Guidelines on Good Practices for Data Rescue

2024 edition

WEATHER CLIMATE WATER

WORLD
METEOROLOGICAL
ORGANIZATION

WMO-No. 1182

Guidelines on Good Practices for Data Rescue

2024 edition

WORLD
METEOROLOGICAL
ORGANIZATION

WMO-No. 1182

WMO-No. 1182

Chair, Publications Board
World Meteorological Organization (WMO)
7 bis, avenue de la Paix
P.O. Box 2300
CH-1211 Geneva 2, Switzerland

Tel.: +41 (0) 22 730 84 03
Email: publications@wmo.int

ISBN 978-92-63-11182-1

CONTENTS

ACKNOWLEDGEMENTS

We would like to thank the following persons for their outstanding contribution to this publication:

Mr Andrew M. Lorrey, National Institute of Water and Atmospheric Research (NIWA), New Zealand
Mr Axel Anderson, Deutscher Wetterdienst, Germany
Mr Denis Stuber, Météo-France, France
Ms Emelyne Roucaute, Météo-France, France
Ms Kate de Smeth, University College Dublin, Ireland
Ms Mariela Vásquez Guzmán, Climatic Data Rescue (CLIMDRE), Chile
Mr Matthey Fry, UK Centre of Ecology and Hydrology, United Kingdom of Great Britain and Northern Ireland
Ms Nancy Westcott, Midwestern Regional Climate Center, Illinois State Water Survey of University of Illinois (retired), United States of America
Mr Paul Poli, European Centre for Medium-Range Weather Forecasts (ECMWF), United Kingdom
Mr Philip D. Jones, University of East Anglia, United Kingdom
Mr Stefan Brönnimann, University of Bern, Switzerland
Mr Stefan Grab, University of the Witwatersrand, South Africa

Mr Gerard van der Schrier, Koninklijk Nederlands Meteorologisch Instituut (KNMI), Kingdom of the Netherlands; Mr Rick Crouthamel, International Environmental Data Rescue Organization (IEDRO), United States of America; Mr Rob Allan, UK Met Office (retired), UK; and Mr Washington Otieno, Ms Johanna Korhonen, and Mr Tommaso Abrate, WMO Secretariat, assisted in coordinating the review of this publication within the Copernicus Climate Change Service (C3S), IEDRO, Atmospheric Circulation Reconstructions over the Earth (ACRE), and WMO hydrology communities.

Many other colleagues contributed to the development of these guidelines. We would especially like to thank Ms Richaihu Wu, WMO Secretariat, for analysing all current WMO and C3S data rescue guidelines and for her significant input into the initial draft of this publication.

Mr Omar Baddour and Mr Peer Hechler of the WMO Secretariat oversaw the overall coordination of the publication.

SCOPE

This technical document is an update of the 2016 edition of the *Guidelines on Best Practices for Climate Data Rescue* (WMO-No. 1182). It incorporates the data rescue guidelines of the European Union Copernicus Climate Change Service (C3S) (Wilkinson et al., 2019; Brunet et al., 2020; Wilkinson and Freemann, 2021) as well as guidelines from the WMO publication *Guidelines for Hydrological Data Rescue* (WMO-No. 1146).

The present Guidelines are intended to provide guidance in the form of recommended good practices. An overview of data rescue is presented, followed by sections on its importance, archiving original media, imaging, digitization and archiving digital images and digital data. Fourteen appendices provide supporting information. Because of the diversity of national meteorological, marine and hydrological services with respect to size and stage of technological development, along with the variability of environmental conditions, some practices may not be useful for every WMO Member. That being said, the Guidelines cover a wide range of guidance that should provide assistance on how to organize and implement data rescue and provide generalized technological solutions for every Member. More specific technological information, as well as informative illustrations and photos, may be found at the C3S data rescue portal, which is maintained by C3S in close collaboration with WMO and other global data rescue players.

