From books to bytes: A new data rescue tool

Victoria Slonosky1 | Renée Sieber2,3 | Gordon Burr4 | Lori Podolsky5 | Robert Smith2 | Madeleine Bartlett5 | Eun Park4 | Jeremy Cullen3 | Frédéric Fabry3,6

1ACRE–Canada, Montréal, Québec, Canada
2Department of Geography, McGill University, Montréal, Québec, Canada
3School of Environmental Studies, McGill University, Montréal, Québec, Canada
4School of Information Studies, McGill University, Montréal, Québec, Canada
5McGill University Archives, McGill University, Montréal, Québec, Canada
6Department of Atmospheric and Oceanic Sciences, McGill University, Montréal, Québec, Canada

Correspondence
Victoria Slonosky, ACRE–Canada, Montréal, Québec, Canada.
Email: victoria.slonosky@mail.mcgill.ca

Present address
Robert Smith, Grenadine Technologies, Montréal, Québec, Canada.
Madeleine Bartlett, Museum of Jewish Montreal, McGill University Library, Montréal, Québec, Canada.
Jeremy Cullen, District 3 Innovation Center, Concordia University, Montréal, Québec, Canada.

Funding information
Geothink SSHERC grant # 895-2012-1023; McGill Library Innovation fund.

Abstract
Historical data provides observational information crucial to our understanding of the evolution of geophysical processes. However, there is a gap between predigital age observations, which are typically handwritten, and data that is discoverable and analysable. The data rescue protocols here address this gap, covering the information lifecycle from handwritten register pages to transcription-ready content, describing the historical data, the database design for the data rescue, and the development of an application design to transcribe the meteorological information directly from an image file to the database. The preparatory steps necessary to organize, curate, image, and structure the meteorological information, prior to transcribing the historical data, are outlined here in an integrated methodology. The initial organization, the development of an image file nomenclature to link the rescued data to the original source, and the description of a metadata schema to optimize the transcription application are all vital to the process of ensuring traceability and transparency in the data rescue process. Taken together, these steps describe best practices guidelines for similar projects. Although we designed the methodology and application to be used in any data rescue context, our particular concern was to accommodate the needs of citizen scientists. We thus focused on making our application easily maintained, flexible, direct to database, clear, and simple to use.

KEYWORDS
citizen science, data rescue, database, historical climatology, historical meteorology

1 INTRODUCTION

Data rescue has become a vital activity within climatology. Climatology has always been an historical science, relying on many years of cumulated observations, with the use and reuse of older data an inherent part of climatology. Examples of collecting and analysing older records date back to at least the 18th Century (Cotte, 1774). Efforts to compile and transmit old weather records continued apace through the 19th Century (e.g., Disturnell, 1867; Dove, 1840; Humboldt, 1817) and into the early 20th Century (Hann, 1903; Mascart, 1925).
Climatologists and others working in data rescue have peeled back the historical layers of weather data, moving from the monthly compilations printed in compendia such as the World Weather Records (Clayton, 1927) or Monthly Climatic Data for the World, which formed the original basis of datasets such as the Global Historical Climatological Network (Vose et al., 1992), to long series of daily weather (Moberg et al., 2000; Ansell et al., 2006) to sub-daily data (Brugnara et al., 2015) and finally down to the original, integral set of weather observations as they were recorded in the original register or logbook. Climatologists have always been interested in extreme weather, variability, and recurring patterns which may help long-term climatic prediction (Mascart, 1925). The rescue and reuse of historical weather data includes using high-resolution subdaily observations to better understand extreme, short-lived and high impact events such as storm and flooding (e.g., Ashcroft et al., 2016; Bosilovich et al., 2013; Dupigny-Giroux et al., 2007; Jourdain et al., 2015; Kaspar et al., 2015; Tan et al., 2005).

Several weather-related citizen science projects have been successful in the data rescue transcription of historical records, such as DataRescue@Home (Kaspar et al., 2015) and Old Weather (Brohan, 2016). Recent advances in computing power and memory have allowed more data to be stored digitally, leading to an explosive growth in weather data rescue projects in the last few decades (Allan et al., 2011; Brunet and Jones, 2011; Jourdain et al., 2015; Kaspar et al., 2015). Historical weather data rescue is a relatively young field compared to other citizen science areas (Silvertown, 2009) and literature documenting projects is still scanty outside technical presentations and recommendations (e.g., Ryan et al., 2018, Thorne et al., 2017; Kaspar et al., 2015, Bosilovich et al., 2013; Allan et al., 2011). We hope this contribution will add to the literature on citizen science-based weather data rescue.

A decade ago, Bronnimann et al. (2006) shared their experiences and provided useful recommendations for archival data rescue in the climatological domain. Their major recommendations included knowing the data source ahead of time; determining, where possible, if the quality of the data to be captured would be sufficient for later analysis; and considering the error rate of different methods of transferring the information from a physical object (usually paper) to a machine-readable format. In the past decade, we have seen advances in digital data storage and retrieval capacity along with increases in digital image technology. Thus, the digital image file of the original paper record has added a vital new link in the data rescue chain and provides a key new element in maintaining data traceability back to the original document so that the data retains context. Brunet and Jones (2011, 30) reaffirm the importance of this linkage to an entry in a register in a physical location because it allows others to assess the data’s accuracy.

Historical data rescue safeguards previous generations of observations not only for today but also for future generations. Even in settings such as governments or universities, institutional memory can fade and key documents can be lost. Documenting the data as it goes through its various transformations from handwritten observations on paper to scanned image file to machine readable data content helps ensure the utility and interoperability of data into the future. This traceability also allows a later “filling in the gaps” if only a portion of the original file (e.g., only pressure observations) has been transcribed into a usable file format. Finally, in designing data structures today, we need to keep in mind the fact that they will need to be flexible and adaptable to ensure long-term and sustainable use, as the needs and evolution of different research communities is often unforeseeable.

