Integrated Data Management System for Critical Zone Observatories

CZOData II

Mark Williams, UC-Boulder.
Anthony K. Aufdenkampe, SWRC.

Kerstin Lehnert, IEDA/Columbia.
Ilya Zaslavsky, SDSC.
David Tarboton, USU
Jeff Horsburgh, USU.
Emilio Mayorga, UW-APL
Goals for CZOData II

- extensive and iterative interaction and feedback from the community of CZO PIs, scientists and data managers
- uniform web portal appearance for the CZO sites and the national CZO program
- development of a consistent metadata strategy for CZO data, supported by a respective collection of data submission forms and tools
- enhancing publication and data discovery workflows for geochemical, hydrologic, spatial and other data
- creating a uniform data discovery portal
- ensuring that the data descriptions follow consistent semantics
- integrating with the EarthChem system
- developing a consistent online data visualization interface for CZO time series data
Community Involvement

• Instigate and Support an Information Management Committee (IMC)
  • 1-2 investigators/CZO + site data managers
  • Monthly telecon & annual face to face meeting w/ CZOData developers
  • Feedback to CZOData team
    • Data use scenarios
    • Meta-data requirements, shared vocabulary, etc.
• Web-based information events and workshops (>2/year); mailing lists
• Subdiscipline workshops (three workshops)
  • Hydrology (all sensor-based data)
  • Geochemistry (all sample-based data)
  • Geospatial data
• Synthesis working group (two workshops)
Get Started Now

• Form IMC and set first workshop date
• Content for new website
• LIDAR to OpenTopography
• Start registering samples with SESAR
• Start registering datasets with IEDA
Challenges to CZO Data Management

- Atmosphere
- Biosphere
- Hydrosphere
- Lithosphere

Many Object & Data Types!

- Diverse media
- Sensor-based
  - Stationary
  - Mobile
- Spectra/photos
- Sample-based
  - Sub-samples
- Preparations/Fractions
- Numeric & Categorical

Hillslope ↔ Catchment ↔ Watershed
Examples from Different Disciplines

- Climate & Hydrology
  - Point observation (sensor) time series
  - Raster observation (remote sensing) time series
  - Vector networks for water routing
- Geochemistry
  - Sample-based lab analyses
- Geophysics
  - Seismic and other subsurface profiles
- Biology
  - Phylogenetic trees
Sensor- vs. Sample-Centric Data

- Sensor-centric Data Models (i.e. ODM)
  - Site $\rightarrow$ DataValue

- Sample-centric Data Models (i.e. EarthChem)
  - Site $\rightarrow$ Sample/Subsample $\rightarrow$ Prep/Batch $\rightarrow$ DataValue
GeoChemical Data Model

- Feature of interest
- Sample
- Analysis
- Observed value
- Data source
- Method/DQ
- Collection, geospatial
- Material preparation, obs. point
Sample Fractions for Soil Geochemistry

Ziplock (~500g)  
Bulk soil horizon or depth increment

DRY SIEVE
2 mm
<2mm

glass vial:
<2mm fines dry sieved

EA-IRMS
FTIR
SA

WET SIEVE, or DENSITY, or SETTLING
(with or without sonication)

The choice here is important. Do we want aggregates or not?

glass vial:
sand + small detritus

SA
XRD
CEC

SPEX mill

glass vial:
silt + clay

SA
XRD
CEC

SPEX mill

Extractions
Dithionite-Citrate extraction
Na pyrophosphate extraction
Ammonium oxalate extraction

>2mm:
(1) Pick out plant roots & detritus, rinse with DI water, oven dry, mill (SPEX?)

glass vial:
plant detritus milled

EA-IRMS
FTIR

(2) Remaining pebbles & rocks, hard grind

glass vial:
pebbles hard ground

EA-IRMS
FTIR

ICP-MS after Li-borate fusion

Al Can (~70 g)  
For Gamma Counting 137Cs
Needed Capabilities for ODM

**Sample** table

- Optional direct link between Sample & Site
  - Need to assign SampleID before data values exist
  - Natural one-to-many hierarchy
    - 1 site $\rightarrow$ many samples, 1 sample $\rightarrow$ many values
- Recursive parent-child relationships
- Sample metadata
  - Medium, fraction, preservation, container, dilution, etc.
Notes:

