Researchers create novel platform to standardize paleoclimatology data

January 24 2020

Researchers from ISI and USC Dornsife create novel platform to standardize paleoclimatology data. Credit: Cassidy Joyes CC-BY-SA-4.0

Sometimes the most unrelated things can produce the most innovative results. Take, for instance, aikido—a Japanese martial art that can be

translated as the "way of unifying energy"—and paleoclimatology, a scientific field examining climate evolution.

Julien Emile-Geay, an associate professor in the Department of Earth Sciences at the USC Dornsife College of Letters, Arts and Sciences, got a direct taste of this in 2011 when staying with a friend for an aikido camp in San Francisco. His friend was developing semantic databases for biomedical data and Emile-Geay found that this approach could also work for the extremely idiosyncratic data collected by paleoclimatologists.

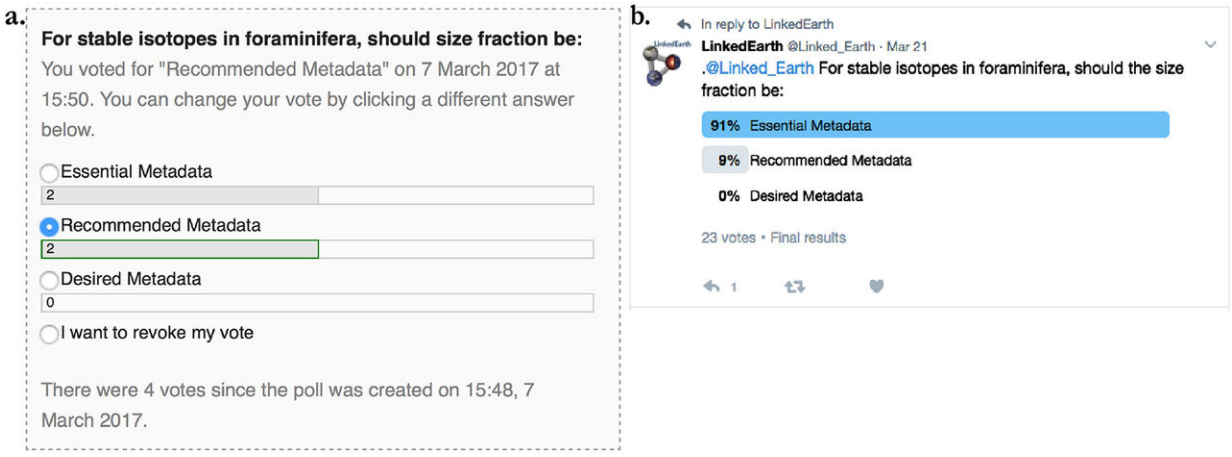
After a serendipitous 2012 meeting with Yolanda Gil, director of Knowledge Technologies at USC's Information Sciences Institute (ISI) and a research professor at USC Viterbi's Department of Computer Science, the researchers created a proposal to integrate Gil's AI expertise with Emile-Geay's Earth sciences background, developing a new platform that gives paleoclimatologists a way of unifying the disparate datasets of paleoclimate data, aikido style.

Along with Emile-Geay, the paleoclimatology group includes Deborah Khider, a postdoc at USC's Department of Earth Sciences and ISI data scientist, and Nicholas McKay, associate professor at the School of Earth Sciences and Environmental Sustainability at North Arizona University. On the AI side, Gil collaborated with Daniel Garijo and Varun Ratnakar, computer scientist and research programmer at ISI, respectively. The teams worked to create a new approach to standardize paleoclimatology data so that Earth scientists can better predict future climate to understand the causes and effects of climate change.

Their research was a feature article in the American Geophysicist's Union (AGU) *Paleoceanography and Paleoclimatology* journal and was highlighted at the AGU Centennial conference, held December 9-13 in San Francisco.

The Lone Wranglers

Paleoclimatology is the study of climate history, with researchers using imprints and indicators to reconstruct past climates. These indicators are usually physical samples collected from natural sources, such as glacier ice cores, tree rings, shells, cave deposits, and lake and ocean sediments. After integrating the resulting diverse datasets, researchers can reconstruct climate variables, like temperatures and rainfall levels. By recreating past climates, Earth scientists are able to predict future climates.



Example of polls on (a) the LinkedEarth platform and (b) Twitter (@Linked_Earth). Credit: *Paleoceanography and Paleoclimatology*

However, ironically, a major issue with the discipline lies in one of its strengths: the diversity of datasets. While the various datasets aid the creation of complicated model simulations to help researchers understand climate progression, the idiosyncrasies of each dataset can be difficult to integrate.