With these principles in mind, here we present a best practices guide to original historical data rescue, using the McGill project Data Rescue: Archives and Weather (DRAW) as a guiding example. DRAW is an interdisciplinary research effort to transform the paper records of the McGill Observatory in Montreal, Canada into a database format. The complexity of the McGill Observatory records is a feature that engages the interest of an interdisciplinary team in devising methods of classifying and transcribing these records (Park et al, 2018; Sieber and Slonosky, 2019). Although we focus on the process of the data rescue of one particular set of records, the meteorological elements observed at McGill followed international standards at the time (Kingston, 1878; Scott, 1908). While the lessons learned in DRAW are applicable to other weather data rescue projects, the process was designed to be adaptable to other archival handwritten records as well, including geophysical, medical, and other complex forms.

In this paper, we consider the steps that are required to organize data rescue projects before and during transcription of the registers. We design and implement an integrated data rescue protocol based on the organization of the image files of the meteorological documents. Key to DRAW’s organization is the various tiers of digital cataloguing we imposed on the information, for example, weather register types and meteorological variables. Equally important is the design of a flexible database and the transcription application. We describe in this paper the process of selecting, cataloguing, scanning, and organizing both the physical records, the digital representation of the records, and the meteorological data they contain (Section 2). We emphasize the importance of linking digital images to the original paper and ink registers for data integrity and transparency. Section 3 describes how the information contained within the weather registers is
structured into a database. We discuss the issues involved in designing a user-friendly open source transcription application based on this complex set of records (Section 4). These elements come together in an overall description of the DRAW project, from the original registers to the transcription application in our conclusions. All these steps are part of a vital, often undocumented, process for ensuring traceability, transparency, and long-term continuity of climatic data as it is transformed from paper to digital to ultimate use. This process is presented here as a best practice guide informed by practical experience gained from an interdisciplinary data rescue effort, combining expertise in archival practices, information studies, data management, public participation, historical climatology, and software design.

2 | HOW TO ORGANIZE THE HISTORICAL CLIMATE RECORDS

Here, we discuss the structure of the original records. The discoveries made in this detailed examination of the records inform our best practices recommendations for (a) metadata encoding in image filenames for traceability, (b) grouping identical data formats together into register types, and (c) developing flexible tools for data transcription.

2.1 | From observations to register books

From 1874 to 1964, professors and students at the McGill Observatory (Figure 1a,b) recorded the state of the weather several times a day, varying from up to nine times a day in the 1870s to twice a day in the 1930s (McGill, 1874). They wrote down over 30 different atmospheric elements (see Section 3) in tens of thousands of pages contained in hundreds of logbooks spanning nearly a century (Figure 1c,d). As the science of meteorology and social needs evolved, the elements recorded were modified to reflect these changes, leading to a variety of register formats (types). The McGill Observatory records are curated by the McGill University Archives, classed under the archival reference of *Accession number 1491*; the accession designates a particular fonds or collection within a given record group. In DRAW, we consider only the registers containing subdaily observations recorded at specific observation times (Figure 2). The complete 1491 archival accession also contains registers of wind profile charts and hourly observations transcribed from self-recording instruments, which do not form part of this project.

For the period from 1874 to 1935, the McGill Observatory register were A2-sized ledger-books, with the recorded observations for a given observing time spread over both sides of the sheet (Figure 2a,b). The recording observer opened the register book to the correct page and began entering observations for the appropriate date and time, filling in the columns from left to right across the sheet (Figure 2a), turning over to the next page, and continued entering observations for the given time and date on the next sheet (Figure 2b). The subsequent entries at the next observing time were entered below the first observation record. The registers were thus filled in from left to right across two double-spread ledger sheets (the equivalent of four sides for 72 elements, see below) for any given observation time, and from top to bottom at each subsequent observation time. Depending on the number of observations per day, there are from 2 to 7 days recorded per register page. These formats were standard across Canada and were typical of late 19th and early 20th century weather registers.

A considerable number of idiosyncrasies were uncovered in the format and organization of the records. As the organization of content changed over time, we created a “register type” to differentiate between the various ways the observations were recorded. We found five major register types in the period from 1874 to 1941, with subtypes within these major groups (Table 1). Reasons to identify new register types or add a new subclassification included the addition or removal of an observed meteorological element (Table 2), a change in the position of the recording of an element on the register page, or a change in the observing schedule. Subgroups identified with the 300 class of register, for example, were mainly due to variations in how humidity was calculated and recorded as experiments with this difficult parameter continually changed observing practices. A register type’s organization and structure provides one lynchpin in the transition from the physical paper and ink documents to the transcription user interface. Correctly identifying the type and subtypes of the register pages ensures that the information is correctly captured when transcribed into the database. These multiple changes in format mean that flexibility must be built into the data transcription process and thus into the database structure.

2.2 | From register books to unique digital image identifiers

The entire 1491 accession from the McGill Observatory was microfilmed with black and white photography in the early 1970s, photographing over 10,000 pages onto 34 microfilm reels. Digital images, in tiff and jpg format, were extracted from selected microfilms in 2003. The images were named 001.jpg, 002.jpg, and so on, within each microfilm reel directory, as a result of the automated imaging process (e.g., Tan et al., 2005; Westcott et al., 2011, WMO, 2016). Not only were these filenames uninformative regarding the information they contained, but they were also not unique.
Metadata is always necessary to describe the content and source of the data but tends to be onerous to collect and to not remain connected to the data it describes (Wilson, 2007). We decided to create a simple metadata standard that was embedded into the filenames (Murdy et al., 2015; Nelson et al., 2012). The metadata information was chosen to ensure a unique identifier for each image file. Elements that were most important and least redundant were chosen to be included in the filename as metadata. Another concern is to minimize the number of additional files or information keys that can get lost or disassociated from the bulk of the image files over time. Similar needs for file renaming to ensure clarity and traceability of weather data stored in image files were discussed by Brugnara (2017). We developed our standard to keep with our best practice goal of maintaining data transparency and traceability. Traceability is a key in maintaining the integrity of the information from the physical register in an archive container to each individual piece of weather data transcribed into the database. Traceability also is important for data quality control and future error checking.

