Report from the IT Working Group at GLEON6

Areas identified that need development/attention

- Need clear path for decisions with respect to IT
- Way for sites to sharing info/experiences
    - archived e-mail list/wiki
- Data discovery
- Lake Info DB
    - Make Lake Info DB searchable
    - Decision on scope of Lake Info DB
    - Add contact, data policy
- Deployment package
    - User interface enhancement
    - Remote deployment (robustness)
    - Interfaces to instruments,
    - Interfaces to database
    - Configuration tools
    - Add data quality screening
- QA/QC
    - Best practices/policies
- Data sharing
    - Data schema survey

Clear there is a need for **ongoing working group in IT**.

Tony Fountain
Kye Ewing
Vanesa Yael Bohn
Luke Winslow
Tony Fountain
Sameer Tilak
Fang-Pang Lin
Barbara Benson
Kenneth Chiu
Hsiu-Mei Chou

Working group co-chairs will be Barbara and Ken.

Summary reports from the breakout discussions with the lake scientists:

Group 1 (Fang-Pang)

- Data sharing is crucial.
- Raw data is more crucial than QA/QC data.
- Local scientists should decide what algorithms.
- Some basic like range-checking should still be provided and

visualization.
- Many people wanted to contribute historical data. Need to think about data archiving for that kind of data.
- Human protocols and guidelines are very important.
- Tutorials are crucial for sure.
- Existing controlled vocabulary is not enough.
- Units, sampling freqs etc. are important.
- QA/QC: Different researchers have own QA/QC.
- IP issues.
- Good to include GIS data and other kinds of data.

Group 2 (Luke):

- No matter what they got in terms of QA/QC, they wanted to know what kind they got.
- Wanted contact information for someone who was responsible for that data.
- Federation? No consensus.
- Search on variables, like DO, temp, windspeed.
- Lakes need to identified by more than name.
- Desire to be able to quickly determine characteristics temporal qualities.
- Need someway to discover and locate associated data, such as local met data.

Group 3 (Tony):

- Three examples:
  - Metabolism: Exchanging particular types of data.
  - Computed derived data from metabolism.
  - Integrate with one more type of data.
- Investigated what steps are necessary to do something like this.
- Models need to be transparent.
- Policy on data sharing.
- Instead of full data sharing, just a catalog.
- Organismal data.
- Time series with transect data.

Group 4 (Sameer)

- Three categories of priorities: A, B, C.
- QA/QC is very important, maybe some canonical algorithms.
- Querying requirements.
- GIS data, spatial data.
- Lake specific metadata and sensor metadata.

- New variable should be automatically propagated.

Appendix: Breakout group reports

Group 1 (submitted by ??)

Overarching: Knowledge sharing vs Data sharing

(1) Storing raw data is critical, but QA/QC methods might vary from one site to the others. Basic data services can be build upon raw data.

(2) Data archiving for historical data.

(3) Protocols and Guidelines required for QA/QC, Instrument Management.

(4) Tutorials

Querying Data

- Scenario Study lake dynamics for say Sunapee, YYL, and Denmark
- Current metadata and control vocabulary is not sufficient:
  - Needs DO, Temp from different lakes in certain time period
  - Additional data needed: water chem, meteorology, land use
  - Also needs to know metadata such as units, sampling frequency, time zone
- Requirements for various QA/QC functions.

QA/QC – concern/needs – identify problem

- where should the QA/QC scheme in data stream happen
- how to handle post-processed/corrected data
- who should be responsible for that

Note: no matter what and how QA/QC will be done, we should always keep one copy of raw data.

Suggest generating GLEON QA/QC Protocols

- by sensor, by condition/site
- must be flexible
- can flag data since calibration
- can use CI help – calibration notices/versioning
- GLEON can provide basic description of these flags

Note: Generate subcommittee to develop QA/QC standards

Other issue related to QA/QC:

- Do we need to coordinate sampling frequencies among site?
- How to handle data aggregation, can GLEON provide advice on variety of ways?
- Data governance: log of data access, report on what have they done

Instrument Management

- calibration history: time stamp/log file for each station
- specific/guideline/protocol template; shared vocabulary for sensors to which drift connections should be applied, supply raw information

Training

workshop/tutorials for analyzing sensor data, QA/QC, bouy deployment and maintenance, etc.

