LiPD and CSciBox: A case study in why data standards are important for paleoscience

IN42A-10

Elizabeth Bradley[1], Izaak Weiss[1], Nicholas McKay[3], Julien Emile-Geay[4], Laura Rassbach de Vesine[1], Kenneth A. Anderson[1], James W. C. White[2], and Thomas M. Marchitto[2]

[1] Department of Computer Science, University of Colorado, Boulder, CO, USA
[2] Institute for Alpine and Arctic Research (INSTAAR), University of Colorado, Boulder, CO, USA
[3] School of Earth Science and Environmental Sustainability, Northern Arizona University, Flagstaff, AZ, USA
[4] Department of Earth Sciences, University of Southern California, Los Angeles, CA, USA

University of Colorado Boulder

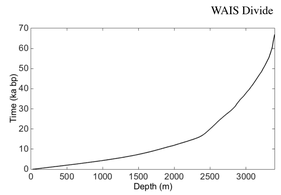
lizb@colorado.edu

Geoscience / CSciBox / Computer science

The main idea:

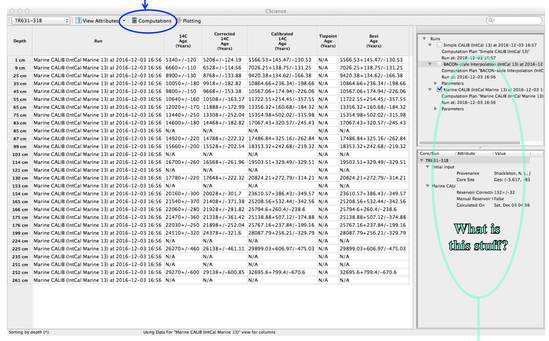
Photo: Dorthe Dahl-Jensen

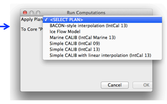
CSciBox



- A powerful analysis and design environment for building and working with age models
- Custom-built browsers and editors for viewing and processing core data and supporting data sets
- Modern cyber-infrastructure that allows fast, tailored access to data and scalable computation
- Built-in analysis workflows that incorporate commonly used tools for multiple types of cores
- A flexible, extensible framework that allows scientists to customize those workflows and tools

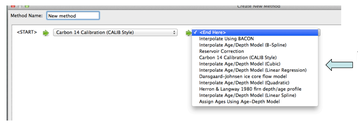
Running CSciBox





CSciBox ships with several useful analysis workflows….

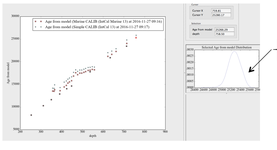
…and a workflow editor:



built-in tools

Custom Graphical User Interface (GUI)

- Sort by any column, ascending or descending
- `Run' shows which analysis workflow produced the corresponding result
- Can show/hide columns, customize attribute names, add new ones, …
- Powerful, flexible plotter; output in many formats
- Automatically applies age models to un-dated depths as they are created



Points on the plot are the mean of the distribution.
Clicking on a point shows the full distribution.

Can customize plot: font, grid, legend, error bars, icon shape/ size/color, …

"Metadata"

Reproducibility is a key component of the scientific method. As **computation becomes more central** to the scientific enterprise, it is urgent to address concerns regarding **reproducibility of computational results**.

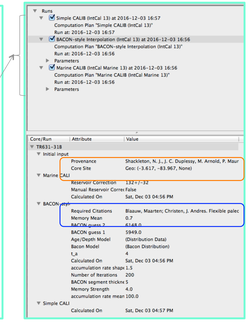
To this end, CSciBox **automatically tracks, stores, and displays** all the inputs and outputs of every computation you have performed on a core, the computation steps themselves, and any other information that you used in the analysis.

This means that:

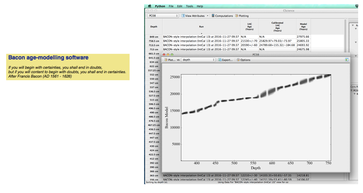
- Record keeping for reproducibility and publication is less burdensome (automatic!)
- Possible age models and other analyses can be viewed side-by-side with no additional work (both inputs and results)
- Variants of analyses are easy to perform:

- **Citation notes** are automatically collected and displayed every time you use another scientist's tool or data
- Metadata is always bundled with the core data
  - This includes data relevant to the core but not necessarily to age model computation (like provenance, location, laboratory procedures, etc.)

And storing all of this stuff is made possible by LiPD!



- CSciBox includes an array of commonly used modeling, calibration, and data analysis tools for building & using sediment & ice core age models
- The workflow editor lets you create new analysis programs from these built-in tools—without doing any coding
- This makes it easy to define, run, and evaluate variants of a given analysis
- People with modest programming skills can easily customize these tools
- CSciBox's "plug in" architecture is designed to make it easy for scientists to add new tools to the arsenal
  - Just implement your algorithm; data storage, user interaction, and background data (like calibration curves) are extensively supported
- "Plug ins" need not be written in Python; currently, for example, BACON is available as a built-in CSciBox workflow element:



- Existing programs can then be used *through CSciBox's GUI,* taking advantage of its features and providing a single consistent interface—and removing the need to learn the individual interface for each tool
- Workflows are stored *with the data, in a single consistent format,* so analyses are transparent, repeatable, and easy to update and share—between software tools and between scientists

CSciBox is open source

GPL Free Software

- Source code (python) available on github
- GNU public license; free to modify/extend/use as you see fit
- But you don't have to know python to run it; we have one-click installers too
- See our website for links to the code, installers, papers, instructions, and tutorial videos

We are trying to build— and support — a user community. Please join us!