The image nomenclature consists of five elements, grouped from the most unvarying to the most varying, or, seen from the catalogue point of view, from the largest grouping entity to the smallest. The accession number “1491” contains all the records of the McGill Observatory (i.e., many register types) so this is the first element. The register types (Table 2) usually span multiple years, and thus normally contain many individual registers, so is the second element of the filename. Each physical register book, which contains many pages, had been assigned an archival item number; this item number is the third element of the filename. Each page spans several days of observations, so the calendar date range spanned in given image file is the next element of the filename. Finally, each record spans two physical pages, so the last element of the file name is the Page Type (1 or 2). The metadata record of the information fields contained within each register type is the only piece of additional information necessary to identify the data contained on a given page from its filename (Figure 3). We coded open source tools in Ruby and Python to inspect the image files and rename each image with what became a unique identifier (see Supplementary Material Table S1).

### 2.3 From named image files to meteorological data elements

Figure 2 illustrates the data contents for register type 120 (cf., Table 2 for a listing of the individual elements). The
left-hand side of the first page of observations (Figure 2a) has the date and time of the observation. The first category of elements contains readings related to atmospheric pressure. The next columns record elements of temperature and humidity. On the right-hand side of a register sheet are columns related to wind, clouds, and precipitation. The second page of the register sheet (Figure 2b) starts again with the date and time on the left. Anemometer readings are the first column of the standard register page. At McGill, however, they provide a classic example of an idiosyncratic use of printed data forms which must be identified and accommodated in any data rescue schema. As wind velocity was recorded in a separate register book, observers sometimes used these columns to record other, ancillary information, in our case barometer readings. The casual use of printed forms and local custom understood by the observers in the 19th century can cause confusion in the 21st century, as the numbers entered are incongruent with the headings on the columns. Aberrant recordings of historical data (e.g., observations in a column that fail to match the heading;

---

| Date       | Temperature | Humidity | Wind | Clouds | Precipitation |
|------------|-------------|----------|------|--------|---------------|
| July 10    |             |          |      |        |               |
| July 11    |             |          |      |        |               |
| July 12    |             |          |      |        |               |
| July 13    |             |          |      |        |               |

---

**FIGURE 2** Examples of register pages from the McGill Observatory records (a) Page 1, July 10–July 13, 1879, (b) Page 2, July 10–July 13, 1879 (McGill, 1874)
unusual abbreviations, ditto marks, brackets covering several observation times) remain considerable obstacles to the transcription of historical data (Westcott et al., 2011).

The next group of variables on Page 2 consists of the minimum and maximum thermometers readings, followed by weather descriptors. The ways the observers recorded descriptive weather observations changed significantly over time. Initially, weather at the time of observation was listed with simple descriptors with qualifiers, such as “Fog,” or “Light Snow.” After June 1878, the international weather symbols (Figure 4) decided upon at the Meteorological Congress in Vienna were used instead. Letters representing different weather types were annotated by adding exponents to indicated intensity qualifiers, with light snow now represented as s0 and heavy rain as r2 (Kingston, 1878; Marriott, 1906; Moore, 2015). These present considerable transcription challenges (Section 4).

Not all the elements printed on a register sheet were actually recorded by the observers. When no precipitation occurred, for example, the precipitation entries were left blank. The implication for data rescue is that, even after the data are digitized, future data users may still need to return to the original records to ascertain their completeness or quality. We aim to capture the data as it is recorded, without interpolating or infilling missing data.

3 | CONVERTING THE REGISTERS TO A DATABASE

Given the size and scope of our data rescue effort (over 3 million potential observations), our platform design allows for multiple users and citizen science data transcribers working simultaneously. As described above (Section 2), the rescue and exchange of historical data is hampered by the lack of uniformity in many historical records. A well-designed database of transcribed observations should serve as a bridge between the individualist nature of each data source and a potentially more standardized version of encoded historical weather data for researchers.

A single hour’s observation record may require up to four sheets with 72 separate elements; we have 20,000 sheets of observations. The sheer number of observations

