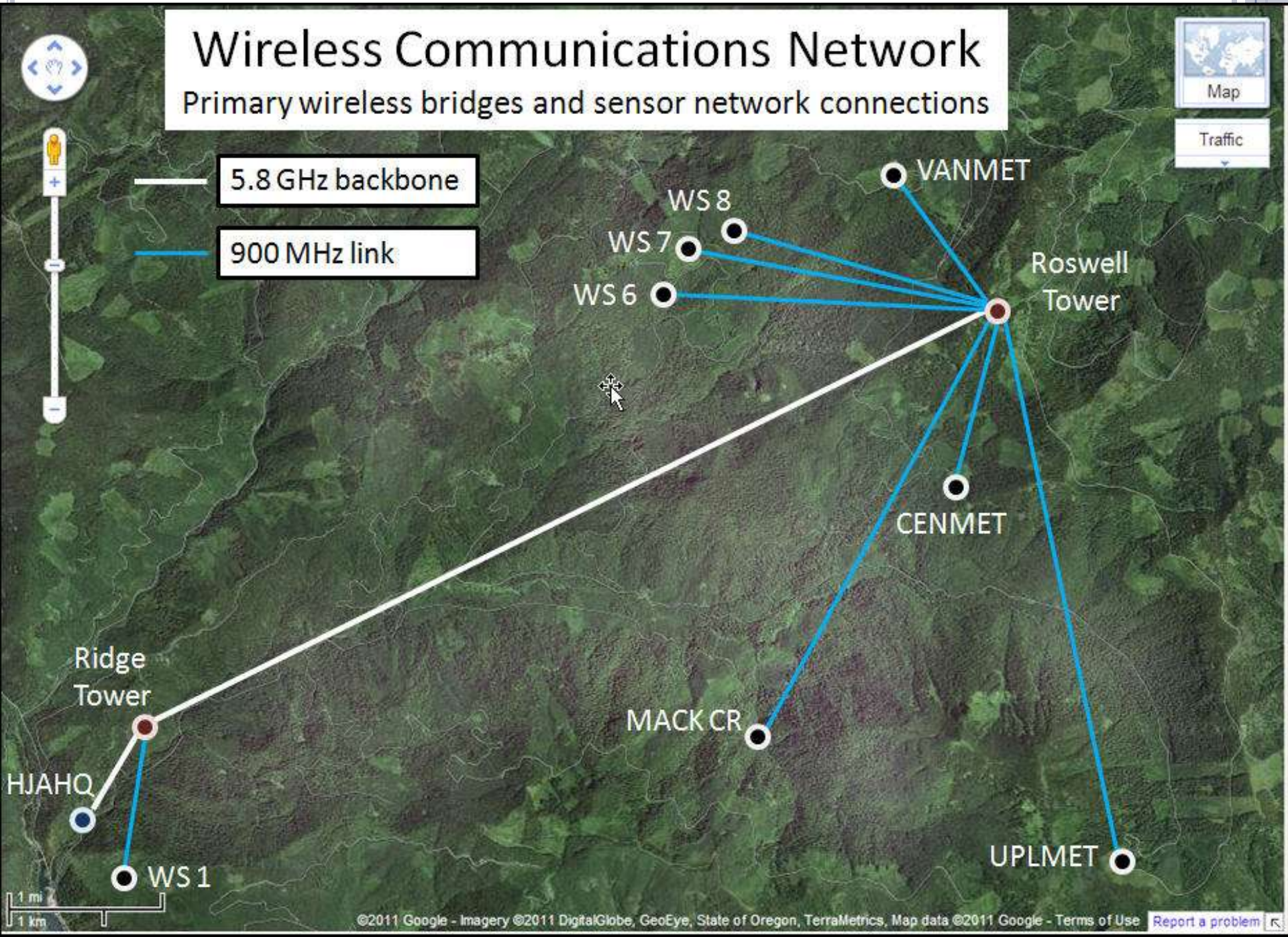
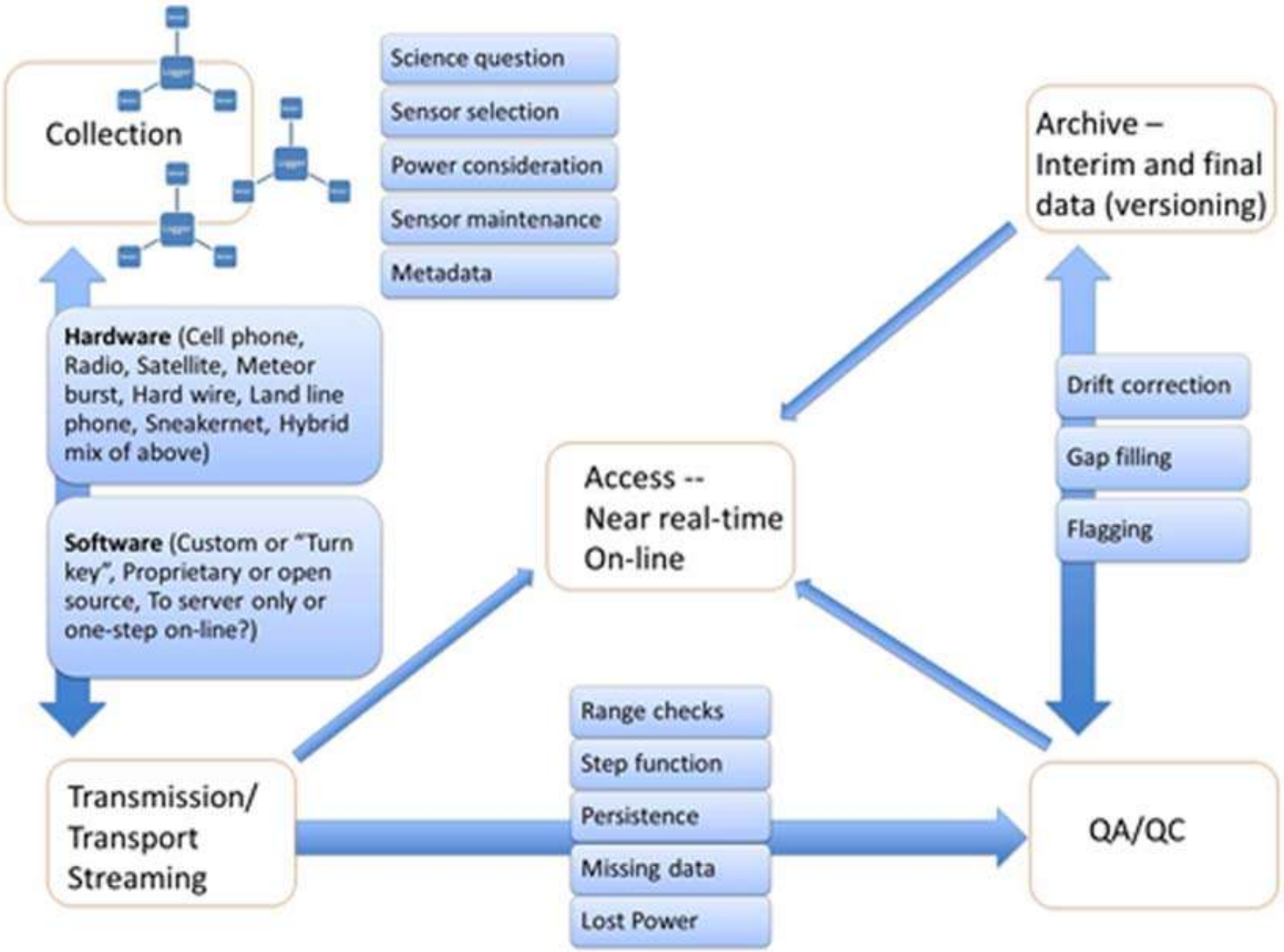
Sensor data management software, requirements and considerations

