
  - Different practices for different biomes
  - Specified field methodologies, units, reporting frequencies
  - Submission via Excel template or web form
  - Exported as cleaned subsets, templates and reports showing availability



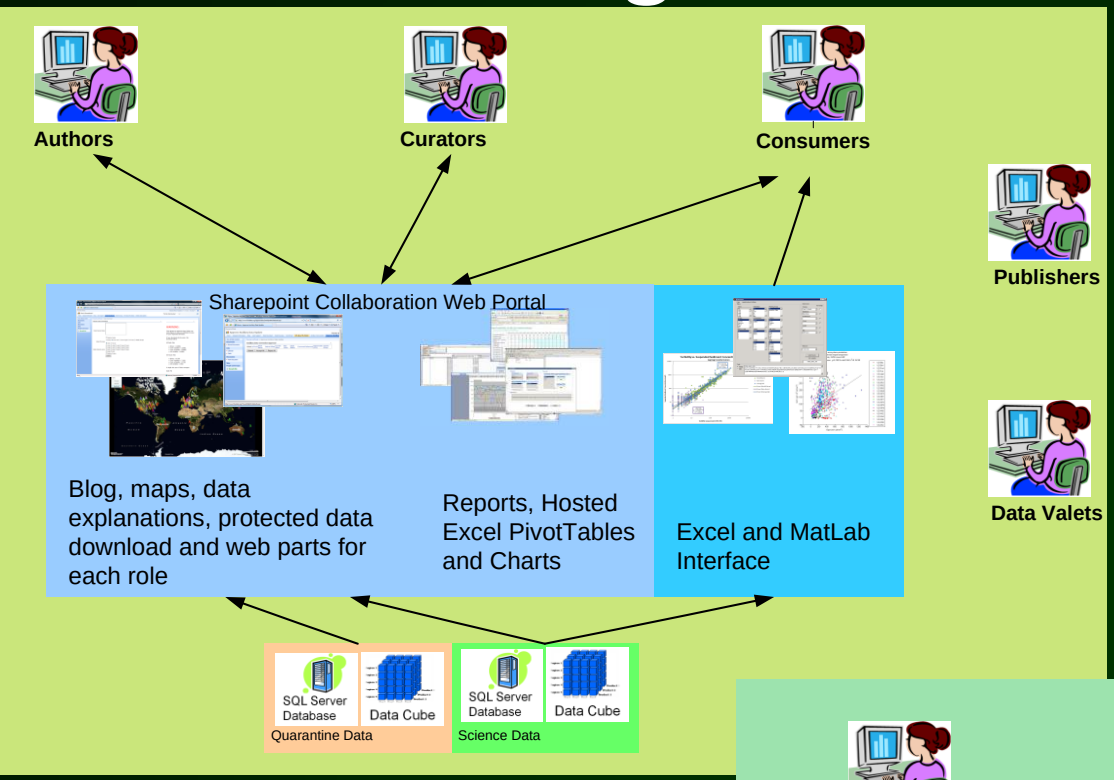
Zotino Sites in Russia

Very close together
Deployments change
How many sites?
Where are they?



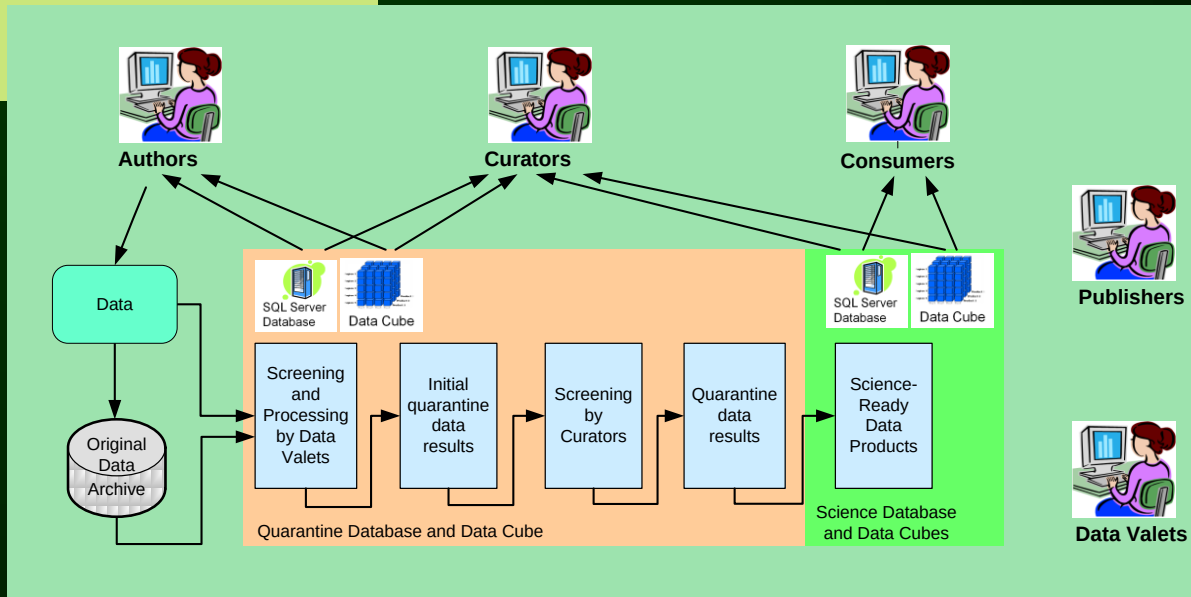
[FAO protocols (Law et al 2008)]