These good practices can be applied to the rescue of scientific data, in particular, the rescue of meteorological, hydrological, marine and other environmental data. The specificities of the environmental data need to be identified and taken into account in close collaboration with the respective communities.

GUIDELINES ON GOOD PRACTICES FOR DATA RESCUE

1. INTRODUCTION – THE SCIENTIFIC IMPORTANCE OF A DATA RESCUE

Data rescue involves accessing, organizing and preserving recorded instrument observations and environmental data at risk of being lost. Data rescue is critically important to ensure that future generations of scientists and other data users have access to all the information necessary to assess environmental variability and change, as well as to provide a range of environmental services. These data help bridge the gap between historical data and current observations. The length of available digital environmental data varies between countries and institutions. While many twentieth century records are available in digital format, there are cases where these data, alongside older records from the nineteenth and eighteenth centuries, exist only as hard copy records that need to be rescued and digitized. The validity of climate models and palaeoclimate data (tree ring data, ice core data, pollen observations and analyses, and so forth) would greatly benefit from long instrumental time series in all regions of the world.

Data-sparse regions remain in the twenty-first century. It is thus crucial to obtain and verify the more recent records. The reliability of environmental observations can be greatly influenced by location and site, as well as the quality control (QC) procedures relating to the observations. Comparing observational data from one site with those from nearby sites and consulting the original observation documentation (and original information on observational practices) will allow scientists to better verify past and present environmental regimes.

Extending the environmental record backwards in time and filling in temporal and spatial gaps has a number of benefits:

(a) It helps make agrometeorological, disease vectorization and hydrological and climatological numerical model representations more credible and enables better projections of future climate;

(b) Combining rescued data with data already available allows current weather and climate to be better placed within a historical context;

(c) It provides a basis to assess historical sensitivities of natural and man-made systems to environmental variability and change, thereby better enabling accurate assessments of the impacts of future environmental variability and change.

Such assessments can serve as input for policymakers to mitigate loss from environmental disasters and provide increased information for economic development.

In order to be accessible to environmental service providers and researchers in any practical sense, the data (including metadata) must be available in digital form (both images and numerical values). Furthermore, all current data also must be preserved and made available in digital form since these observations are the historical data of the future. Original documentation is crucial for the verification of digital values.

2. OVERVIEW OF DATA RESCUE: ARCHIVING, IMAGING AND DIGITIZING

Data rescue involves accessing, organizing and preserving environmental data at risk of being lost due to deterioration, destruction, neglect, technical obsolescence or simple dispersion of environmental data assets over time. Non-digitized data are at risk owing to the vulnerability of the original paper/media record. Data rescue includes: organizing and imaging paper, microfilm, microfiche and other outdated media records; keying numerical and textual data and digitizing stripchart data into an appropriate format; and archiving the data, metadata and quality-control outcomes and procedures. An overview of the components of data rescue activities is provided in Table 1. Common to each of these components are stages of planning, implementation and quality control to ensure successful and high-quality data rescue. Taking time at the beginning of a project/component to identify the necessary steps, data and equipment requirements and to develop the necessary quality control measures allows practitioners to identify potential issues and establish appropriate procedures.

Most countries have systematically taken observations of the environment, recorded their observations either manually or automatically, transcribed them onto paper forms or converted them into microfilm or other storage media, and eventually entered them into some form of computer media for easy access and analysis. Unfortunately, in some cases, these data were not transcribed from the paper forms, either because of a lack of funds and/or insufficient personnel or because the data were held elsewhere, sometimes overseas. In addition, many of the paper forms, microfilm/microfiche and other outdated media copies are at risk of being lost due to rapid deterioration of the medium or because of the undocumented relocation of portions of archives.

It is the purpose of these Guidelines to provide advice on the steps required to organize, image, digitize and preserve such environmental data. Appendices 1–3 provide additional information about data rescue as a global activity (Appendix 1), types of documents containing weather and climate, hydrological and marine data (Appendix 2) and a checklist for setting up new data rescue activities (Appendix 3).

