Patterns

CellPress
OPEN ACCESS

Opinion

Building a Traceable and Sustainable Historical Climate Database: Interdisciplinarity and DRAW

Victoria Slonosky[1,2,*] and Renée Sieber[3]
[1]Atmospheric Circulation Reconstruction Over the Earth - Canada, World Meteorological Organization, Montreal, QC, Canada
[2]Centre for Interdisciplinary Research on Montreal, McGill University, QC, Canada
[3]Department of Geography, School of Environment, School of Computer Science, McGill University, Montreal, QC, Canada
*Correspondence: victoria.slonosky@mail.mcgill.ca
https://doi.org/10.1016/j.patter.2020.100012

Turning historical meteorological observations into usable data is a challenging process that is immeasurably enriched when it encompasses interdisciplinarity. Here, the McGill DRAW (Data Rescue: Archives and Weather) project shows how climatologists, geographers, archivists, data scientists, and coders together built a citizen-science-based transcription platform to transform the McGill Observatory paper records into a traceable and sustainable database.

We're awash in historical weather data. Meteorological instruments like the thermometer were invented in the 1600s, and people have been measuring the weather ever since. We've now accumulated centuries' worth of climatological data, which are of utmost importance in understanding climate change, climatic variability, and extreme weather and climate events. Although they only (!) go back to the 1600s, these observations are neither proxy data nor model output. These data provide the only direct measurements of our actual climate.

Despite their value, there are difficulties to using historical weather records. First, they're not data. The original weather observations, in handwritten diaries and registers, are neither digital nor discoverable, so are not in a form we can presently use. Second, we have more content to convert into a machine-readable format than we have individuals to transcribe it. We often need to reach beyond the credentialled scientific community to digitally capture the records. Third, they're not standardized. Over the course of the past 4 centuries, the weather and climate have been measured with diverse instruments and been recorded in many, many different formats. This history requires a deep understanding of the assumptions and context in which the meteorological observations were made. Documents that contain elaborate cursive handwriting and specialist symbols are ill suited to machine learning. Fourth, we want to start thinking beyond the goal of using these records, once they're turned into data, solely for immediate scientific analysis. Data rescue projects are often of necessity focused on specific targets, but longer-term issues such as archiving for future users, data capture for non-climatologists, and data traceability are also important.

5 years ago, a group of people from different disciplines came together to think about the McGill Observatory records and how to transform them into a database of 4 million records. This is our story of how interdisciplinarity became essential to building a sustainable, traceable, and scalable data creation. Much of what we have been discovering in the first few years of DRAW (Data Rescue: Archives and Weather; https://citsci.geog.mcgill.ca/en/) is the immense amount of behind-the-scenes organization involved in turning source material into digital data.

The Story of DRAW

McGill University's Observatory (1864–1963) kept astronomical and geophysical as well as meteorological observations, and McGill's archival collections have weather records going back to the 1790s. The amount of information contained in the collections is enormous; the current project focuses on the daily weather registers with observations up to nine times a day starting in 1874 (Figure 1), a subset containing an estimated 4 million observations. We need to collect, organize, and handle all this content.

One approach could have been to type the observations from scanned images of the registers into spreadsheets. Our climatologist ran a previous data rescue project (ACRE-Canada) where she sent image files of weather diaries, custom-designed spreadsheets formatted to match the arrangement of observations for each weather diary, and instructions and FAQs to individual volunteer transcribers via DropBox; the volunteers usually emailed the completed spreadsheets back. This process is not uncommon for the often shoestring or volunteer efforts of historical climate data rescue and is currently recommended best practice by the Copernicus Climate Data Services Organization.[1] Some 500,000 data points from eastern Canada were transcribed, but keeping track of everything became untenable, and she couldn't continue as a one-person operation.