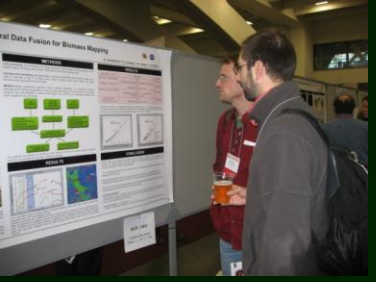
Enabling Virtual Conversations



Collaboration Centric View

Data Centric View





Current Challenge: Connecting Data, People and Publications

- Data browsing, curation, and archive is only the beginning
- Publish data with papers
  - Capture contributor information, fair-use criteria, and acknowledgments
  - Capture analysis artifacts and connect them to both data source and paper learnings
  - Capture then current data corrections, gap-fills, quality assessment inferences for subsequent forensics

Global Convergence in the Temperature Sensitivity of Respiration at Ecosystem Level

Miguel D. Mahecha[1,2*], Markus Reichstein[1], Nuno Carvalhais[1,2], Gitta Lasslop[1], Holger Lange[4], Sonia I. Seneviratne[2], Rodrigo Vargas[5], Christof Ammann[6], M. Altaf Arain[7], Alessandro Cescatti[8], Ivan A. Janssens[9], Mirco Migliavacca[10], Leonardo Montagnani[11,12], Andrew D. Richardson[13]

[1]Max Planck Institute for Biogeochemistry, 07745 Jena, Germany.

[2]Institute for Atmospheric and Climate Science, ETH Zurich Universitätsstrasse 16, 8092 Zurich, Switzerland.

[3] Faculdade de Ciências e Tecnologia, FCT, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal.

[4]Norsk Institutt for skog og landskap, N-1431 Ås, Norway.

[5] Department of Environmental Science, Policy and Management, University of California, Berkeley, CA 94720, USA.

[6]Agroscope ART, Federal Research Station, Reckenholzstr. 191, CH-8046 Zurich, Switzerland

[7]McMaster Centre for Climate Change, McMaster University, Hamilton, Ontario, Canada.

[8]European Commission, Joint Research Center, Institute for Environment and Sustainability, Ispra, Italy.

[9]Department of Biology, University of Antwerpen, Universiteitsplein 1, 2610 Wilrijk, Belgium.

[10]Remote Sensing of Environmental Dynamics Laboratory, DISAT, University of Milano-Bicocca, Milano, Italy.

[11]Servizi Forestali, Agenzia per l'Ambiente, Provincia Autonoma di Bolzano, Bolzano, Italy.

[12]Faculty of Sciences and Technologies, Free University of Bozen-Bolzano, Piazza Università 1, 39100, Bolzano, Italy.