3. ARCHIVING PAPER MEDIA[1]

3.1 Searching and locating

The first task of data rescue is to locate the data. Before beginning preservation procedures, it is of paramount importance that historical datasets be identified and recognized as meaningful environmental data. They may be conserved in national archives and services, at observing stations, or in private collections and other locations which are often unknown to the expert community. Some data may be held in overseas archives, particularly those of countries with a colonial history. Without undertaking such a search, many potentially valuable datasets may remain unidentified and inaccessible. Searching the following locations and sources may reveal valuable environmental data that have not already been organized, digitized and well managed in an existing data management system and which may therefore otherwise have been overlooked:

(a) NMHSs and other government departments or agencies, which often have data holdings;

(b) Oceanographic and polar institutes, marine fisheries, marine observatories, coast guard and marine environmental protection agencies;

(c) Astronomical, magnetic and geophysical observatories;

(d) National park services, telecommunication and postal services, airport and aviation services, commercial company archives;

(e) Hydrological institutes and agencies as well as electricity supply companies with hydropower projects and other companies associated with rivers and lakes;

(f) Government agencies, national, regional and local archives, diplomatic and consular services;

(g) Universities, colleges, high schools and individuals such as long-serving staff in meteorological and related organizations;

[1] Archiving digital media is addressed in section 6.

Table 1. Components of data rescue activities

Components of data rescue	*Activities*	*Keywords*
Archive paper, microfilm/ microfiche and other outdated media records (section 3)	Search and locate	National Meteorological and Hydrological Service (NMHS), observation sites, universities, aviation and maritime agencies, hydrological and agricultural organizations, international libraries and databases, national archives
	Preserve and store	Clean media; place in acid-free, labelled archive boxes, safe from dust, moisture, pests
	Create electronic paper/microfilm or outdated media records holdings inventories	Catalogue all paper media; estimate scope of imaging and digitizing effort
Image media (section 4)	Create master image inventory	
	Image and validate Update master image inventory Create image file inventories on each digital storage device/ directory	Update master image inventory after images are validated as readable, including metadata Cross-check external storage file inventories with master image inventory
Digitize data values (section 5)	Create digital data inventory	
	Key entry, chart trace	Input data into data management system
	Quality-control data	Update digital data inventory as data are digitized and pass through various QC tests
Archive digital media (section 6)	Cross-check printed media, images and digital data	Compare image and digital data inventories with original electronic paper/microfilm or outdated media record holdings inventories
	Back up electronic media	Daily
	Disperse multiple copies of images and digital data archives	Various locations
	Refresh media and migrate technologies	Every 5 to 10 years

(h) International data centres;

(i) Libraries or local historians, whose archives may have paper, microfilm/microfiche or even digital copies of data;

(j) Agricultural organizations, such as large corporate food companies, maintaining plantations throughout the world with meteorological records dating back hundreds of years;

(k) Military agencies and transportation ministries, which may have records that can be obtainable, especially aviation and maritime weather records;

(l) Religious organizations, particularly those with an observational or scientific interest, such as teaching orders, and individual clergy;

(m) Scientific societies, especially organizations focused on historical data (natural history societies);

(n) Museums;

(o) Ships' logs, historical newspapers and personal diaries, which can also be valuable sources of weather and climate information.

The search process should cross-check available digital records to determine whether the records have already been digitized, to identify significant gaps, and to provide clues to the dates on which observing programmes may have begun in particular regions. This includes checking international databases, such as:

(a) The Global Historical Climate Network;

(b) The International Surface Pressure Databank;

(c) The Integrated Surface Database;

(d) The Global Runoff Data Centre;

(e) The Global Groundwater Information System.