### TABLE 1 The McGill Observatory records, archival accession 1491: An overview

| Register type | Subtype | Date range | Archival item numbers | Number observations per day | Estimated expected number of observations |
|---------------|---------|------------|------------------------|----------------------------|------------------------------------------|
| 100           | 110     | August 1, 1874–June 2, 1878 | 1–8                    | Mon–Sat: 9; Sunday: 3      | 488,376                                  |
|               |         |            |                        |                            |                                          |
| 120           |         | June 3, 1878–Dec 31, 1879  | 8–11                   | Mon–Sat: 9; Sunday: 3      | 189,126                                  |
|               |         |            |                        |                            |                                          |
| 130           |         | 1880       | 12–13                  | Mon–Sat: 6; Sunday: 3      | 85,176                                   |
|               |         |            |                        |                            |                                          |
| 140           |         | Jan 1, 1881–Oct 31, 1884  | 14–21                  | Mon–Sat: 6; Sunday: 3      | 327,600                                  |
|               |         |            |                        |                            |                                          |
| 150           |         | Nov 1, 1884–Dec 31, 1899  | 21–51                  | Mon–Sat: 6; Sunday: 3      | 1,295,658                                |
| 200           | 210     | Jan 1, 1900–May 7, 1914  | 51–93                  | 3                         | 628,128                                  |
|               |         | August 27 1914–April 24 1915 |                        |                            |                                          |
|               |         | August 22, 1915–July 14, 1920 |                    |                            |                                          |
| 220           |         | May 8, 1914–July 4, 1914  | 80                     | 3                         | 4,320                                    |
|               |         | July 26, 1914–August 29, 1914 |                    |                            |                                          |
| 230           |         | July 5, 1914–July 25, 1914 | 81                     | 3                         | 1,008                                    |
| 240           |         | April 25, 1915–May 29, 1915 | 82                     | 3                         | 5,712                                    |
|               |         | May 30, 1915–August 24 1915 | 83                     |                            |                                          |
| 250           |         | July 15–Dec 31, 1920      | 94                     | 3                         | 8,112                                    |
| 260           |         | 1921–1934              | 95–108                 | 3                         | 265,872                                  |
| 300           |         | 1935                 | 109                    | 2                         | 26,280                                   |
| 400           |         | 1936–1937            | 110–111                | 2                         | 59,696                                   |
| 500           |         | 1938–August 3, 1941    | 112–114                | 3                         | 198,744                                  |
| Total number expected observations | | | | | 3,610,034 |
| Field Group (Tab) | Element                        | Register type |
|------------------|--------------------------------|---------------|
|                  |                                | 100 s | 110 | 120 | 130 | 140 | 150 | 200 s | 210 | 220 | 230 | 240 | 250 | 260 |
|                  |                                |       |     |     |     |     |     |       |     |     |     |     |     |     |
| Barometer        | Observed                       | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | Attached thermometer           | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | Corrected                      | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | Reduced to 32F                 | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | Reduced to sea level           | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
| Temperature of the air | Observed                  | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | Corrected                      | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
| Humidity         | Wet bulb observed              | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | Wet bulb corrected             | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | difference                     | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | Pressure of vapour             | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | Relative Humidity              | X     |     |     |     |     |     |       |     |     |     |     |     |     |
|                  | from wet and dry               | X     |     |     |     |     |     |       |     |     |     |     |     |     |
|                  | Relative humidity              | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | from hydrograph                | X     |     |     |     |     |     |       |     |     |     |     |     |     |
|                  | Dewpoint                       | X     | X   | X   | X   | X   |     |       |     |     |     |     |     |     |
|                  | Hygrometer                      |       |     |     |     |     |     |       |     |     |     |     |     |     |
| Wind             | Direction                      | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | Velocity                       | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | Force                          | X     |     |     |     |     |     |       |     |     |     |     |     |     |
|                  | S or G                         | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | St’y or Var                    | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
| Clouds           | Upper type and amount (2 letters) | X   |     |     |     |     |     |       |     |     |     |     |     |     |
|                  | Upper type and amount (1 letter) | X   | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | Upper direction                | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | Lower type and amount (2 letters) | X   |     |     |     |     |     |       |     |     |     |     |     |     |
|                  | Lower type and amount (1 letter) | X   | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | Lower direction                | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | Total cover/tenths             | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
| Rain             | Time began                     | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | Time ended                     | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | Duration                       | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | Amount/depth                   | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
| Snow             | Time began                     | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | Time ended                     | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | Duration                       | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
|                  | Amount/depth                   | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |
| Rain and melted snow | Amount                       | X     | X   | X   | X   | X   | X   | X     | X   | X   | X   | X   | X   | X   |

(Continues)
### Table 2 (Continued)

| Field Group (Tab) | Element                                | Register type |
|-------------------|----------------------------------------|---------------|
|                   |                                        | 100 s         | 200 s         |
|                   |                                        | 110 120 130 140 150 | 210 220 230 240 250 260 |
| Anemometer        | Reading of dial                        | X X X X X X | X X X X X X |
|                   | Miles since last observation           | X X X X X X | X X X X X X |
|                   | Miles in 24 hours                      | X X X X X X | X X X X X X |
| Max thermometer   | Observed                               | X X X X X X | X X X X X X |
| read at ___       | Corrected                              | X X X X X X | X X X X X X |
| Min thermometer   | Observed                               | X X X X X X | X X X X X X |
| read at ___       | Corrected                              | X X X X X X | X X X X X X |
| Radiation         | Grass min                              |               | X X |
|                   | Solar radiation/black bulb             |               | X X |
|                   | Total hours of bright sunshine         |               | X   |
| Thunderstorms     |                                        |               | X   |
| Phenomena at (time of) |                          | X  | X  | X  |
| observation (words)|                                        |               |     |
| Phenomena at time of observation (symbols)|                  | X  | X  | X  | X  | X  | X  |
| Phenomena and time of occurrence (symbols)|                  | X  | X  |     |
| Sleigh'g          |                                        |               | X   |
| Total snow on ground |                                  | X  | X  | X  | X  |     |
| Weather (words)   |                                        |               | X   |
| Weather (symbols and 1-letter abbr.) |            | X  | X  | X  | X  | X  | X  | X  |
| Weather since last observation (symbol) |                  | X  | X  |     |
| Weather and Phenomena since last observation (symbol and words) | | X  | X  | X  | X  |
| Ozone             |                                        | X  | X  | X  | X  |
| Aurora            | Class                                  | X  |     |
|                   | Time of occurrence                     | X  |     |
|                   | Class and time of occurrence           | X  | X  | X  | X  | X  | X  | X  |
| Remarks           |                                        | X  | X  | X  | X  | X  | X  | X  | X  |
| Initials of observers |                                | X  | X  | X  | X  | X  | X  |     |