- Method Type should not include analysis method, b/c its in the values table.
- LIMS info is recorded in Values table (i.e. sample amount, budget #, dilution ratio, sample location, container type)
- Analysis “Batch” or “Run” is treated as a sample group
- ParentSample table allows for composite samples
Notes:

- Sample Groups: Analysis Batch, Profile, etc.
- Site Groups: Transect, station, observatory, etc.
- Value Groups: ???
- Person Groups: ?Research Teams, etc.?
<table>
<thead>
<tr>
<th>ODM v1.1</th>
<th>Suggested</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sources/Institution</td>
<td>Person → Institution</td>
</tr>
<tr>
<td>Soil/Sed intervals</td>
<td>OffsetValueMin &amp; OffsetValueMax</td>
</tr>
<tr>
<td>Only one offset value</td>
<td>add horizontal offsets?</td>
</tr>
<tr>
<td>Horizon Descriptions?</td>
<td>Add DataValueNote field to Data Values</td>
</tr>
<tr>
<td>Methods table insufficient</td>
<td>add MethodType, PersonID, etc.</td>
</tr>
<tr>
<td>CensorCode insufficient</td>
<td>Need value field (i.e. Method Detection Limit)</td>
</tr>
</tbody>
</table>

**Other outstanding issues:**
- Do spatial offsets also belong in samples table? [yes]
- Spectral data, photos?
- Dataset versioning
Importance of Sample/Site Tracking

• CZO scientists share samples!
  • Data often needs to be merged at level of subsamples

• SWRC’s biggest data management headaches always come from merging data from different instruments/labs by sample and by site.

• International Geo-Sample Number (IGSN) is the answer!
### Existing
- Core
- Core half round
- Core quarter round
- Core piece
- Core section
- Core section half
- Core sub-piece
- Core whole round
- Cuttings
- Dredge
- Grab
- Hole
- Individual sample
- Oriented core
- Other
- Rock powder
- Terrestrial sample
- Trawl

### Considering?
- **Sampling events:**
  - holes, cores, dredges, stratigraphic sections
- **Individual samples:**
  - Specimens, rocks, minerals, fossils, precipitates, synthetic material, etc.
  - Fluid samples: seawater, hydrothermal fluids, groundwater, etc. (to be completed)
  - Particulates: aerosols, suspended matter
  - Soil pedons and samples thereof
- **Sub-samples of any of above:**
  - processed samples such as mineral or fossil separates, leachates, thin sections, etc.

http://www.geosamples.org/sampletypes
CZO Geo-Object Types

• Site/Location (x,y, z treated via vertical offset)
  • surface water station, well, lysimeter, piezometer, soil pit, borehole, monument, meteorological station/tower, tree?

• Fluid Sample (Water Sample?)
  • stream/river water, pond/lake water, wetland surface water, groundwater, soil water (unsaturated), sediment porewater, sap?

• Gas Sample (also fluid?)
  • atmospheric gas, dissolved gas, soil gas

• Soil/Lithology/Sediment Sample (need help with names)
  • Surface grab, core, auger interval, pit interval, rock, saprolite?, bedrock?, cuttings?

• Plant Sample
  • Whole plant, tissue, ???
CZO Sample Fraction Types

• Subsample
  • Duplicate or split that does not fractionate whole sample

• Size Fraction
  • i.e. > 2 mm, 63-2000 um, <63 um

• Extracted Fraction
  • Acid soluble, total lipid extract, dithionate-citrate-bicarbonate extract

• Extraction residue
  • etc.
Generalized, Extensible ODM Suggested v2

• After discussions and on plane ride home
## ODM v1.1

**DataValues**
- ValueID (PK)
- DataValue
- ValueAccuracy
- LocalDateTime
- UTCOffset
- DateTimeUTC
- SiteID (FK)
- VariableID (FK)
- OffsetValue
- OffsetTypeID
- CensorCode
- QualifierID
- MethodID (FK)
- SourceID (FK)
- SampleID (FK)
- DerivedFromID (FK)
- QualityControlLevelD