Earth scientists have their own approaches, processes, and data collection and coding methods that may not always be complementary or intuitive, and transforming the data into a usable format for research and analysis, or "data wrangling," can be a cumbersome task. Some researchers can spend up to 80% of their time wrangling data, such as identifying outliers and missing values or looking for dispersed records in multiple databases. The need for standardization in the field was clear. "Life without standards is miserable!" Emile-Geay said. "Imagine needing a different plug type for every single item in your house-that's currently the state of paleoclimate data, forcing early-career folks who want to integrate their data to spend months of their life reinventing the wheel every time they do something." Especially as funding is getting scarcer, Emile-Geay noted, this data wrangling is essentially a waste of time. "We were sick and tired of it and wanted to save future generations from wasting their Ph.D. brains that way."

A Socio-Technical Approach

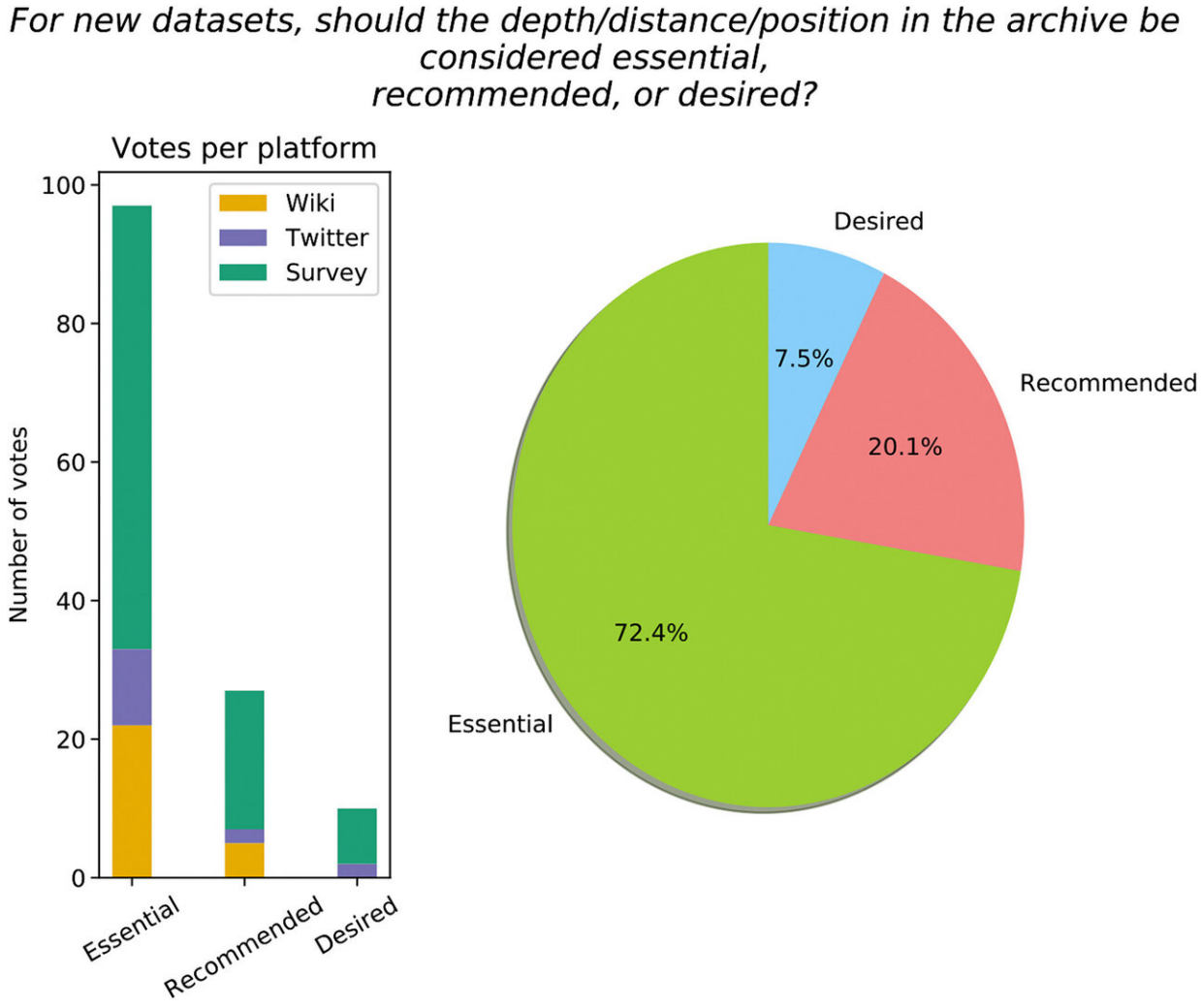
To address these concerns, the paleoclimatology and AI teams developed a novel platform. This new platform is part of the NSF's LinkedEarth project (funded by EarthCube), and is based on a "controlled crowdsourcing" approach, where the crowd (i.e., the paleoclimatology experts using the system) can develop terms, or properties, to code their data, which are then made available immediately to other users. By creating new properties, users can choose the appropriate terms to define the dataset they're working with.

