*mayorga@apl.uw.edu

https://github.com/BiG-CZ/CZIMEA

EarthCube Integrated Activities https://earthcube.org

http://bigcz.org

http://odm2.org

CUAHSI HydroInformatics Conference, Tuscaloosa, AL, 2017 Sept. 26

Cross-Site Soil & Microbial Ecology Cyberinfrastructure for the CZIMEA EarthCube Project

Emilio Mayorga[1]*, Landung (Don) Setiawan[1], Keshav Arogyaswamy[2], Miguel Leon[3], Emma Aronson[2], Aaron Packman[4]

[1]Applied Physics Laboratory, University of Washington, Seattle, WA; [2]University of California, Riverside, CA; [3]University of Pennsylvania, Philadelphia, PA; [4]Northwestern University, Evanston, IL

1. Introduction

The **CZIMEA (Critical Zone Integrative Microbial Ecology Activity)** EarthCube Integrative Activities project is carrying out cross-site soil sampling and analysis involving many universities affiliated with the **10 Critical Zone Observatories (CZO) across the US**. The scientific goal is to gain insights into the differences between soil microbial communities as they vary across ecosystems and with depth, using a wide range of soil and environmental methods and both metagenomic and amplicon high-throughput sequencing to analyze nearly 200 unique soil samples.

To facilitate management, integration and dissemination of the cross-site metadata and data generated by this activity, we are storing information using the **Observations Data Model 2 (ODM2)** system with the **ODM2-Admin User Interface**, deployed on a Cloud instance managed by the CUAHSI Water Data Center. ODM2 is enriched with cross-linkage to external data systems using universal identifiers that include DOI's, ORCID's and IGSN's, in addition to ODM2-managed controlled vocabularies available online as SKOS. The CZIMEA ODM2 relational database holds extensive, structured sampling and sample metadata; it will soon store environmental measurements. Genomic results will be stored externally in dedicated systems, linked to via universal identifiers.

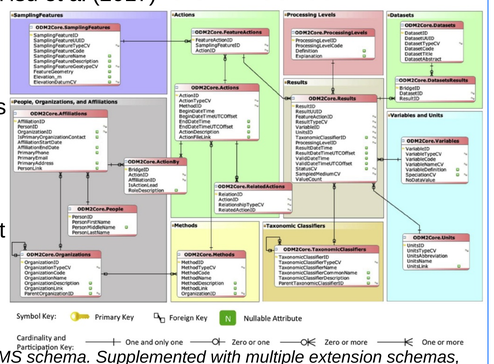
The ODM2 data system is designed to enable access via web services. We plan to assess and leverage discovery and other capabilities being developed by EarthCube. The project is also collaborating with the BiG CZ project (Software System for Integration and Analysis of Bio- and Geoscience Data in the Critical Zone) as a concrete use case, and to leverage the capabilities under development by that allied project. We will describe the current cyberinfrastructure being used and upcoming enhancements.

2. Background: ODM2 and ODM2 Admin

ODM2

http://odm2.org

- An **Information Model** and Software Ecosystem for Spatially-Discrete Earth Observations, designed to enhance the integration and management of **sensor and sample-based data**, and interoperability across scientific disciplines and domain cyberinfrastructures
- Horburgh et al (2016); Hsu et al (2017)
- A practical profile for OGC's Observations and Measurements
- Controlled vocabularies
- Multiple **physical implementations** (file, database, object), including 4 Relational Database Management Systems (**RDBMS**): SQLite, MySQL, PostgreSQL, SQL Server



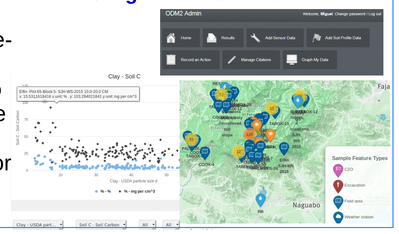
Core RDBMS schema. Supplemented with multiple extension schemas, including Sampling Features, Results, Annotations, and External Identifiers

- **Open-source software ecosystem** (**http://github.com/ODM2**): Underlying Python SQLAlchemy and Django APIs; GUI Desktop and Web Applications; Web services
- **Growing community:** initially funded by NSF; continued funding from many sources; 26 contributors on GitHub. Used by: *HydroShare* (**https://www.hydroshare.org**), *EnviroDIY* (**http://data.envirodiy.org**), *iUtah* (**http://iutahepscor.org**), *Luquillo CZO* (**http://criticalzone.org/luquillo/**), others

ODM2 Admin

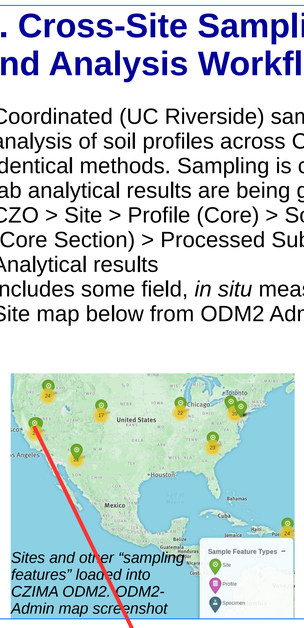
https://github.com/miguelcleon/ODM2-Admin

Web-based (Django) data management, access and visualization application developed by **Luquillo CZO** (part of the NSF Critical Zone Observatories (CZO) network) to manage their specimen and sensor time series data from the critical zone.