Thus, on a winter's day in 2015, at the McGill's Centre for Interdisciplinary Research on Montreal, the DRAW project began with a core group of researchers, each interested in different aspects of expanding the use of the weather registers. Our archivists are interested in the physical book: how it's stored and how to document and trace both the book and the information it contains through different media transformations. For them, DRAW offers a prototype in increasing accessibility of archival collections. For our climatologist and meteorologist, the main interest lies in the scientific content: how can we analyze the observations once they become data, what can they tell us about the weather and climate of the past, and how can that inform us as

Figure 1. An Open Register Book
The books on the shelf are the contents of one box; the boxes on the shelves behind contain yet more registers.

to the weather and climate of the future? Our information scientist (IS) is interested in data structuration and data life cycles. How do these affect data usability? For our geographer, her interest is in the process of engaging non-expert volunteers in the transcription of this weather information. Adopting a citizen-science approach means understanding what motivates individuals to volunteer their time and how to design a site that encourages citizens to keep coming back. Our educator is interested in how the DRAW transcription experience helps students learn about climate, research methods, and the scientific process. The various disciplinary lenses through which we all viewed the same physical object came together in the project to transform the McGill Observatory registers into a digital database with the help of citizen science.

How Interdisciplinarity Solves Problems Differently

We chose an interdisciplinary approach to transform the weather content into data. As we went forward, we found that the project was inherently interlocking: each part of the project implicated other aspects. One key example is in naming the image files. An image file was taken of the two pages (left-hand side and right-hand side, Figure 1) of an open register book. At first glance, how to name the image files might have seemed to be a decision for the archivists alone. It soon became clear, though, that the image file is used in every aspect of data rescue. In addition to her place-based interest, our geographer wanted the file name to serve as metadata because, in her experience (with the Federal Geographic Data Committee's efforts on metadata handling), metadata are easily separated from the data source. Our archivists wanted the file-name metadata to provide provenance for the original archival logbooks, an argument endorsed by our climatologist, who wanted traceability for meteorological data verification and contextual climate research purposes.

During the file-naming process, we discovered that the arrangement of meteorological content on a page was not uniform across time. This would not be such a problem if we were looking to only capture certain meteorological variables on the page such as temperature and pressure (as many data rescue projects do). Our climatologist wanted the entire contents of the page because many of the weather observations such as precipitation time, cloud cover, or weather conditions are compatible with older weather diaries and can be used to build up centuries-long, detailed pictures of climate change. With the complete record, it will be possible to look closely at the whole weather picture and analyze relationships between variables such as cloud type, temperature, precipitation, and weather condition.

We realized that the file name also needed to include these changes of information layout. Although our archivists didn't require this variation in typology—a link to the original physical book was all they needed—they were actively engaged with the needs of the archive users, in this case our climatologist and our app developer. Our app developer needed each file name to have the page layout it contained specified so the appropriate transcription settings could be linked to the layout of the observations on the page. The final image file name includes a register-type element that feeds into a data schema, which is then deployed in the transcription app.

The data schema provides an *a priori* categorization of the content to maintain the topology of the data as it's transcribed. We found at least 15 different register types, that is, 15 different arrangements of observation characteristics, variances in number of characteristics (from 38 to 49) collected, and frequency of sub-daily observations. We needed internal structures to maintain the relations between different items in a particular register (e.g., cloud type compared to precipitation) but also variations on an instrumental measure (initial and temperature-adjusted versions of the barometer). The schema also needed to incorporate transcriber and individual transcription information (e.g., is the page completed?).

The development of the data schema led to an unexpected scaling advantage. If an original, physical register was printed, then it was usually intended for mass distribution for the period during which it was in use. Once a pattern of information organization is "mapped out" and labeled, it can be used for any location using the same printed form. The idea of a register type as a metadata unit that contains information such as the variables recorded on a page (and thus contained within an image file) has been used not only for Canadian data but also for the Smithsonian Volunteer Network, whose meteorological observation forms were used across North America. The principle applies for any printed and widely distributed form.

We have yet to build a database-to-database crosswalk that will provide a single measure of, say, relative humidity for each sub-daily observation over the century of observations. Measurement standards, instruments, and techniques varied over the time period, and the observations are not all directly comparable. We thus need to maintain all these data points separately, rather than immediately reconciling them into a single data series.