## Suggested v2

**DataValues**
- ValueID (PK)
- DataValue
- LocalDateTime
- UTCOffset
- DateTimeUTC
- SiteID (FK)
- VariableID (FK)
- CensorCode
- MethodID (FK)
- SourceID (FK)
- QualityControlLevelD

**DataValuesExtension**
- DataValueExtensionID
- DataValueID (FK)
- AttributeID (FK)
- DataValueAttributeValue

**Attributes**
- attributeID (PK)
- attributeType (CV)
- attributeDescription
- Units (FK)

**Example Attributes:**
- Offset, OffsetMin, OffsetMax,
- QualifierID, DerivedFromID,
- ValueAccuracy,
- InstrumentType, InstrumentID (FK), SensorID (FK)
- AnalysisNote, DataValueNote,
- ProjectName, CensorType,
- CensorLimitValue

**AttributeTypes:**
- Correspond directly to table that is being extended.
- i.e. Site, Sample, Value

**Bold fields in tables are required**
**Non-bold fields are optional**
**ODM v1.1**

**Samples**
- SampleID (PK)
- SampleType
- LabSampleCode
- LabMethodID (FK)

**Suggested v2**

**Samples**
- SampleID (PK)
- SampleCode
- SampleNote
- IGSN
- IsFieldSample
- LocalDateTime
- UTCOffset
- ObjectTypeID (FK)
- FractionTypeID (FK)
- MethodID (FK)
- SourceID (FK)

**Table Notes**
- alpha-numeric, ~ 20 char
- ~200 char
- Intl. Geo-Sample Number
- Y/N, to distinguish ultimate parent
- Creation, when container filled
- Corresponding to IGSN Object Types
- i.e. whole sample, >63 um, acid extract
- Types: collection or prep., not analysis
- Who performed method above

**SamplesExtension**
- SampleExtensionID (PK)
- SampleID (FK)
- AttributeID (FK)
- SampleAttributeValue

**Example Attributes:**
- VerticalOffset, VerticalOffsetMin, VerticalOffsetMax, HorizontalOffset, HorizontalOffsetDirection (deg.), Medium, AlternateSampleCode, FieldCampagneName, Amount, StorageLocation, ContainerType, DilutionRatio, CollectionNote, PreparationNote, FractionNote, IsExperimentalSample, ExperimentID

**ParentSampleXRef**
- ParentSampleID (FK)
- ChildSampleID (FK)

**SiteSampleXRef**
- SiteID (FK)
- SampleID (FK)
<table>
<thead>
<tr>
<th>ODM v1.1</th>
<th>Suggested v2</th>
</tr>
</thead>
</table>

### Sites

<table>
<thead>
<tr>
<th>SiteID (PK)</th>
<th>SiteCode</th>
<th>SiteName</th>
<th>Latitude</th>
<th>Longitude</th>
<th>LatLongDatumID (FK)</th>
<th>Elevation_m</th>
<th>VerticalDatum</th>
<th>LocalX</th>
<th>LocalY</th>
<th>LocalProjectionID (FK)</th>
<th>PosAccuracy_m</th>
<th>State</th>
<th>Country</th>
<th>Comments</th>
</tr>
</thead>
</table>

### SitesExtension

<table>
<thead>
<tr>
<th>SiteExtensionID</th>
<th>SiteID (FK)</th>
<th>AttributeID (FK)</th>
<th>SiteAttributeValue</th>
</tr>
</thead>
</table>

**Example Attributes:**

- **From ODM 1.1:** SiteDescription, LocalX, LocalY, LocalProjectionID (FK), PosAccuracy_m (LatLongAccuracy_m?), City/Township, State/Province, Country, Comments
- **From IGSN/SESAR:** Physiographic feature, Name of physiographic feature, Location description, Locality, Locality description, Field Program/Cruise, Platform type, Platform name
- **From Sue Brantley:** annlPrecip_mm, anlMeanTemp_oC, slopeDeg, aspect, landscapePosition, landUse, vegSpecies, parentLithology, exposureAge, erosionRate, depthToRock_m, soilTaxonomy, SSURGO_ID (FK), siteNote, ContactName
- **From SWRC:** AlternateSiteCode, WatershedName, HUC, ElevationAccuracy,
### ODM v1.1

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
</table>
| **GroupDescriptions** | GroupID (PK)  
GroupDescription |