3. Cross-Site Sampling and Analysis Workflow

- Coordinated (UC Riverside) sampling and analysis of soil profiles across CZO's, using identical methods. Sampling is completed, lab analytical results are being generated
- CZO > Site > Profile (Core) > Soil Horizon (Core Section) > Processed Subsample > Analytical results
- Includes some field, *in situ* measurements
- Site map below from ODM2 Admin UI



Sites and other "sampling features" loaded into CZIMA ODM2, ODM2-Admin map screenshot

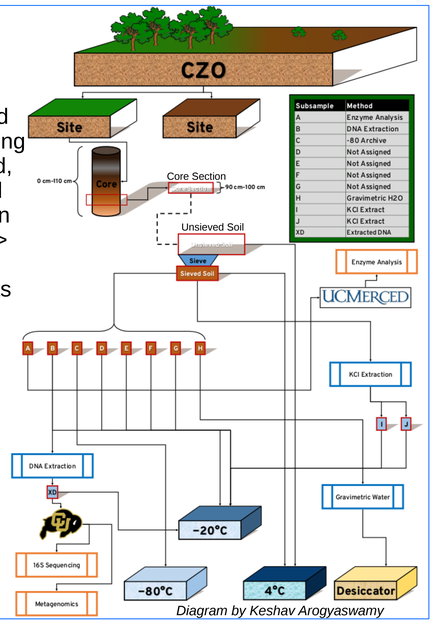
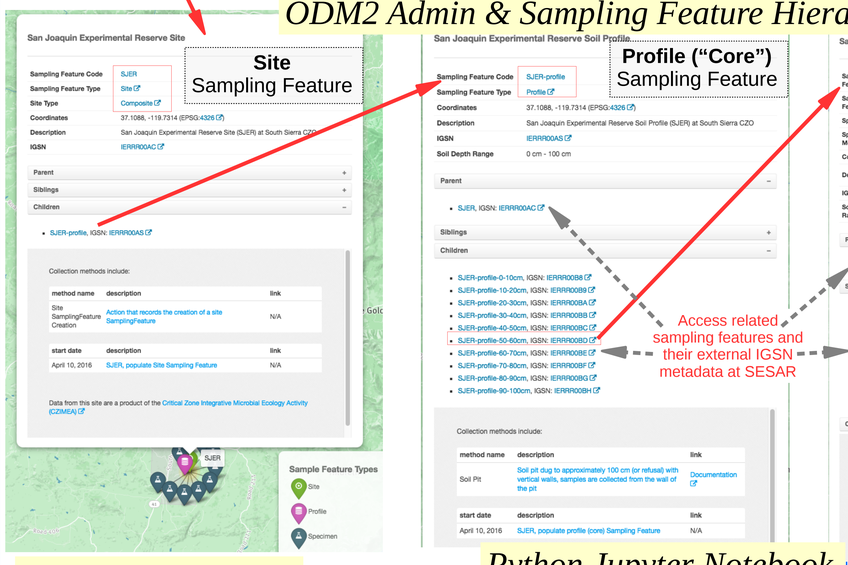


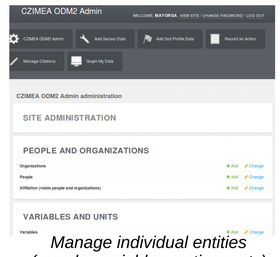
Diagram by Keshav Arogyaswamy

4. CZIMEA Data Management with ODM2 Ecosystem

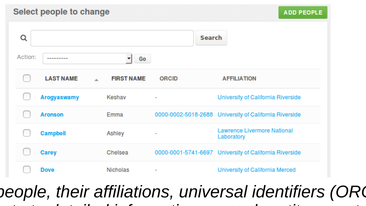
ODM2 Admin & Sampling Feature Hierarchy



Other ODM2 Admin Features



Manage individual entities (people, variables, actions, etc)