The complexity of the data schema prompted us to build our own transcription app for DRAW (more details in Slonosky et al.[2] and Sieber and

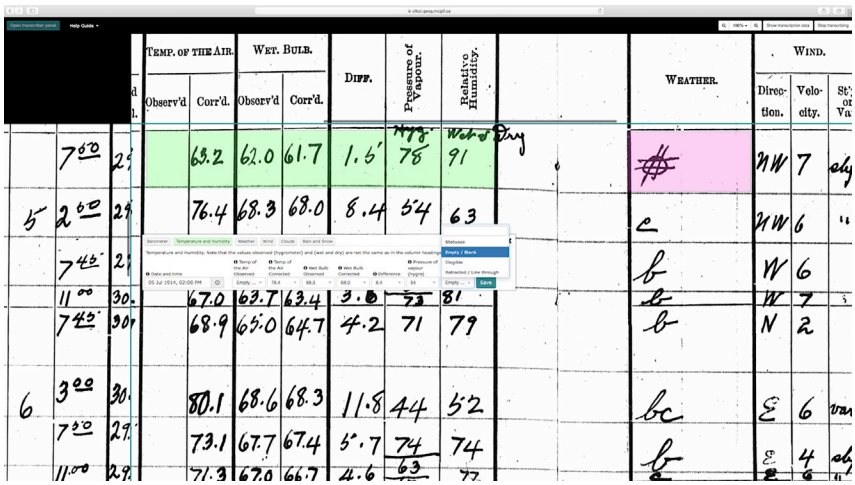

Figure 2. The Transcription Environment

Slonosky[3]). The app has a system-administrator interface and an end-user interface. The system-admin interface mediates file management and the configuration of each register type for data entry. Building this interface required us to develop knowledge about IT and learn about user interface/user experience (UI/UX).

The end-user interface encompasses both the transcription environment (Figure 2) and content to provide help, interest, and historical context, such as period photos and blog posts, as well as other outreach links. The team did extensive UI/UX testing, with surveys carried out by our archivists, IS, environmental sciences and studies (ESS) students, library staff, and general public volunteers. Our goal of capturing the entirety of observations on the page led us to some difficulties when we first encountered weather symbols (Figure 2). Drop-down select menus were suggested by our student, later volunteer, app developer as a solution. By having transcribers match printed weather symbols from 19th-century manuals with what they saw on the page, we could avoid having to direct them to legends or look up tables. Once drop-down menus were included as a transcription feature, we could also test whether limiting options in complex fields such as cloud type to a drop-down select menu helped avoid mistakes. Accuracy is an important issue in citizen science.[4] In evaluating the transcriptions, ESS students found DRAW transcribers to have preliminary accuracy rates of 96.14%.[5]

Interdisciplinarity, especially on shoestring efforts, required that we make significant investments to keep everyone engaged and support their interests. Some investments include allocating time to understand different vocabularies (e.g., wetbulb temperature, crosswalks) or deciding on acceptable terminologies when the same words meant different things to different disciplines (“digitization” or “classification”). Interdisciplinarity comes at a cost, mostly in terms of efficiency; the file-naming convention alone took several months of thought. It was well worth the time, as we now have a standard that travels across the entire data rescue life cycle.

Other investments included attempts to find funding for non-students or former students who wanted to remain part of the project after graduating or having to obtain permission from five different departments for permissions for some aspects of the project. Funding is a perpetual issue for a project that crosses discipline boundaries, not uncommon in citizen science or data rescue. Lack of funding also means that no one can work on the project for more than a small fraction of their time, and people often had to drop out, sometimes for considerable periods of time, to concentrate on other work. Not using proprietary and well-known software also has a cost. Our developer is committed to maintaining the software as open source on GitHub (https://github.com/open-data-rescue/climate-data-rescue), another platform of a contributory nature.

