The Director’s View of the CUAHSI Water Data Center

Alva L. Couch
Acting Director,
CUAHSI Water Data Center
acouch@cuahsi.org
Purpose of this talk

• To **define the roles** of the Water Data Center and explain how strategic decisions are made.

• To **open a long-term dialog** between the Water Data Center and CUAHSI users, data publishers, developers.

• To **develop and define language** with which to express and discuss the “burning issues” in Water Data Center activities.
Water Data Center Status

• Funded by NSF effective April 1, 2013.

• Two software engineers hired:
  – Steve Pilot (started April 1)
  – Marie Martin (started April 21)

• User support continues to be provided by
  – Jon Pollak
A year of transition

• From “prototype” to “product”.
• From “informal” to “more formal” agreements with data publishers and users.
• From “unmoderated” to “curated” data sources.
• From “best effort” services to “Service Level Objectives” (SLOs).
What isn’t changing?

• Infrastructure support and updates for:
  – The HISCentral Catalog and discovery services.
  – HydroServer, WaterML and WaterOneFlow services.
  – Translators from non-WaterML to WaterML.

• User support for:
  – HydroDesktop, HydroExcel, other data clients.
  – (HydroDesktop development continues separately).
Short term goals

• Take control of HIS Central.
• Improve reliability via Azure cloud.
• Improve search speed via parallel computing.
• Increase precision of search queries and results.
• Improve site registration/harvesting processes.
• Improve catalog quality via rigorous curation, computed metadata.
• Enable “linked data” in catalog and servers.
Improving Discovery Precision

• Ensure completeness of metadata via curation.
• Allow users to search for variables instead of just concepts.
• Allow users to specify other known facets during discovery.
• Compute derived data quality metadata: values/month, frequency of NULLs, etc.
Improved registration and harvesting

• Current registration process involves “hurry up and wait”.
  – Register server.
  – Wait for harvest.
  – Tag variables.

• Proposed process:
  – Gives immediate harvesting feedback upon registration.
  – Allows easy correction of errors and immediate tagging using the community-moderated variable Controlled Vocabulary.
  – A “one stop” process.
Insuring Catalog Quality

• One can measure data discovery in terms of two factors: precision and recall:

• **Precision** is the extent to which what you get is precisely what you want.

• **Recall** is the extent to which what you want is part of what you get.

• Catalog curation improves recall, but does not improve precision.

• One way to improve precision: “linked data.”
“Linked Data”

- “Linked data” refers to the process of including semantic references in data that relate that data to other similar data.
- Simplest form: linking via metadata.
Enabling linked data

• Currently, metadata contains “words” that describe data.
• In the future, these “words” become Uniform Resource Indicators (URIs).
• Unlike a word, a URI can embody
  – Meaning.
  – Synonyms.
  – Synonyms in other languages.
Before linked data

Linkage by use of common terms

Dataset 1

metadata

“Nitrogen”

Some synonyms

e.g., HIS Central

Dataset 2

metadata

“Nitrogen”

Other synonyms

e.g., GEOSS
After linked data

Linkage by use of **cross-referenced URIs**

![Diagram]

- Dataset 1
  - CUAHSI
    - metadata
      - CUAHSI SKOS URI for concept “Nitrogen”
      - Synonyms we use
  - cross-reference

- Dataset 2
  - GEOSS
    - metadata
      - GEMET SKOS URI for concept “Nitrogen”
      - Synonyms proposed by others, or in foreign languages
The importance of linked data

• Before linked data, synonyms are a local concept, in the mind of the local publisher or publishing infrastructure.

• After linked data, synonyms are a global concept, with cross-references to other global repositories of synonyms, even in other languages.
Short-term plan for enabling linked data

- Expose variables via SKOS services
- Embed SKOS links in HIS Central Metadata
- Return SKOS references to HIS Central queries
- Cross-link CUAHSI SKOS other SKOS servers:
  - GEOSS GEMET
  - CUNY EDSC
On the immediate horizon

• WaterML 2, ODM 2
• Interoperability with OGC GEOSS
WaterML 2, ODM 2

• Expecting ODM 2 definition this summer.
  – Requires changes to HydroServer.
  – That in turn produce native WaterML 2.
  – That in turn requires changes in the catalog.

• Most important change: “linked data”
Toward a world water data infrastructure


• Linkage to OGC GEOSS linked data vocabulary servers.
Longer term goals

• Partnerships with scientists and industry to extend the state of the art
  – HydroShare: enable sharing of bundles of data.
  – EarthCube: enable interoperability between services providing different kinds of data.
  – Plug and play sensors: enable vendor-specific sensors to be deployed instantly and seamlessly.
Why?

• At this point, you might be overwhelmed with details of what we’re planning to do.
• But I’d also like you to understand “why” we do things.
A typical data center quandary

• Comparing *cost* and *value*.
  
  – Defining areas for improvement and Service Level Objectives (SLOs).
  