**Figure 3** The nomenclature structure of the meteorological archive filenames

1491_120_11_1879-07-13-1879-07-15_1.jpg
FIGURE 4 Examples of weather symbols in 1888: (a) Canadian symbol for overcast sky (10/10ths cloud cover)*, Beaufort letter c; (b) International symbol for rain, Beaufort letter r; (c) International symbol for snow, Beaufort letter s; (d) International symbol for thunderstorm, (e) Canadian symbol for snowdrift (the international symbol has the vertical arrow pointing upwards)**; (f) International symbols for fog, Beaufort letter f; (g) Canadian symbol for clearing skies*; (h) International symbol for aurora; (i) handwritten example of exponents modifying Beaufort letter code for cloudy, signifying light cloud; (j) as in (i) for heavy rain; (k) as in (i) for light snow; (l) as in (i) for heavy thunderstorm. International weather symbols as taken from Scott (1908); Canadian symbols from Kingston (1878); handwritten examples from the McGill Observatory weather registers [McGill Observatory]. *Added to Canadian weather symbols by Kingston (1878), but “not included in those recommended by the Vienna Conference.” **Note that this symbol is rotated clockwise 90° in the Canadian registers compared to Scott (1908).

FIGURE 5 The historical weather data structure represented as an entity-relation model. Each box refers to a data table. The solid lines refer to a direct link between data tables. The dotted lines refer to an intermediate data table that serves the sole function of relating fields. Updated from Park et al (2018) as the flexible model evolved.
TABLE 3  Details and examples of the relational data structure (see also figure 5):

| Entity | Contains Attributes | Purpose is | Example of |
|--------|---------------------|------------|------------|
| User (transcriber) | User ID; User Name; Email; When signed up; Last logged on; Number transcriptions | Contains necessary information about the transcriber | 001; redacted; email redacted; since Jan 1, 2013; January 13, 2018, 14:23; 1000 |
| Register Type | Type ID; Associated Page Type(s) | Associate appropriate page types with a register | Type 120; Page 1_120; Page 2_120 |
| Page Type | Register Type ID for each page number; set of Field Group IDs | For each page type of a register (e.g., page 1, page 2) contains links to field groups on that page | Register Type 120, page 1; Barometer_Group_Type120; AirTemp_Humidity_Group_Type120; Wind_Group_Type120 Cloud_Group_Type120; Rain_Snow_Group_Type120 |
| Page | Page ID, URL to the image file, Image Height and Width, Link to Register Type, Metadata (extracted from image filename) | Links to and provides information for the original image file | 1491_120_11_1879-07-01_1879-07-03_1.jpg; 2234x2700; Register Type 120, Page 2; Tuesday, July 1, 1879, Wednesday, July 2, 1879 & Thursday, July 3, 1879; https://citsci.geog.mcgill.ca/en/admin/pages/337 |
| Page Day | Date ID for Page P 1; Date ID for Page 2, Links to number of observation per day | Method to link the pages of the same hour of the same day together Date of observation | Page 1 of July 01, 1879 to July 3, 1879; Page 2 of July 01, 1879 to July 3, 1879 Date and number observations per page 9 observations per day Monday-Saturday, three observations on Sunday, three days per page (cf., Figure 2; July 13th, 14th, 15th 1879) |
| Transcription | For Person_Page: Numerical & text entries, Bounding box coordinates | For any one user on any one page, tracks the activity, values digitized and bounding boxes of every annotation entered on a page (image file) | |
| Annotation | For a field group of a person page: Date, Time, Links to Numerical & text entries, Contains Bounding box Cartesian coordinates for one box (x,y; width, height) for a field group | Collects the data entered and Cartesian coordinates of bounding box surrounding the data on the image file | July 13, 1879; 7:48am; ("Empty";55.4;"Illegible";53.3,2.1 379,86); 100,100; 400,100 |
| Data Entry | Entry ID; number/string | Collects a single transcribed meteorological observation | 55.4 |
| Field Group | Name; set of Field IDs | Groups meteorological observations corresponding to the original physical layout of information on a page | Temperature Extremes [name for field group]; therm obs, max therm corrected, min therm obs, min therm corrected |
| Field | Name; Type (string, number) | Classifies all possible meteorological observation categories | Barometer Corrected for Instrumental Error |
| Field Option | ID; Name; …; Hint | Lists all potential values for a given field | Register type 100: “overcast”, register 120 Figure 4g |

on a given page at a given observation time can be overwhelming. As will be seen in Section 4 on the user interface (UI), an entire row of observations could not all be entered at once, for reasons related to both technical programming and ease of use for transcribers. To solve this transcription issue, a database entity called “field group” was created to group similar observations together in manner that reflects the organization of data on the page. This grouping is then reflected in the transcription environment described in Section 4. For example, the five observations under the printed heading “Barometer” (Observed, Att’d Ther., Corrected for instrumental error, Reduced to Temp.
32°, Reduced to Sea-level; see Figure 2a) are grouped together under the field group “Barometer”. As the data recorded changes between register types, new field groups are created whenever there is a change in the recording of the observations (Table 2).

The DRAW data rescue process moves from one that associates image or pages with a spreadsheet mirroring the layout to a database that is aligned with standard database practices. A relational data model is employed, where the data (e.g., “minimum thermometer corrected”) is structured into collections of types of data (e.g., “temperature extremes”). Each collection—an entity—is well-defined and thus separated from other collections. Figure 5 shows the relationships among the entities in what is called an entity-relationship model (Chen, 1976), and Table 3 shows the collection of entities in our database, with examples. The following are examples of “1 to n” relations shown in Figure 5: one page (cf., Figure 2a) has many annotations (individual data entries); one page type contains many pages; one register type has a Page 1 (Figure 2a) and Page 2 (Figure 2b) associated with it.