3.2 Preservation and storage

Paper and microfilm/microfiche holdings and holdings of other outdated media records should be organized in a logical manner and stored in acid-free archive boxes on sturdy shelves or in filing cabinets, and the following should be kept in mind:

(a) Depending on the previous location and condition of paper charts or forms, the paper copies may have to be dusted or vacuumed prior to storage;

(b) Correct work, health and safety procedures need to be followed in handling possibly contaminated records;

(c) Specialist equipment and training are required for dust removal from fragile records;

(d) Special care needs to be taken to ensure operator safety if records are mouldy or have been treated with mould inhibitors or pesticides;

(e) Fragile documents should be handled with clean and dry hands by holding the edges, as appropriate; clean nitrile or cotton gloves may be used to protect the operator's health;

(f) Depending on the amount of dust, protective clothing and masks may be necessary.

Special rooms are often designated for the storage of archived paper and microfilmed or otherwise stored data. In this situation, the following points should be taken into consideration:

(a) Ideally, the rooms should be temperature- and humidity-controlled;

(b) Data must be protected from insects, rodents, mildew, fire, flooding, dust, theft and all other dangers;

(c) Avoid the presence of wood in the room to minimize the dangers of fire and insect infestation;

(d) A professional archivist could help guide this process and a manager should be assigned to periodically inspect the room.

Data are usually stored in acid-free archive boxes or in filing cabinets by type (for example, form or chart type) and by station, year and month. When data are stored in this manner:

(a) The number of years of data and the number of stations may determine the most logical manner in which to organize the paper copies, whether by year or by station;

(b) Label the archive boxes and/or filing cabinets and add the storage location (the box or file drawer number), as well as the station/year/month, media type and form type to the electronic paper/microfilm/other media holdings inventory so that the data can be easily found for potential imaging and digitizing.

For long-term storage, the following should be considered:

(a) After critical data are imaged, the original documents must not be discarded. If resources do not permit continued storage, approach national archives, agency/service libraries, or university libraries for storage. If these institutions are unable to store the archives, and, if it is permitted by the national archive agency, international libraries or storage institutions should be approached (for a further discussion, see Appendix 5);

(b) Microfilm is not suitable for a permanent archive. Microfilm will deteriorate unless housed in special air-conditioned rooms with dehumidifiers;

(c) Digital images are also not a permanent archive (see section 6.2).

Appendix 4 contains a further discussion of the required infrastructure, supplies and personnel. Appendix 5 offers suggestions on maintaining paper archives.

3.3 Creating electronic inventories of paper/microfilm and other outdated media holdings and images

As paper media are organized and stored, a series of inventories is required to facilitate moving from paper copies to electronic images to digital data. This is necessary so that once the paper data are discovered and organized, they can be accessed for imaging. Later, once images are made, they can be digitized, and once the images are digitized, both the original paper data and the electronic images can be easily retrieved for quality-control verification of digitized values. Ideally, the naming protocol at each step should conserve as much information as possible. Table 2 summarizes the steps needed to develop an archive of original media (paper and outdated media) holdings, and to follow through with imaging, the accounting of images and the location of image files.

In summary, as environmental data are found, they should be incorporated into an electronic paper/old media holdings inventory to determine what data exist, what rescue efforts have already been made and what remains to be done. The electronic inventory holdings should include the station, year, month, media type, form number or type, element type, number of pages (including metadata pages) and, in particular, the box/file drawer location and the unique identifiers that are used to denote the documents. The variable type should also be included.

Once an electronic inventory for historical environmental records has been created, it should be maintained and updated as new data are received, and new image locations should be added to the master image inventory as imaging proceeds. The master image inventory can be used to estimate the progress of the imaging task. More details of the imaging task are discussed in section 4.

4. IMAGING ORIGINAL MEDIA

After the paper and microfilmed or otherwise stored copies are collected and catalogued, planning can begin to determine the order of imaging and digitization. Where to begin and how these tasks will proceed will be based on the needs of the individual NMHSs. Prioritization will likely be different for imaging than for digitization. See Appendix 6 for a discussion of items to consider when prioritizing the order of imaging and digitizing.