### Suggested v2

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
</table>
| **Groups** | GroupID (PK)  
GroupCode  
GroupDescription  
GroupTypeID (FK) |

| Table Notes | alpha-numeric, ~ 20 char  
~200 char  
Types: Value, Sample, Site, Person?, etc. |

| **Groups** | GroupID (FK)  
ValueID (FK) |
| **SiteGroupsXRef** | SiteID (FK)  
GroupID (FK) |
| **SampleGroupsXRef** | SampleID (FK)  
GroupID (FK) |
| **ValueGroupsXRef** | ValueID (FK)  
GroupID (FK) |

### Notes:

- Sample Groups: Analysis Batch, Profile, Experiment, etc.
- Site Groups: Transect, station, observatory, etc.
- Value Groups: Profile, Analysis, Spectra
- Person Groups: ?Research Teams, etc.?
<table>
<thead>
<tr>
<th>ODM v1.1</th>
<th>Suggested v2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Methods</strong></td>
<td><strong>Method</strong></td>
</tr>
<tr>
<td>MethodID (PK)</td>
<td>MethodID (PK)</td>
</tr>
<tr>
<td>MethodDescription</td>
<td>MethodCode</td>
</tr>
<tr>
<td>MethodTypeID (CV)</td>
<td>MethodDescription</td>
</tr>
<tr>
<td>MethodLink</td>
<td>SourceID</td>
</tr>
<tr>
<td>Sources</td>
<td>Sources</td>
</tr>
<tr>
<td>SourceID (PK)</td>
<td>SourceID (PK)</td>
</tr>
<tr>
<td>Organization</td>
<td>SourceDescription</td>
</tr>
<tr>
<td>SourceDescription</td>
<td>SourceDescription</td>
</tr>
<tr>
<td>SourceLink</td>
<td>SourceLink</td>
</tr>
<tr>
<td>ContactName</td>
<td>Corresponding PersonID (FK)</td>
</tr>
<tr>
<td>Phone</td>
<td>PersonID (PK)</td>
</tr>
<tr>
<td>Email</td>
<td>LastName</td>
</tr>
<tr>
<td>Address</td>
<td>FirstName</td>
</tr>
<tr>
<td>City</td>
<td>Phone</td>
</tr>
<tr>
<td>State</td>
<td>Email</td>
</tr>
<tr>
<td>ZipCode</td>
<td>InstitutionID (FK)</td>
</tr>
<tr>
<td>MetadataID (FK)</td>
<td>PersonLink</td>
</tr>
<tr>
<td>Institutions</td>
<td></td>
</tr>
<tr>
<td>InstitutionID (PK)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
DataSeries: better sample integration?

- Uses DataSeries table modified from HydroDesktop (next page)
- DataSeries Table can act as a XRef Table, but requires creation of a data series upon registration of the FieldSample
- A DataSeries from/for a single sample can be viewed as equivalent to EarthChem’s Analysis table
- Joins do not require passing through huge DataValues table
**HydroDesktop**

**Suggested ODM v2**

**DataSeries**
- SeriesID (PK)
- SiteID (FK)
- VariableID (FK)
- IsCategorical
- MethodID (FK)
- SourceID (FK)
- QualityControlLevelD
- BeginDateTime
- EndDateTime
- BeginDateTimeUTC
- EndDateTimeUTC
- ValueCount
- CreationDateTime
- Subscribed
- UpdateDateTime
- LastCheckedDateTime

**DataValues**
- ValueID (PK)
- SeriesID
- DataValue
- ValueAccuracy
- LocalDateTime
- DateTimeUTC
- UTCOffset
- OffsetValue
- OffsetTypeID (FK)
- CensorCode
- QualifierID
- SampleID (FK)
- FileID (FK)

**Example DataSeriesAttributes:**
- BeginDateTime, EndDateTime, BeginDateTimeUTC, EndDateTimeUTC, ValueCount, CreationDateTime, Subscribed, UpdateDateTime, LastCheckedDateTime, InstrumentType, InstrumentID (FK), SensorID (FK), PlatformID (FK), AnalysisNote, UTCOffset,
Thank You

Funded by:
US National Science Foundation
Figure 3. Excerpts from tables illustrating the population of ODM with streamflow gauge height (stage) and discharge data.