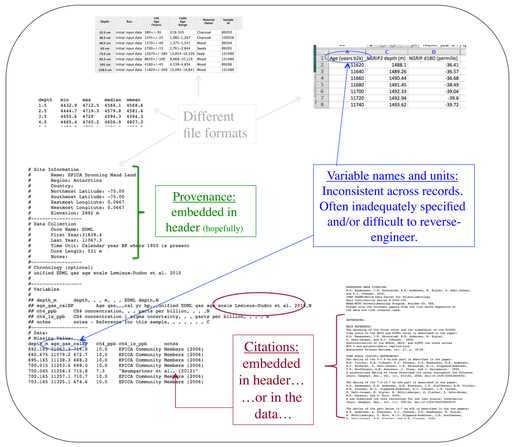
www.cs.colorado.edu/~lizb/cscience.html

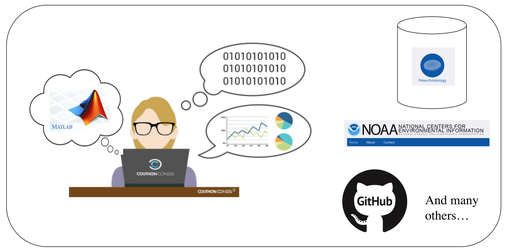
This material is based upon work sponsored by the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

IMPORT

Core issues for this poster

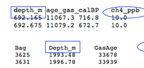
EXPORT



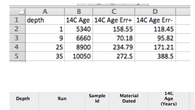


Paleodata Challenges:

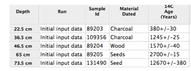
- File formats — xls, csv, txt, mat, …. $#&*%(#!
- Variable names, units
  - Different names for the same quantity
  - Different *units* for the same quantity
  - And units may not even be specified…



- Variable relationships
  - Implicit in the column names; no semantic link (e.g., that columns C and D are the +/- 1σ bounds on the error in column B)



- Provenance, procedure
  - Critical to reproducibility



- Citations
  - Generally stored separately from data (if at all)

- Interoperability, data sharing
  - Requires common formats

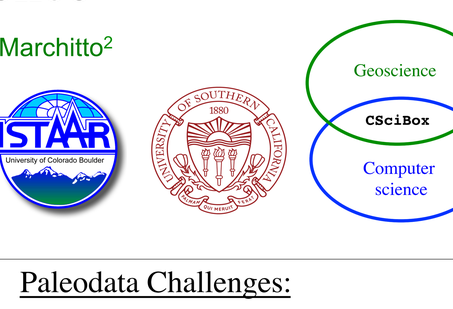
Solution to all of this: a data standard!

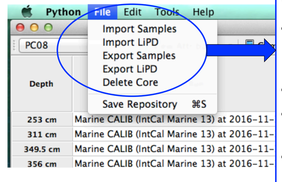
"Data standards are the rules by which data are described and recorded. In order to share, exchange, and understand data, we must standardize the format as well as the meaning." (usgs.gov)

Linked Paleo Data:

a container for paleoclimate data & metadata

- All paleorecords share common features, allowing for a common, meaningful structure
- LiPD is a flexible JSON container wherein paleodata travel with all the relevant metadata, linked semantically
- The container is structured hierarchically, so information is easy to access





{JSON}
JavaScript Object Notation: leading format for data sharing on the web

What this gives CSciBox:

- Uniform file format:
  structured, not flat: explicitly captures semantic relationships between variables
- Consistent variable names: coming soon
- Specific units with known meanings:
  CSciBox uses this to assure consistency
- Provenance is stored with the data:
  location, material, lab procedure, ...
- Citations are stored in appropriate parts of the LiPD record:
  linked to the data record, the method, etc.
- *Analysis steps* stored with the record, too…

CSciBox was the first third party adopter of LiPD, and that interaction directly resulted in…

- Capacity to store and restore probability distribution data
- Robust structure for method documentation to enable replication
- Seamless incorporation of multiple implementations or realization of chronological models
- Method and column-specific references to publications

Linked PaleoData

N. P. McKay and J. Emile-Geay, "Technical Note: The Linked Paleo Data framework—A common tongue for paleoclimatology," *Clim. Past Discuss.* **11**:4309-4327 (2015)

Conclusions:

- CSciBox + LiPD = single, easy-to-use "front end" to access and use the best community tools; support for using those tools in informed and appropriate ways
- All steps of age model construction & use can be performed within CSciBox, using a single intuitive graphical user interface
- Seamlessly interoperable with any software that uses LiPD
- Structured nature of the LiPD record captures data, provenance, results, analysis details, etc. ➡ complete documentation, "free" reproducibility