Paper and/or microfilmed or otherwise stored pages are photographed or scanned to create an image or digital picture which can then be preserved from further deterioration and made

Table 2. Steps to develop and maintain electronic inventories

Assemble all records and metadata for storage	
Organize the records according to a logical archival plan	By station/form type/year By form type/station/year By year/station If chart data, also by meteorological element (wind, precipitation, etc.)
Create an electronic inventory of paper/outdated media holdings	Including station, year and box/file drawer location, medium type, form type Include variable types Define missing periods Estimate quantity of medium types, form types and image volume to determine a workplan
Create master image inventory from the electronic holdings inventory	Identify station, year, month, media type, form type or number Define periods of record Estimate number of pages to be imaged
Create image file inventories as images are created on each digital storage device/directory for easy location of image files or in each computer directory	
Include pages containing only metadata (information on networks, observing practices, etc.)	
Maintain the master image inventory as imaging proceeds	Enter number of pages imaged as images are validated Enter digital storage device/directory location or computer location
Cross-check the master image inventory with the image file inventories for the number of images on digital storage devices/directories	

widely available. All pages of original meteorological documents (including blank pages and covers) should be imaged via a digital camera or optical scanner. This is to ensure that all available metadata (for example, observing practices, site and data descriptions) are rescued. See Appendix 7 for a discussion of the relative merits of digital cameras and scanners, desirable camera features, as well as the dos and don'ts of imaging.

In all cases, the metadata (the description of the data and the observing practices, location and site characteristics) must be rescued and digitized together with the images and digital values.

4.1 Basic steps in the imaging process

Imaging practices and techniques may vary, but all require documentation tracking and file tracking. Special care is needed to handle fragile documents to avoid further deterioration during the imaging process (including transport from the archive to the place of imaging and back); a professional archivist may be consulted in cases of doubt (this includes the challenge of imaging pages from books or diaries without damaging the binding). Some basic procedures that are strongly recommended are:

(a) As indicated in Table 2, create a master image inventory from the electronic paper/old media holding inventory. Include the number of pages to be imaged. Update the master image inventory when imaging and validating the images by adding the number of pages imaged and validated and by adding the computer directory or digital storage device/directory location (provide the template);

(b) Regularly validate the imaged files for completeness and readability as imaging proceeds to make sure the images are readable and the image content matches the filename, and regularly organize and back up imaged files according to an established protocol to avoid confusion among or loss of files (see also section 4.2 below);

(c) A large number of image files will be created by a digital camera or scanner. For each image that is created and downloaded to the computer, the camera/scanner software will generate a filename. These filenames may not be related to the file content. Some may be repeated if images are downloaded numerous times to a computer on the same day. It is recommended that software be used to generate a filename that is related to the image content (see the example in (d) below). Software is available with some cameras and scanners to perform this task or can be purchased separately. See the C3S data rescue portal and Appendix 7;

(d) The image filename should include the station identification number (SID), an acronym for the form type (ACR), the date (YYYYmmDD), and the page number (PPP). These files are typically .png, .jpg or .tif files (for example, SID_ACR_YYYYmmDD_PPP.png);

(e) Store images in an appropriate structure, such as station/year/month directories . Maintain a master list of filenames included in the main directory and similar file lists in individual directories or digital storage devices so that images can be easily located and retrieved;

(f) Update the master image inventory with the number of pages imaged, and compare the number of pages originally counted with the number of images actually produced and stored on digital storage devices/directories. The number of pages counted and entered into the inventories should equal the number of images produced. Enter the location of the images (digital storage device/cloud storage space or directory name);

(g) Documenting the imaging process, including the file naming convention and any particular issue encountered during the imaging process is crucial, as the personnel involved in imaging may change over time.