Reflections

Traceable. Transparent. Technically sustainable. Scalable. These are concepts that we keep returning to when we talk about our goals for the data. They’re words that are easy to invoke but hard to implement. We’ve been working on DRAW for 5 years, and we’re still not quite there yet. We know what we don’t want: for another data rescue team to return in 50 years and have to redo everything. We want future users to be able to build on our data for whatever the needs, interests, and curiosity of 2070 will bring. Interdisciplinarity exerts a cost, but with everyone contributing and everyone learning, we’ve found it to be a good way to create a trusted data source.

ACKNOWLEDGMENTS

Gordon Burr, Frederic Fabry, Rob Smith, Lori Podolsky, Eun Park, and Stéphan Gervais have been part of DRAW since the beginning. Drew Bush and Geoff Pearce have expanded our scope by bringing DRAW into the classroom,whereas Rachel Black and Yves Lapointe have enthusiastically supported outreach. Many thanks to all the volunteer transcribers, workshop participants, ENVR-401 student project teams, SIS, and work-study and summer work students who have contributed so much to DRAW. DRAW has been supported by Geothink SSHRC grant 895-2012-1023, the McGill Library Innovation fund and FQRNT grant Équipe 2019-PR-253338. As always, discussions with the ACRE (Atmospheric Circulation Reconstructions over the Earth) community and support for data rescue from Xiaolan Wang at ECCC helps keep historical data rescue moving forward.

WEB RESOURCES

Climate Data Rescue, https://github.com/open-data-rescue/climate-data-rescue.

Data Rescue: Archives and Weather, https://citsci.geog.mcgill.ca/en/.

REFERENCES

1. Wilkinson, C., Brönnimann, S., Jourdain, S., Roucaute, E., Crouthamel, R., IEDRO Team, Brohan, P., Valente, A., Brugnara, Y., Brunet, M., et al. (2019). Best Practice Guidelines for Climate Data Rescue (Copernicus Climate Change Service).

2. Slonosky, V., Sieber, R., Burr, G., Podolsky, L., Smith, R., Bartlett, M., Park, E., Cullen, J., and Fabry, F. (2019). From books to bytes: a new data rescue tool. Geosci. Data J. *6*, 58–73.

3. Sieber, R., and Slonosky, V. (2019). Developing a Flexible Platform for Crowdsourcing Historical Weather Records. Historical Methods: A Journal of Quantitative and Interdisciplinary History *52*, 164–177.

4. Eveleigh, A., Jennett, C., Blandford, A., Cox, A.L., and Brohan, P. (2014). Designing for dabblers and deterring drop-outs in citizen science. In CHI ’14: Proceedings of the SIGCHI

Conference on Human Factors in Computing Systems (Association for Computing Machinery), pp. 2985–2994.

5. Brinkerhoff, C., Albano, A., Feddersen, B., Becker, S., Kruglova, K., Tsynkevych, M., Hernandez, E., Nicoll-Griffith, K., Sieber, R., and Slonosky, V. (2017). Data quality of citizen science: learning from the past to inform the present. Project report for DRAW (Data rescue: Archives and weather) (McGill University).

About the Authors

Victoria Slonosky is a historical climatologist. She works on reconstructing past climates and analyzing climatic variability from historical records in Canada and Europe. She leads McGill's Data Rescue: Archives and Weather (DRAW) interdisciplinary citizen-science project and is an affiliated member of the Centre for Interdisciplinary Research on Montreal. She also leads the Canadian chapter of the Atmospheric Circulation Reconstruction over the Earth (ACRE) project; ACRE-Canada volunteers have transcribed over half a million Canadian historical weather records. Her book, *Climate in the Age of Empire: Weather Observers in Colonial Canada*, recounts Canada's scientific heritage in the field of climatology.

Renée Sieber is a professor of geography and environment (jointly appointed) at McGill University in Montréal, Canada. She is also affiliated with McGill's School of Computer Science, McGill's Digital Humanities Working Group, and the Global Environmental and Climate Change Centre of Quebec. Sieber works at the intersection of social theory and computer code. She is best known for her research on public participation GIS/participatory geographic information systems (GIS), the use of computerized mapping to facilitate citizen participation. She is currently researching the impact of automation (AI) on citizen participation in government and science.