Provide guidance for how scientists/students use/share common data
Suggest using tools like wiki, internet courses to accumulate and share knowledge
Knowledge sharing by having workshop, or using internet tools/ videoconferencing tooks, or site exchanges.

Group 2 (submitted by Thorsten Blenckner):

- Provide a template, so we can set example
- storage of data
- link in other data into the dataset, historical data
- tag sensors for exchanging sensors
- contact name/info in the meta-databata very important
- information on how the data are downloading, and how often they are downloaded
- search for multiple variables, variable groups
- ability to get GPS coordinates of sites
- quick overview of what data (temporally) are available, like USGS
- have google earth image with each of the sites
- additional information of additional data which could be available elsewhere - added to site information

Group 3 (submitted by Tony Fountain)

3 proposed science studies to derive IT requirements.

1. Cross-site metabolism study
* main requirement is to collect data from multiple sites
* want to be able to query across variables, measurements, sharing policy, and quality metrics

2. More advanced study. Similar to #1 with additional need to compute GPP/R. So, need additional analysis tools.
3. More advanced yet. Similar to #2 with additional requirements of accessing more data, especially 3rd party or agency data. Perhaps similar to CUASHI web services for agency data. Also need visualization tools.

Workflow for performing analyses:

1. Discover data (and resources)
2. Select
3. Integrate
4. Select model
5. Configure model
6. Check data for quality and appropriateness to analysis/model
7. Analyze data
8. Visualize results

Registration system -- possible first step towards a fully automated system is a registry of system resources. A searchable system

that contains info on data/measurements, coverage, policies, and contact info.

Description vs. standards -- for some data it may be better describe the data fully rather than enforcing standards. For example, time. As long as the specific time format is described, then it can be translated.

Need additional types of data in addition to sensor data, including organismal data and transect data.

Need ability to query specific measurement values/ranges that satisfy specific criteria. For example, ranges of temp values.

Need to keep raw data and calibration info and calibration info for QA/QC. Need automated sanity checks, e.g., sensor drift.

Group 4 (submitted by ?)

- QA/QC [A]
    - Very important
    - Algorithm for canonical data – thermistor chains.
    - Filters (9999) to automate data anaysis.
    - Where should it happen – site/network/both etc.?
    - Science folks need to give canonical QA/QC algorithms. Have GLEON set of policies – set of recommendation to sites about how they implement. Some did not agree with that and said "GLEON should not control the sites in terms of QA/QC." Some said, "Some basic QA/QC must happen for all the network-level data and GLEON should help sites/researchers with that. " Need access to both raw and processed data.
- Metadata about the sensors important [A]
- Need for IT committee for ratification. Not all members are informed. Formalizing the process and need to inform everyone. [A]
- Querying requirements [A]
    - Fundamental questions: What, where, When is it measured?
    - Compare multiple datasets – different time scales, units etc. [DbBadger should be sufficient.]
    - Should we propose some standard interpolation procedures? [DbBadger should be sufficient.]
    - Is there is a data overlap across two sites measuring temp during this time scale. Google interface to show sites etc.? [Important but not high priority. B+]
    - Add-ons developed by individual site should be given back to the network. [C]
- Controlled vocabulary (variable names, units etc.) needs ratification.
- Training
- Display/Visualization
- Flexibility of DB schema

- Site Autonomy and configuration changes
- Data discovery
- Spatial data. Current system single location time series data. Do we have requirements to access multiple data sources including spatial/organism related data? Yes. Aggregate not just variables, but also sites.
- What GLEON network can provide for general public?
- Do we need to bring into other data – digital elevation models, spatial data? Should GLEON provide resources to that end? [].
- Extend the system by adding custom plug-ins.
- Data modeling and event detection/anomalies?

There is a need a mechanism to connect buoy data with other supporting data. This is not a simple process but should be a part of intermediate or long-term plan. Add way to add supporting data. Each site has the ability to point to their data. Lake-specific metadata more categorized and should be more searchable.
1 day before GLEON training session for eco<->It folks. [A]