4.2 Validation and storage of image files

View and check imaged files before writing them onto a digital storage device/cloud storage space or into a climate data management system (CDMS) (see *Climate Data Management System Specifications* (WMO-No. 1131)), and keep the following points in mind:

(a) Images may be out of focus due to camera position;

(b) Images may be too pale and penned/pencilled corrections may not be visible;

(c) Images may be too dark because of camera/scanner settings;

(d) Portions of imaged data may be obliterated by a bright light. Using direct lighting with a camera flash or with camera stand can reflect the light back into the camera lens, obliterating some of the imaged data, especially when the medium is glossy paper.

An effective approach to quality assurance (QA) is to periodically (for example, every 300 pages) check a few images to make sure they are readable. Checking the completeness of the images and comparing them with the paper archive to make sure no pages of data were missed is also recommended. It is important to confirm that the image filename and the image contents match at least once per month/year/station.

If digital storage devices are used, they must be labelled.

(a) When used, the filenames of image documents must be entered into a file inventory, along with various identifying metadata (station name and number, year, month, form type, number of pages) and digital storage device label/cloud storage space location;

(b) A copy of the filenames associated with each digital storage device should be included on that digital storage device and in the master directory;

(c) The digital storage device label should be included in the master image inventory.

It is of paramount importance that duplicates (backups) be created of the image archive on digital storage devices/cloud storage spaces and/or in a CDMS/hydrometric data management system (HDMS). If digital storage devices/cloud storage directories are copied to CDMSs/HDMSs, each digital storage device/cloud storage directory should have a directory, and all file inventories must also be copied.

A fundamental requirement of good data management is to store backup copies of the data and metadata off site as insurance against loss in case of damage to buildings and computers from natural or man-made disasters.

(a) It is important to distribute copies of the entire electronic image archive and the master image inventory;

(b) If data are processed at a regional or multinational centre, copies of the data should be distributed to the originating countries and stored at the regional centre. When planning to process data at a regional centre, there must be in place a prior agreement between the countries involved stipulating what is to be done if natural or man-made disasters damage buildings or computers. The WMO Secretariat can assist with this type of agreement.

5. DIGITIZING DATA VALUES

Both the tracing of charts (including autographic stripcharts) and the transcription of textural data into a digital form that can be stored as numbers readable by computers are referred to as digitization. The former is addressed in Appendix 8; digitization of alphanumeric records by keying is addressed in the present section. Digitization is often performed at NMHSs. It is sometimes also performed at regional centres by companies or via citizen science initiatives and is usually guided by subject-matter experts, such as climatologists, hydrologists, and so forth. If printed data are located within a national archive, it is likely that digitization may need to be done there. When planning to process data outside the NMHS – or even outside the country – an agreement between the parties must be approved beforehand. Again, the WMO Secretariat can provide assistance with an agreement of this type.

Digitization usually takes much more time than imaging, and because digitizing is such a lengthy process, limited resources may prohibit all data from being digitized. The priority of data to be digitized should be based on scientific, technical and socioeconomic considerations. When it is not possible to capture all the information on a form digitally, it is especially critical to retain the original forms and copies of originals. See Appendix 6 for a further discussion on digitization priorities.

In some instances, when funds are not available to digitize large quantities of data, a crowdsourcing approach is used. Volunteers may key the data into an electronic form or trace time-series plots from charts on specially designed websites. Another approach is the use of optical character recognition (OCR) software to digitize printed material. This approach is undergoing testing and may become more effective and widely used in the future. At this time, OCR is of limited value as it requires specialized forms for readability. For a further discussion on crowdsourcing, charts and OCR, see Appendix 8 and the C3S data rescue portal.

A first step in digitizing data is to design a digital data inventory similar to the one created for imaging to follow the progress of the digitization process. Data should be organized by station, year, month and data type.

Regardless of the method to be employed (keying or tracing charts), procedures describing the steps to digitize the data and quality control the digital output should be developed using a subset of the records at the start of the project. By looking closely at this sample, practitioners can identify commonalities in their data (for example, faded or unclear entries, data gaps, historical annotations) and decide on an approach for dealing with problematic records and recording relevant information, as required. Providing good instructions to the digitizers also helps to avoid data quality problems, particularly if the digitizers are not experts. Instructions used historically for previous data collection and processing efforts should be collected and examined and language issues and domain-specific terms considered before writing instructions to present-day digitizers. The instructions should be as detailed as possible (see Appendix 9).