An example of a many to many (n:n) relationship is field options and fields. Direction, an instance of a field option, can be used for both wind and cloud. Clouds, an instance of a field, can have multiple field options like direction and cloud type.

The intent of this database design is to replicate the layout of a register type in terms of the number and schedule of observations per day and days per image file. This allows us to trace data back to the original register books and images as well as forward to the UI design for citizen science data entry.

4 | FROM DATA STRUCTURE TO DATA TRANSCRIPTION

The design of a platform that allows transcription of the data directly into a database is a vital component of the data rescue process. As the best practices guidelines described for data rescue can be applied to any consistent, handwritten set of records, here we describe the user and administration interface of the open source application designed for the DRAW project, in order that reader can adapt it to their purposes. It is important for data transcription applications to be both flexible, to accommodate changes in data structures, and as easy to use as it is possible to program.

Our response to the challenge posed by the complexity of the data contained in the weather registers, both for the number of elements and the variety of ways they were recorded, including numbers, text abbreviations, and symbols, was to develop our own custom-built software. The original code structure was inspired by the open source Zooniverse platform, used by many other citizen science applications, but we quickly found it needed to be redesigned to fit register-type documents, such as those found in the McGill Observatory archives. The DRAW software is open source and can be found at https://gitlab.com/openarchives-data-rescue/climate-data-rescue. Given the interdisciplinary nature of the project, we also designed for easy repurposability to a variety of data rescue goals (e.g., other weather registers, medical records, student records). We developed two interfaces, one for the project administrator and one for the transcriber, both described below (Figure 6).

4.1 | From Database to System Administrator View

One way to ensure flexibility is to design a database management system that links the data structure (i.e., the meteorological elements described in Section 3) to the user transcription interface (Section 4). We created an administrator interface that is easily repasurable for multiple configurations of fields, field groups, pages types, and register types. This flexibility to easily create new configurations in response to variations in historical sources is a key attribute we strove to implement as a best practice goal from the beginning of the project.

The administrator UI of the database management system (Figure 6a) allows the transcription “environmen,” which is the transcriber UI (Figure 6b), to be populated with the fields listed in Table 2 for each register page type. The project administrator builds up each register type in the environment by adding in possible field variables, fields, then field groups, pages types, and register types (Figure 7). Figure 7 also reflects the organization of the physical structure in the page, document (book), and catalogue of register types. Most fields are numerical, such as the recording of barometric pressure or temperature; others are free-entry text like “Remarks,” where transcribers can type in any value. Selectable options such as “empty” or “illegible” were added to every field.

As another best practice recommendation, we suggest preemptively providing drop-down select menus for fields which are symbolic in nature or where a limited vocabulary makes it feasible to give users a list of options to choose from. In the case of DRAW, drop-down selects were made for three fields: cardinal directions, cloud types, and weather conditions. Figure 6a shows an example of the administrator UI, where all possible variables are entered into the database for the selectable fields. The hope is that this specialized and restricted vocabulary will reduce potential transcription errors by showing transcribers only those values which we know a priori are possible. In the case of weather symbols for DRAW, we searched through the
registers and contemporary weather observation manuals (Kingston, 1878; Marriott, 1906; Scott, 1908) to identify all the possible abbreviations, symbols and combinations used for wind direction, cloud type, and weather descriptors.

The next issue was to determine how transcribers would enter what in relational data modelling are called “compound fields.” Compound fields are those in which multiple items have been recorded in a single column; cloud types are one such example. A typical entry might read “4CuSt,” which means the sky was 4/10ths covered with cumulostratus clouds. In this case, the transcriber would have to pick “4,” then “Cu,” then “St” from the drop-down menu to complete the entry with multiple drop-down selections made in the same entry box. The design process for the project is iterative, with new solutions to unexpected problems being deployed as new issues, such as new

FIGURE 6 User Interfaces of the DRAW app: (a) Project Administrator View for adding pulldown options for Wind; (b) Transcriber View for entering wind data
standardized vocabulary, symbols, or abbreviations are encountered.

4.2 From database to the transcriber view

When the transcriber opens a selected image file, information regarding the structure of data on the page is recognized by the app by parsing the image filename. Based on the metadata, the appropriate transcription environment with the variables to transcribe for each register type is loaded into the transcription environment. The page and register types act to configure the transcriber bar, which floats over the image (Figure 6b).

Field groups in the database (Section 3) translate into tabs on the transcription bar; each field group is assigned to a transcription bar tab. This reduces clutter in the transcription environment and enables the transcriber to focus on a small portion of the page. The work and organization that went into the file nomenclature described in Section 2.2 and the database structure described in Section 3 combine here to make the transcription task easier for the transcriber. The addition of the field group entity, which is not strictly necessary from a databases point of view, further demonstrates the iterative process developing the data rescue via a web app: a process that was introduced to address an issue for the transcriber interface design had effects leading back towards the database design. Citizen science-based design and usability issues are evaluated in a companion paper (Sieber and Slonosky, 2019).

5 CONCLUSIONS

The aim of this paper is twofold: first, to describe the best practices guidelines informed by an interdisciplinary historical data rescue effort, developed in cooperation with experts in archives, data and information studies, programming, geography, historical climatology, and meteorology; and second, to present an open source data transcription application and associated tools for data rescue. Each step described throughout this paper, from the cataloging of original records, to the database structure for meteorological elements contained in all registers, to the file nomenclature and finally to the transcription environment associated with each register type and page, are necessary components that work together to enable the rescue of this old weather data. Our best practices recommendations are as follows:

1. Data structure: Data rescue methodology comprises both data structure and data management; indeed, it is hard to see how these can be extricated. In the DRAW example, the layout of the information contained in each page (the register type; Section 2.1) needs to be minutely detailed to be accurately reflected in the transcription environment (Section 4). The software must be able to parse the unique metadata information contained in the filename (Section 2.2); so, it can assign the correct date and register information to the transcription environment. These elements are all reflected in the data structure (Section 3).