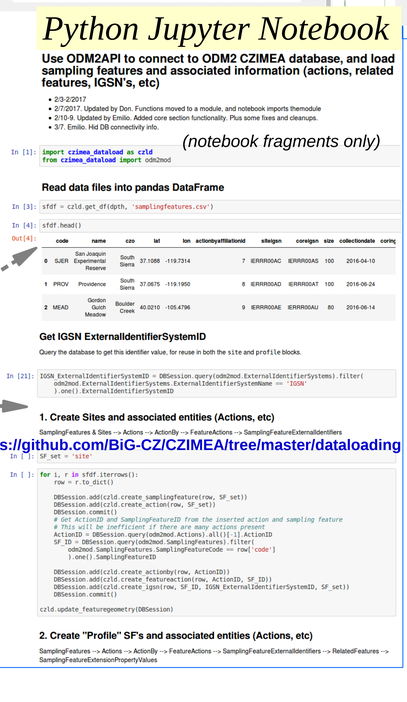
View people, their affiliations, universal identifiers (ORCID). Navigate to detailed information on each entity, or external resources (ORCID page, institution web site, etc)

Data Management

- **ODM2 PostgreSQL-PostGIS** database with ODM2 Admin web UI, hosted on Azure cloud instance (with Ubuntu) provided by the **CUAHSI Water Data Center (WDC)**, managed by Luquillo CZO Data Manager
- Metadata-rich database
- **Currently, only loaded metadata:** sites, sampling, samples & associated methods, people, timestamps, etc; including relationships between sampling features
- (Meta)data loading
  - Access via remote database connection
  - From various **spreadsheets**, using **odm2api** and **Jupyter notebook**
  - Directly via ODM2 Admin UI
  - Directly via SQL and RDBMS UI
- Specific needs from this use case have led to **substantial enhancements to ODM2 Admin**, collaborative development
- Storage, exposure and linkage of multiple **universal identifiers (DOI, ORCID, IGSN)**
- Latest CZIMEA-driven ODM2 Admin enhancements deployed June 2017; see **https://github.com/BiG-CZ/CZIMEA**

Python Jupyter Notebook

(notebook fragments only)



https://github.com/BiG-CZ/CZIMEA/tree/master/dataloading

5. Next Steps and Future Work

- **Finalize deployment of CZIMEA ODM2 "production" version and view-only access,** on CUAHSI WDC Azure cloud
  - Load remaining subsample metadata
  - Start loading results from in-situ and instrument analyses; and landscape properties from geospatial sources (soil taxonomy, land use, etc)
  - Develop Jupyter notebooks demonstrating access to metadata and data in database
  - Assess CZIMEA ODM2 Admin usability with project domain scientists
- **Genomic results**
  - Start designing linkage to external genomic results
  - Generate CZO Environmental Package metadata (adapted from Genomic Standards Consortium standard)
- **Web service access**
  - ODM2 REST API, **https://github.com/ODM2/ODM2RESTfulWebServices**
  - Make Sampling Feature metadata accessible via OGC WFS and WMS (with GeoServer)
  - Map soil environmental measurements to WoSIS schema and web services (OGC WFS & WMS); **http://www.isric.org/explore/wosis**
  - EarthCube GeoWS web services?
- **Potential linkage to EarthCube Cyberinfrastructure**
  - Linkage of individuals and projects to EarthCollab?
  - Explore semantic mappings via GeoLink / Earth System Bridge?
  - Discoverability via CINERGI / EarthCube Data Discovery Hub?
  - BCube service brokering?
- Continued **BiG-CZ/ODM2** engagement

6. References and Acknowledgements

ODM2 References

- Horsburgh, J. S., Aufdenkampe, A. K., Mayorga, E., Lehnert, K. A., Hsu, L., Song, L., Spackman Jones, A., Damiano, S. G., Tarboton, D. G., Valentine, D., Zaslavsky, I., Whitenack, T. (2016). Observations Data Model 2: A community information model for spatially discrete Earth observations, *Environmental Modelling & Software*, 79, 55-74, doi:10.1016/j.envsoft.2016.01.010
- Hsu, L., Mayorga, E., Horsburgh, J. S., Carter, M. R., Lehnert, K. A., Brantley, S. L. (2017), Enhancing Interoperability and Capabilities of Earth Science Data using the Observations Data Model 2 (ODM2), *Data Science Journal*, 16(4), 1-16, doi:10.5334/dsj-2017-004
- Horsburgh, J., A. K. Aufdenkampe, K. Lehnert, E. Mayorga, I. Zaslavsky (2017). ODM2: An Information Model and Software Ecosystem for Spatially-Discrete Earth Observations, HydroShare, *(Powerpoint presentation)* **http://www.hydroshare.org/resource/95458e53fe7e474f85642d6a711729b6**
- Horsburgh, J. (2017). ODM2 IPython Notebook Examples, HydroShare, **http://www.hydroshare.org/resource/ff79d7926f6040c9acd004636b4e4d38**

Acknowledgements

NSF EarthCube Awards 1541044 & 1541047. BiG-CZ Project (NSF Award 1339834). Luquillo CZO and CUAHSI WDC, for support with CZIMEA ODM2 Admin instance.