5.1 Minimizing keying errors

There are a number of techniques to minimize data entry (keying) errors. The first is to double- or triple-key the data (have two or more individuals key the same data). This can be done by having multiple keyers enter the exact same data and then comparing each data value. Alternately, one keyer or group of keyers can enter daily data while another enters monthly data; both sets of data can then be evaluated for consistency in monthly totals and average values. Double-keying can be done outside a CDMS and data imported after comparison.

A second technique to reduce keying errors is to create a template of the form being keyed. Use of a template allows the keyer to know exactly where to find the data on a page, reducing the number of errors. A template may also include sanity checks for the data entered into the form, further reducing transcription errors. Additionally, data keyed into a template can be directly imported into a database. Even when keying only a portion of a form, it is recommended that a keying template be created for each form type.

Thirdly, it is critical to transcribe the data exactly as they are printed (that is, to make a one-to-one transcription). This is known as “key as you see” in Australia and “saisir ce que l’on voit” in France. Avoid all forms of “on the fly” coding or unit conversion. Computers are very efficient at recoding, changing units, averaging, and even re-recoding. Keying as printed can ensure that there is a definitive record of the formulas used to recode the data. If coding must be used, take care that it is well documented and reversible (ensure that it is possible to return to the original data). Data must be keyed exactly as they appear on the original observation form, even if it is obvious that the observer made an error. These data are occasionally used in legal cases, and it is important to maintain the original data. Of course, erroneous values must also be flagged and possibly corrected in the further processing of the data. A copy of both the original and the error-corrected digital data should be maintained as two different datasets.

Double/triple keying, the creation of a template, and one-to-one transcription will greatly reduce keying errors. After the data have been keyed or imported into the CDMS as provisional data, a quality control should be performed (see Appendix 9). This cannot be overemphasized.

Finally, it is recommended that all data on a given page (for example, latitude, longitude, observer name, observation time, other meteorological variables and all numerical and textual data) be keyed. This is largely because it is unlikely that there will be another opportunity to return to that page and key any omitted data. Digitizing all variables also facilitates better quality control of the record.

5.2 Comparison with international databases

Acquiring digital climate data stored in international databases obtained from SYNOP or METAR codes may be attractive to an NMHS. Often, these data were not stored in-house at the time they were generated and may not be easily available due to the bandwidth required to download very large data files. Such data are desirable as they are already digitized and may contain more frequent observations and more variables than those stored in an in-house climate database. Although these datasets are likely to have been subject to quality-assurance procedures, they can still be plagued by keying, coding and transmission errors. However, if a value is not clear on the original paper form, it may be possible to recover it from the SYNOP code. Digitizing the original media and comparing it with data from international databases will lead to a more complete and accurate version of the climate data. Further data provenance is ensured by including data from international databases and newly digitized records and images, wherever possible.

5.3 **Metadata**

Metadata contain information and provide details about data, for example, how, where, when and by whom the data were recorded, gathered, transmitted and managed (Aguilar, et al., 2003).

5.3.1 ***Capturing meteorological metadata***

The minimum requirements for meteorological metadata comprise: information on station/platform identifiers, the geographical coordinates and elevation of the observing sites, and data-processing features (elemental units, special codes, calculations, algorithms, and so forth).

Good practices require more complete metadata such as:

- Stations/platform identifiers (for example, start/end date of the station, type of station, organization responsible for the data);
- Geographical data (for example, topographical information, location);
- Local environment (for example, local land use/land cover, obstacles, soil type);
- Instrumentation maintenance (for example, type of instrument, instrument sheltering and mounting, instrument calibration results and inspection results);
- Observing practices (for example, observing times and schedules, corrections);
- Data processing (for example, methods and algorithms);
- Historical events and data transmission (for example, formats, reporting period).