  – Quantifying the cost and value of each improvement and SLO.
Water Data Center Governance

• Final authority: CUAHSI Board of Directors, with assistance from:
  – Users Committee: user priorities and objectives.
  – Informatics Standing Committee: infrastructure priorities and objectives.

• The Data Center director’s view of governance
  – “The Board” and designees set priorities (**value**).
    • Scientific worth/merit/enabling potential.
  – “The Data Center Director” predicts and tracks **cost**.
    • Labor, infrastructure, services.
  – There is an **ongoing dialogue** to balance cost and value.
The director’s view of costs

• Everyone knows that
  – Labor isn't free.
  – Hosting isn’t free.

• But also:
  – Data storage isn’t free.
  – Performance isn’t free.

• Their cost depends upon user expectations, also called Service Level Objectives (SLOs).
Data storage isn’t free

• One pays for
  – Volume of storage (gigabytes).
  – Time of storage (months).
  – “Transactions” when storage is accessed.
  – US unit of storage cost: $/GB/month

• Some benchmarks:
  – Least expensive high-reliability, high-availability cloud storage: $.07/GB/month.
  – Most expensive: $.15/GB/month.
  – Nominal: $.10/GB/month.
  – (Breaking news: Azure just reduced its cost from $.15 to $.10 to compete with Amazon)
The cost of persistence and reliability

• A common myth: buy a $150 3-TB drive, put it on the web... done!
• What’s wrong with this picture?
  – What happens if it fails/is struck by lightning/is stepped upon?
  – How long does it take to recover when it fails? Can it be recovered at all?
  – How quickly can several concurrent users access it?
• Making a $150 drive reliable, persistent, and highly available costs a lot more than $150!
An interesting aside

• In the high-capacity disk market, there is a non-linear and extreme jump in cost at around 10 TB.

• < 10 TB: home video. < $50/TB
  – relatively low reliability/capacity. No recovery strategy.

• > 10 TB: multi-user access. > $500/TB
  – high reliability, fast recovery, high request capacity.
Components of Data Storage Cost

- What is one paying for?
  - **High reliability**: the data source will not stop working.
  - **High persistence**: when you put something there, it stays there.
  - **High availability**: the data source handles multiple simultaneous uses.
  - **Quick recovery**: in the unlikely event of a failure, data becomes available again quickly.
“Best Effort” versus “Service Level Objectives”

• The typical SLO for prototypes: “Best effort:” if something goes wrong, we’ll fix it when we get “a round tuit.”

• The typical SLO for a production service: if something goes wrong, we’ll be back online in one hour/day/week...
Moving away from “best effort”

• Currently, CUAHSI services expose about 70 TB of data.

• In my opinion, about 10 TB is “at risk” of disappearing without notice:
  – When funding ends.
  – When graduate students leave.

• The quandary: is the value of backing up the 10 TB worth the cost?
Hosting new data products

• Traditionally, hosting of hydrologic data has been a “gigascale” problem:
  – At most GB of data.
  – Current catalog: about 100 GB.

• Now, people are coming to me with “terascale” data they want us to host:
  – Satellite derived products.
  – Model outputs.

• Question: is the value to the community worth $.10/GB/month = $100/TB/month?
Inside $.10/GB/month

• It might seem that the “cloud rate” of $.10/GB/month is very expensive.
• ... until one compares it to the lifecycle cost of setting up a large disk array ...
• Same rough lifecycle cost, plus **astronomical** initial hardware and implementation cost.
Beyond $.10/GB/month

• There are other options, including modified SLOs:
  – “Nearline storage” ($.01/GB/month + access charges based upon access speed): store something on tape, move it to disk when requested (typically one day turnaround).
  – “Unreliable caching”: store in a large rack of slow, lower-quality disks, refresh from web periodically, do not back up. Approximate cost: $100/TB/year; one must include network and electricity costs!.

• But are either of these acceptable to users?
## Components of Value and Cost

<table>
<thead>
<tr>
<th>Value</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Number of users enabled.</td>
<td>• Labor.</td>
</tr>
<tr>
<td>• New capabilities/uses.</td>
<td>• Computing cycles.</td>
</tr>
<tr>
<td>• Scientific/transformational impact.</td>
<td>• Data storage fees and storage transaction fees.</td>
</tr>
<tr>
<td>• Education, outreach and visibility.</td>
<td>• Software licensing.</td>
</tr>
<tr>
<td></td>
<td>• Electricity.</td>
</tr>
<tr>
<td></td>
<td>• Network connectivity.</td>
</tr>
</tbody>
</table>
A “big board” of cost and value: all positions approximate

- Adapt satellite/radar data products
- Interoperate with GEOSS
- Catalog and Data curation
- Improve search performance
- “Linked data”
- Search for variables
Questions?

Alva L. Couch  
Acting Director  
CUAHSI Water Data Center  
acouch@cuahsi.org