2. Traceability: Each archival record is by its nature unique (O'Toole, 1994); however, our filename nomenclature system is designed to adapt to any source. A geographical location identifier, an observer's name, or other relevant metadata can be added in or substituted for the archival fonds number used in the example given here. We recommend that each individual physical item be given a reference number to allow for traceability in the process. This traceability supports data quality, data confidence and future data repurposability. A crucial filename element for our app is the register type, which contains information concerning the meteorological variables and their placement on the page, followed by the date range of observations covered on the page. This process presumes the physical content has some initial cataloguing. If not, the content needs to be appraised, arranged, and described. The procedure described here also presumes the physical documents have been or will be digitally photographed as part of the data rescue effort, and gives consideration to the filename organization (metadata) of the digitized images.

3. Data capture: We recommend capturing the entirety of the records when practicable. Capturing the entirety of the meteorological information on each register sheet enables us to have the full complement of fields on which to conduct research, both now and for future researchers, whose potential needs cannot be predicted today. Furthermore, capturing all data enables error
checking and validation by ensuring different variables are mutually compatible (i.e., there is cloud cover when rain is recorded). Transcribing all the data meant that our platform needed to be custom-built with accommodations made for numerical, textual, and pictorial representations of data. We aim to maximize the transcriber’s effort and time by providing a list of known options when feasible, rather than have transcribers unfamiliar with raw weather records expend effort deciphering values for which only a limited range of options exist. This represents a departure from traditional spreadsheet data entry as for some fields the project administrator must preselect all possible instances of that field. This use of preselection needs to be balanced with free entry; free form entry gives the transcriber more freedom but also the potential to introduce error and waste effort.

4. Transcribe directly to a relational database: With the DRAW software described here, our goal was to create a database linked to the transcription environment (UI) in a way to be easy for the both transcriber to use and for the project administrator to manage. The transcription environment, which consists a transcription bar containing the field to be entered hovering directly over the image of a register page, allows the meteorological information to be transcribed directly to a database, without passing through intermediary steps such as spreadsheets or text files. This more direct process should help reduce errors. The flexibility of the project administrator UI, with simple addition of new fields and creation of new page formats, makes the application adaptable to a variety of historical data formats as it eliminates the need to create new templates when the historical format changes. Changes linked to a new format type, once entered into the appropriate field in the administrator UI, are automatically updated in the transcription bar, making the process invisible to the transcribing users.

5. Flexibility: The metadata structure, the database, and the transcription application must be sufficiently flexible to accommodate new configurations that may be found during the transcription process. With DRAW, although the collection of image files has been intensively examined, not every page of each of the tens of thousands of images has been scrutinized. We expect that, as the project proceeds, transcribers will discover as yet unidentified idiosyncrasies. The software must be able to easily and quickly adapt to these new discoveries.

Our decision to capture all the meteorological elements on the page (recommendation 3) allows us to build up a complete portrait of the atmospheric environment at the time of observation. Entering the data into a relational data structure (recommendation 4) also allows for relationships to be built among data elements. This can later allow for quality control and cross-checking between elements, or even for transcribers to check relationships as they type, such as seeing whether temperatures are near or below freezing when snow was observed. Ultimately, downstream users of the data will be able to investigate relationships between variables, such as how cloud cover and type relate to precipitation amount and type.

The commitment to faithfully record and transmit all the information as historically recorded means that further work will be necessary to provide data in a standardized encoded format for use by other researchers and to integrate into global datasets. Conforming individual datasets to international data standards is not trivial and requires further interdisciplinary conversations between historical climatologists and data scientists. The data from the DRAW project, for example, will be both hosted locally at McGill and integrated with other worldwide historical climate databases, such as the International Surface Pressure DataBank (ISPD; Cram et al., 2015) and the current Copernicus Climate Change Service data rescue initiative (Allan, 2017; Thorne et al., 2017). Collaboration is also ongoing for hosting at various image data repositories (e.g., MERIT: MEteorological Records Image porTal, https://met-acre.net/MERIT/, M. Bennoy, pers. comm April 26 2018).

Two principles which we strongly recommend as guides to any overall data rescue process are data traceability (recommendation 2) and software flexibility (recommendation 5). The need for traceability when transforming centuries-old observations from a physical paper and ink format to a 21st century format arises both from scientific and archival considerations. The scientific consideration is for data quality to be verified at any point either in the data rescue process or in later use of the data. Archival concerns imply we acknowledge we are one link in the chain of people interested in this data. We owe it to both our historical predecessors and our eventual successors to conserve and document records for future as well as current uses. The complexity and variety of historical records that have accumulated for over a century at the McGill Observatory testify to the many different instruments, methods, and purposes meteorological data has served through history. The McGill Observatory dataset, for example, has become well-documented as a result of the interaction between climatologists interested in data rescue and archivists. As the climatologists showed interest in the archival documents and indicated their scientific value, the multidisciplinary team including archivists, meteorologists, information, and data specialists were able to use this information to create cataloguing, metadata, and database structures. We expect that other historical climatic datasets, and other scientific datasets in general, are equally complex. Any data rescue efforts have to start with a survey of the physical
documents, and if necessary, the creation of an itemized catalogue. Capturing these shifts through time requires a flexible and easily adapted data rescue framework. As well as knowing one’s data, one also has to be able to organize, and if necessary, create traceability within the dataset in a way that acknowledges the needs of old and new users.