The process is controlled in that a select group of users representing a wide array of paleoclimatology fields establish an editorial board, which reviews requests for new or changed properties and determines whether the users' proposals are to be incorporated into the Paleoclimate Community reporTing Standard, or PaCTS. All decisions made regarding PaCTS involve the input of paleoclimatology researchers,

making it a transparent, inclusive and bona fide community effort.

The system implements AI to help draw links between data and make them more accessible. "The AI techniques that we use are semantic technologies that allow us to represent scientific knowledge," explained Gil. "We also construct what we call the "Linked Earth knowledge graph' that expresses connections among datasets, researchers, locations, publications, etc." She noted that, additionally, users can ask "sophisticated queries of the ontologies and knowledge graph to easily access the data they are interested in."

The platform is described as a socio-technical system. Along with all the technical aspects, the approach has strong social aspects, as the value of the platform relies on information sharing. A key incentive for users is that they receive recognition for everything they contribute to the platform, which is tracked and displayed on their profile pages. Additionally, they can upload metadata specifications and existing datasets in multiple standards formats, making it easier to contribute to, access, and unify the data.



Example of a survey question for a new data set. The histogram represents the number of votes on each platform (orange: LinkedEarth, purple: Twitter, and green: Google survey). The pie chart represents the fraction of the votes for essential (green), recommended (pink), and desired (blue). Credit: *Paleoceanography and Paleoclimatology*

Setting the Standard

Developing the platform was no walk in the park. Khider explained, "One of the challenges was to come up with the framework for the

standard," which is made up of three elements: data representation, vocabulary and reporting requirements. "The second [challenge] was to get the community engaged," she continued. "We all want standards to advance the science, but no one really wants to talk about them." Another issue was figuring out where and how to start. As Khider noted, "In the end, we decided that the standard should reflect the needs of a specific community in order to do the most rigorous and exciting science."

There were also hurdles from an AI perspective. "The biggest challenge is that scientific knowledge is always evolving, so as scientists develop a better understanding of the data and their models, they may change how they want the data to be described and organized in the Linked Earth platform," Gil said. "[We needed] to accommodate the evolution of the ontologies and knowledge graph while not losing the work that users had done in the platform using previous versions of that knowledge."

But the hard work paid off. Not surprisingly, the platform has received positive feedback from the paleoclimate community. As of 2019, the controlled crowdsourcing wiki has 692 datasets, with 150 registered users and over 50 contributors. More than 14,000 pages have been created, as the paleoclimatology and AI teams continue their work to improve the platform and get more users involved.

The recognition from the AGU came after the project was implemented. "The editors at *Paleoceanography and Paleoclimatology* were instrumental in getting this project visibility within the community by selecting the manuscript for their Grand Challenges series," Khider remarked. "Having publishers pushing for standards is helping with community engagement for the second version of the standard, since they see interest in this type of work."

The platform can also be applied to other fields. "We are using [the

platform] now to describe neuroscience data in an NIH-funded project that we have with the ENIGMA collaboration," said Gil. "A novel aspect of this domain is that each dataset describes data for a cohort of people who are part of a study, and contains a collection of observations and not just a particular one."

Moreover, PaCTS is only one third of the standardization process, as it accounts for the reporting requirements. Standardizing data representation and terminology round out the process. The latter entails vocabulary and associated spelling, Khider noted, as most of the databases contain identical concepts spelled out in different ways, which can make querying for a particular dataset challenging. "The most obvious next step is to build a library of exemplar notebooks showing how these standards and code help solve common research problems in paleoclimatology, and how they open the door to new investigations," Emile-Geay said. "It's now time to make these standards work for [scientists]."

More information: D. Khider et al. PaCTS 1.0: A Crowdsourced Reporting Standard for Paleoclimate Data, *Paleoceanography and Paleoclimatology* (2019). DOI: 10.1029/2019PA003632

Julien Emile-Geay et al. Toward a semantic web of paleoclimatology, *Geochemistry, Geophysics, Geosystems* (2013). DOI: 10.1002/ggge.20067

Provided by University of Southern California