Don Henshaw
H.J. Andrews Experimental Forest


Wireless Communications Network
Primary wireless bridges and sensor network connections
5.8 GHz backbone
900 MHz link
VANMET
WS 8
WS 7
WS 6
Roswell Tower
CENMET
MACK CR
UPLMET
Ridge Tower
HJAHQ
WS 1
Map
Traffic
1 mi
1 km
©2011 Google - Imagery ©2011 DigitalGlobe, GeoEye, State of Oregon, TerraMetrics, Map data ©2011 Google - Terms of Use
Report a problem


Collection
Science question
Sensor selection
Power consideration
Sensor maintenance
Metadata
Archive – Interim and final data (versioning)
Hardware (Cell phone, Radio, Satellite, Meteor burst, Hard wire, Land line phone, Sneakernet, Hybrid mix of above)
Drift correction
Gap filling
Flagging
Access -- Near real-time On-line
Software (Custom or "Turn key", Proprietary or open source, To server only or one-step on-line?)
Range checks
Step function
Persistence
Missing data
Lost Power
Transmission/ Transport Streaming
QA/QC

Common Themes from participating sites

Joint NERC Environmental Sensor Network/Sensor NIS Workshop, Hubbard Brook Experimental Forest, NH, October 25-27TH, 2011

- Approaches
  - Top down (NEON, USGS)
    - More uniform, faster implementation, less flexible
  - Bottom-up (LTER network sites, individual sites)
    - Less standardization, customized approaches and software
    - Adopting solutions – how do you decide?
      - Reluctance to invest time and energy
      - Lack of mature software
      - Steep learning curve
    - Do we need light-handed standardization?
      - E.g., software, methods, controlled vocabulary, units, etc.

Common Themes from participating sites

Joint NERC Environmental Sensor Network/Sensor NIS Workshop, Hubbard Brook Experimental Forest, NH, October 25-27TH, 2011

- Greatest Needs
  - Middleware between sensor/data logger and database /applications
  - Programming support
  - Training workshops to disseminate knowledge & solutions
  - Ways to share experiences with software and tools that are useful
    - Clearinghouse for sharing code and solutions
    - Knowledge Base (web page) organized by topics (http://im.lternet.edu/resources/im_practices/sensor_data)

- Dataloggers
  - Campbell Scientific - most common (http://www.campbellsci.com/)
  - Hobo (www.onsetcomp.com)
  - Nexsens Technology (http://nexsens.com)
  - GRAPE - NEON (http://www.neoninc.org/)

Sensor Data Management Middleware/Software

- Purpose of middleware
  - Data storage / data handling
  - Data aggregation, formatting, filtering
  - Documentation
  - Automated QA/QC on data streams
  - Archiving
- Software/middleware – Proprietary
  - Campbell LoggerNet (most common)
  - Hobo
  - Vista Data Vision
  - YSI EcoNet
  - Custom applications: Matlab, Excel, SQLServer
- Open Source
  - Environments (see next slide)
  - Custom applications (Python, PHP, MySQL, etc.)

Sensor Data Management Middleware Open Source Environments for Streaming Data

- Matlab GCE toolbox (Proprietary/ limited open source)
  - GUI, visualization, metadata-based analysis, manages QA/QC rules and qualifiers, tracks provenance
- Open Source DataTurbine Initiative
  - Streaming data engine, receives data from various sources and sends to analysis and visualization tools, databases, etc., TiVo-like functionality
- Kepler Project (open source)
  - GUI, reuse and share analytical components/workflows with other users, tracks provenance, integrates software components and data sources
- R-project libraries (open source)
  - Statistical and graphical capabilities, analysis tools, code reuse and sharing, integrated environment

Sensor Management System Components

- Develop protocols for installation of sensors
- Develop calibration / maintenance schedules
  - System alerts (nagging system)
- Build and maintain sensor network metadata
  - Data collection documentation
  - Annotation of sensor events / Sensor history
- Data workflows /processing
- Quality Control / Flagging data
- Archiving
  - Creating a “citable” database
  - Versioning (monthly vs. annual ; provisional v. final)
  - Periodic snapshots or queries
  - Tracking changes to the data, e.g., audit trail

Key Metadata for Sensor Networks

- Sensor descriptions
  - Sensor relocations or replacement (automate w/barcodes?)
  - Sensor events and failures
  - Calibration events / maintenance history
- Data collection documentation
  - Sensor deployment information, incl. station-level
  - Geo-location / operational time span
  - Sampling frequency
  - Methodology changes, e.g., temperature radiation shield change
  - Photo points for station, e.g., hemispheric photos, track local conditions and changes
- Data processing documentation
  - System requirements / Hardware configuration / history
  - Data processing workflow
  - Datalogger program versions / wiring diagrams with labels
  - Attribute / Flag definitions

Sensor Management

- SensorML standard
  (http://www.opengeospatial.org/standards/sensorml)
  - Framework for observational characteristics of sensors
  - Covers station, deployment, sensor, parameter
  - Tools being developed - Lacking production-grade software?