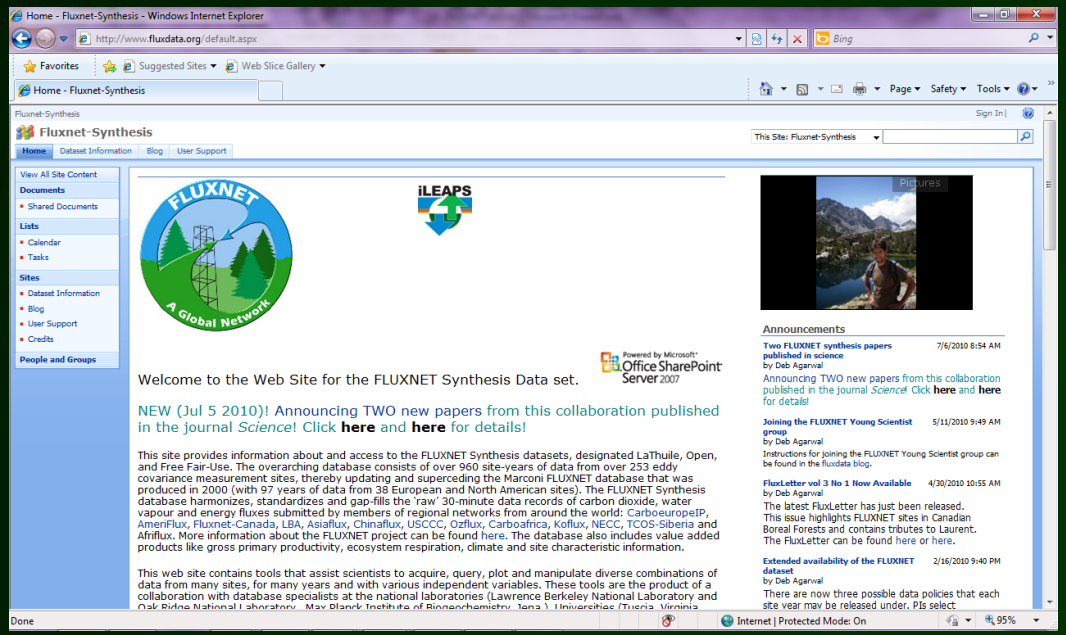
[13]Department of Organismic and Evolutionary Biology, Harvard University, HUH, 22 Divinity Avenue, Cambridge, MA 02138, USA.

*To whom correspondence should be addressed; E-mail: mmahecha@bgc-jena.mpg.de.

1

Increase data sharing, data reuse and, most important, science learning

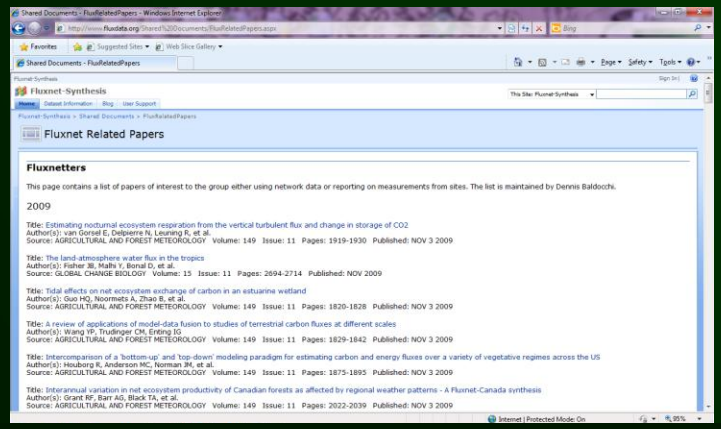
Data Centric Publication Aware Collaboration Portal (1/2)