**OPEN PRACTICES**

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://citsci.geog.mcgill.ca. Learn more about the Open Practices badges from the Center for Open Science: https://osf.io/tvyxz/wiki.

**ACKNOWLEDGEMENTS**

The authors thank Sara Allen, Stéphan Gervais, Charles-Antoine Julien, Jesse Dineen, John Lindsey, Jeremy Cullen Lily Dwyer, Michelle Savard, Tiffany Hirsch, and the manuscript reviewers and editor.

**ORCID**

Victoria Slonosky [http://orcid.org/0000-0001-7631-387X](http://orcid.org/0000-0001-7631-387X)

**REFERENCES**

Allan, R.J. (2017) Current status of ACRE and C3S DRS. 10th ACRE Workshop, NIWA, NZ 7th December 2017.

Allan, R.J., Brohan, P., Compo, G.P., Stone, R., Luterbacher, J. and Brönniman, S. (2011) The International Atmospheric Circulation Reconstructions over the Earth (ACRE) initiative. * Bulletin of the American Meteorological Society*, 92, 1421–1425. https://doi.org/10.1175/2011BAMS3218.1

Ansell, T.J., Jones, P.D., Allan, R.J., Lister, D., Parker, D.E., Brunet, M., et al. (2006) Daily mean sea level pressure reconstructions for the European-North Atlantic Region for the period 1850–2003. *Journal of Climate*, 19, 2717–2742. https://doi.org/10.1175/JCLI3775.1

Ashcroft, L., Allan, R.J., Bridgman, H., Gergis, J., Padmenzky, C. and Thornton, K. (2016) Current climate data rescue activities in Australia. *Advances in Atmospheric Sciences*, 33, 1323–1324. https://doi.org/10.1007/s00376-016-6189-5

Bosilovich, M.G., Kennedy, J., Dee, D., Allan, R. and O’Neill, A. (2013) In: On the reprocessing and reanalysis of observations for climate. In: Asrar, G.R. and Hurrell, J.W. (Ed.) *Climate science for serving society: Research, modeling and prediction priorities*. Springer, Dordrecht, the Netherlands.

Brohan, P. (2016) Free at last. Available at: https://blog.oldweather.org/2016/06/29/free-at-last/. [Accessed 18 January 2018].

Brönniman, S., Annis, J., Dann, W., Ewan, T., Grant, A.N., Grieser, T., et al. (2006) A guide for digitising manuscript climate data.
Moore, P. (2015) *The Weather Experiment: The Pioneers who sought to see the future*. London, UK: Chatto Windus (Random House).

Murdy, J., Orford, J. and Bell, J. (2015) Maintaining legacy data: Saving Belfast Harbour (UK) tide-gauge data (1901–2010). *GeoResJ*, 6, 65–73.

Nelson, G., Paul, D., Riccardi, G. and Mast, A.M. (2012) Five cluster tasks that enable efficient and effective digitization of biological collections. *ZooKeys*, 209, 19–45.

O’Toole, J. (1994) On the Idea of Uniqueness. *The American Archivist*, 57, 632–658.

Park, E.G., Burr, G., Slonosky, V., Sieber, R. and Podolsky, L. (2018) DRAW (Data Rescue: Archives and Weather): preserving the complexity of historical climate data. *Journal of Documentation*, 74(4), 763–780.

Ryan, C., Duffy, C., Broderick, C., Thorne, P., Curley, M., Walsh, S., *et al.* (2018) Integrating data rescue into the classroom. *Bulletin of the American Meteorological Society*, 99(9), 1757–1764. https://doi.org/10.1175/BAMS-D-17-0147.1

Scott, R.H. (1908) *A new and revised edition of Dr. R.H. Scott’s Instructions in the use of meteorological instruments*. London, UK: Meteorological Office.

Sieber, R. and Slonosky, V. (2019). Developing a flexible platform for crowdsourcing historical weather records. Historical Methods.

Silvertown, J. (2009) A new dawn for citizen science. *Trends in ecology & evolution*, 24, 467–471.

Tan, L.S., Burton, S., Crouthamel, R., van Engelen, A., Hutchinson, R. and Nicodemus, L. *et al.* (2005) Guidelines on Climate Data Rescue. World Meteorological Organization Technical Report no.1210.

Thorne, P.W., Allan, R.J., Ashcroft, L., Brohan, P., Dunn, R.H., Menne, MJ *et al.* (2017) Toward an integrated set of surface meteorological observations for climate science and applications. *Bulletin of the American Meteorological Society*, 98, 2689–2702.

Vose, R.S., Scmoyer, R., Steurer, P., Peterson, T.C., Heim, R. and Karl, T.R *et al.* (1992). The Global Historical Climatology Network: Long-term Monthly Temperature, Precipitation, Sea Level Pressure, and Station Pressure Data (Technical Report No. ORNL/CDIAC-53; NDP-041). TN (United States): Oak Ridge National Laboratory.

Westcott, N., Andsager, K., Stoecker, L., Spinar, M., Smith, R. and Obricht, R. *et al.* (2011) Quality Control of 19th Century Weather Data. Illinois State Water Survey, Institute of Natural Resource Sustainability. University of Illinois at Urbana-Champaign Contract Report, 2011–02, p. 49.

Wilson, A.J. (2007) Toward Releasing the Metadata Bottleneck. *Library Resources & Technical Services*, 51, 16–28

WMO (World Meteorological Organization). (2016) Guidelines on Best Practices for Climate Data Rescue. WMO-No.1182, 9pp+ appendices.

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of the article.

---

**How to cite this article**: Slonosky V, Sieber R, Burr G, et al. From books to bytes: A new data rescue tool. *Geosci Data J.* 2019;6:58–73. [https://doi.org/10.1002/gdj3.62](https://doi.org/10.1002/gdj3.62)