- CUAHSI HIS / Observation Data Model
  - Relational database model for individual observations
  - Provide maximum flexibility in data analysis through the ability to query and select individual observation records
  - Record level metadata

Quality Assurance – preventative measures

- Sensor redundancy
  - Ideal: Triple the sensor, triple the logger!
  - Practical: Cheaper, lower cost, lower resolution sensors, or correlated (proxy) sensors
  - Side Effect: establish user-confidence in data products
- Routine calibration and maintenance
  - Schedule or stagger to minimize data loss
- Continuous monitoring and evaluating of sensor network
  - Early detection of problems
  - Limited budgets and increased volume of data precludes past manual sensor auditing practices
  - Automated alerts or streaming QC

Quality control on Streaming Data: Quality Levels

- Quality control is performed at multiple levels
- Quality level important to describe in metadata,
  - Description differs among programs
  - Examples: NASA, CUAHSI, Ameriflux
- http://ilrs.gsfc.nasa.gov/reports/ilrs_reports/9809_attach7a.html
- http://his.cuahsi.org/documents/ODM1.pdf p.18

- Level 0 (Raw streaming data)
  - Raw data, no QC, no data qualifiers applied (data flags)
  - Preservation of original data streams is essential
  - Some datalogger conversion of units and formats may be acceptable (Level ½)

Quality control on Streaming Data: Quality Levels

- Level 1 (QC'd, calibrated data, qualifiers added)
  - Provisional level (near real-time preparation)
    - Typically for internal use, if released, provisional data must be labeled clearly
    - Data qualifiers are added from initial QC
    - Infill and flag missing datetimes
  - Published level (delayed release)
    - QC process is complete
    - Data is unlikely to change – data set is unique
- Comments:
  - Only logger missing value codes should be deleted from streams, e.g. contention that even impossible values may have information
  - *Immediate* Qc is important in near real-time, and *subsequent* QC as trends become apparent (e.g., sensor drift, degradation)
  - Foremost, identify what QC has been done
    - Identify streaming QC methods, thresholds, assumptions
    - If no QC, that should be made clear too

Quality control on Streaming Data: Possible quality control checks in near real-time

- Timestamp integrity (Date/time)
  - Sequential, fixed intervals, i.e., checks for time step or frequency variation
- Range checks
  - Sensor specifications - identify impossible values; not unlikely ones
  - Seasonal/reasonable historic values
    - Highly dependent on the sensor – should be based on domain expertise
- Variance checks – indicator of sensor degradation
  - Running averages or change in slope checks, e.g., outlier detections, spikes
  - Sensitivity is specific to site and sensor type
- Persistence checks
  - Check for repeating values that may indicate sensor failure
    - E.g., freezing, sensor capacity issues
- Internal (plausibility) checks
  - E.g., TMAX-TMIN>0, snow depth> SWE
  - Consistency of derived values
- Spatial checks
  - Use redundant or related sensors, e.g., sensor drift

Quality control on Streaming Data: Data Levels: Gap Filling

- Level 2
  - Gap-filled, estimated, or aggregated data
  - Involves interpretation – multiple algorithms possible
    - different methods will lead to different products
    - some researchers may still want to download Level 1 data to apply preferred methods

- Obligation to provide gap filling?
  - Controversial – can seriously compromise stats, analyses and lead to misinterpretation
  - Desirable when generating summarized data, but transparency critical
  - Probably unsuitable for streaming data – much later in data cycle with expert attention
  - Most critical to document gap-filling, and flag all estimated values to allow removal

Quality control on Streaming Data: Data Qualifiers

- Many vocabularies – desirable to harmonize, but impractical (may crosswalk across vocabularies)
- Good approach
  - Rich vocabulary of fine-grained flags for streaming data – intended to guide local review
  - Simpler vocabulary of flags for "final" data for public consumption, e.g.,
    - 'Verified', 'Accepted', 'Suspicious', 'Missing', 'Estimated'
    - Pass-fail indicator (include in analysis?)
- Certain types of qualifiers may be better as data columns
  - Method shifts, sensor shifts
  - Place key documentation as close to data value as possible

Knowledge Base: A Best Practices Manual for Networked Sensors

- Online guide which summarizes the community's collective knowledge
- Organize by topics
  - Summary of topic
  - Populate through community crowdsourcing
    - 1-pagers to highlight expertise, experience
  - Cite various protocols, e.g., USGS, NOAA, NERRS, NEON, US Forest Service, CUAHSI
  - Let a page "manager" be responsible for updating the summary periodically
  - Discussion blog associated with each topic

http://im.lternet.edu/resources/im_practices/sensor_data