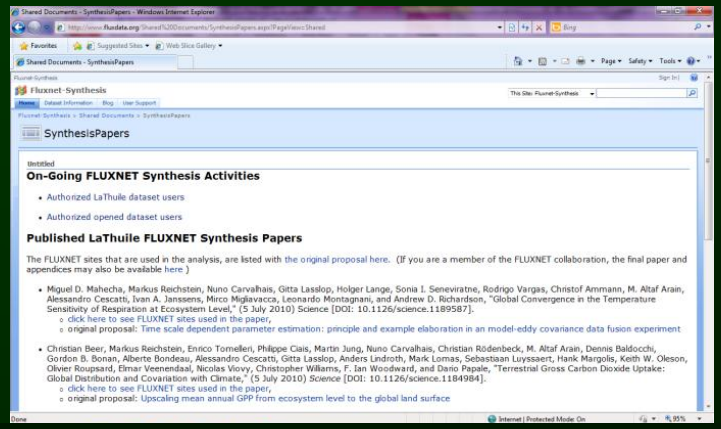


Much larger public area

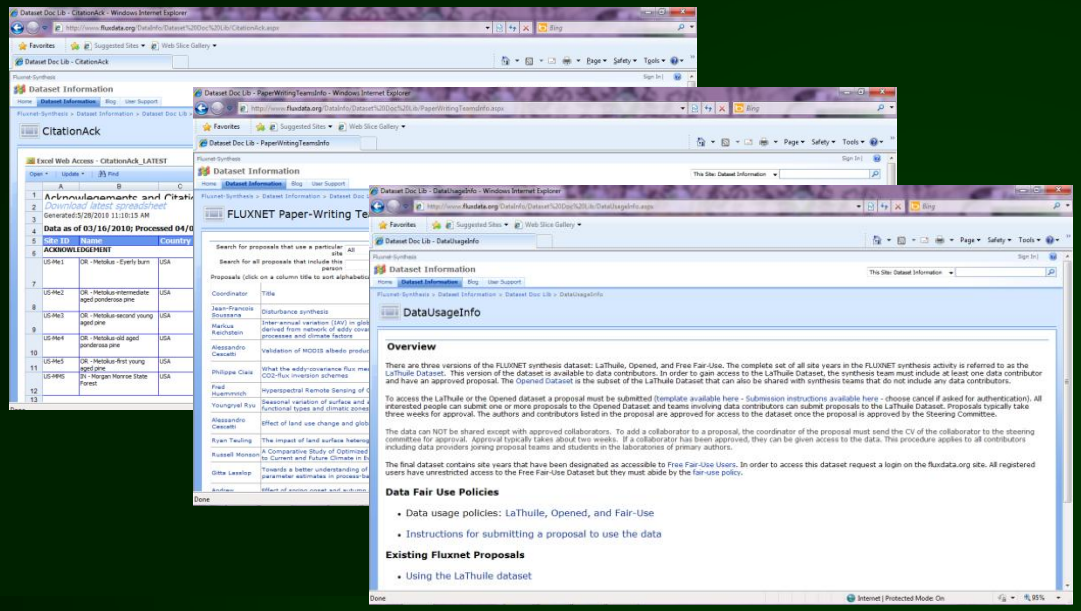
- Recent papers including others of related interest
- Collaboration meetings
- Pictures!

Young Scientist area





Data Centric Publication Aware Collaboration Portal (2/2)

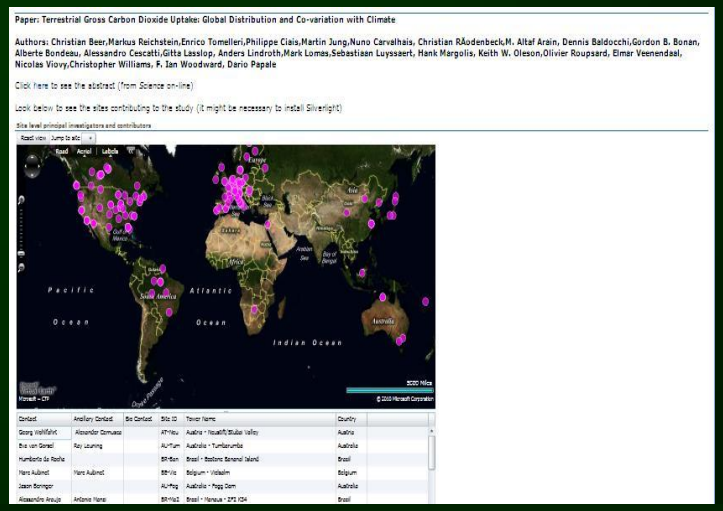


Three level data sharing

- Policy explanation
- Specific areas for data access and publications tracking for each
- Enumerated citations and acknowledgements

Enabling a "FLUXNET Collaboration" Author

- Ability to track the then current tower authors and data processing authors used by each paper
- Tower owners can discover and track papers that use their data



Learnings to Date

- FLUXNET is not one collaboration but rather several different sub-groups with different cultures
  - Ontologies, provenance, curation all mean different things
  - Some publishing ready to use data; others provide "caveat emptor" data access
  - Users expect clean science ready data at all times
- Tracking publications and data usage together enables a virtuous cycle
  - Data represents both a cost and asset to the originator
  - Publication of data is more costly than access
- Plan for data set evolution
  - Algorithms improve
  - Data can only be cleaned when used – science curators are at best a sanity check
  - Campaigns driven by specific papers or analyses give the best motivation