5.3.2 ***Specificities of metadata attached to ship observations***

Date and time of observations

Different countries use different date formats, and this will be reflected in the original documents. For instance, the 4th of July of 1921 might be written as 4/7/1921 in British documents and 7/4/1921 in those from the United States. In older documents, the month may be written out in full or a Roman numeral might be used, for example, 4 VII 1921. Dates must be keyed from the original document in a consistent and unambiguous format. A keying template must have separate cells for the year, month and day on the spreadsheet or on the web interface. The format (YYYY, MM, DD) in separate columns is consistent with the format used for terrestrial data and avoids ambiguities.

An issue with keying dates from ship logbooks or meteorological forms concerns the assignment of the date when a marine vessel crosses the International Date Line in either direction. When keying data, it is important to note that the date will change when the line is crossed. However, the date change may not have been implemented immediately by the logbook writer, and a day may be either missing or duplicated, depending on the direction of travel. This will need to be corrected by a quality assurance test.

Additionally, determining the day and time of day within logbooks as late as the nineteenth century can be bewildering if data rescue personnel are not familiar with the structure of the ship’s day or nautical day. Up until the nineteenth century, the ship’s day revolved entirely around noon, as this was the time when the ship's position would be determined and a revised course calculated. In effect, the ship's day ended at noon, and a new day began, meaning that the ship's day was 12 hours ahead of the civil day. If this is overlooked in the keying process, then sub-daily observations will be assigned to the wrong day and date. It is advisable to make adjustments to the keying format to better accommodate this structure.

In some much older logs and journals, the day may be divided into three parts (first, middle and third/last), denoting the times between noon to 8 p.m., 8 p.m. to 4 a.m., and 4 a.m. to noon. Assigning a time to these observations is necessary. The noon observation, which, from the point of view of the ship's officers, is the most important observation of the day, is clearly at the end of the third part of the day, and therefore the two other observations should be assigned to 8 p.m. and 4 a.m., which is the corresponding endpoint of the first and second parts of the day.

When keying meteorological observations from twentieth century deck logbooks, managers should consider how to deal with ships that adjust their clocks to the time zone they are passing through. This complicated issue is dealt with at the data-processing stage; nevertheless, keyers will need to know how to enter information where a time zone is indicated so that the data processing can be undertaken efficiently.

Where adjustments to ship clocks are noted in the logbook or where a ship records the time zone kept, a provision must be made in the keying template to record this information. Keyers should record the time zone as stated in the logbook (for example, GMT, +3, -8, and so forth), regardless of the actual position of the vessel, and should ensure that each time zone other than GMT has +/- to avoid ambiguity. Keyers must also record (in the same time zone column or set of cells) when the ship clocks are advanced or retarded, and by how much. With this information, all necessary adjustments can then be made during the data-processing stage.

Position of observations

The second essential parameter to key is the ship's position. This is the most complex and potentially problematic parameter that needs to be keyed. At sea, a ship's position is expressed as a latitude and a longitude, in degrees and minutes, and either prefixed with or followed by a letter denoting the hemisphere. Note that at times, the latitude and longitude were also estimated. In the past, it was standard practice to record estimates of daily noon position, called 'dead reckoning', when it was not possible to determine a more exact position, and observed positions were included when a noon sighting of the Sun was possible. Where there is no observed position recorded, the estimated position should be keyed. Where both are recorded, both should be keyed. In other cases, the ship's position between daily noon positions can be estimated from the steered course and distance recorded in the logbooks.

Latitude is always north or south of the equator and denoted by N or S either before or after the position. Longitude was not expressed in such a simple way until the late nineteenth century and early twentieth century. Today, longitude is expressed as being either east or west of the Greenwich meridian, and scientifically by using a notation whereby west longitude is a negative number and east longitude is a positive number. In historical records, longitude is always expressed as being either east or west of a meridian, but that meridian can be assigned to any point on the Earth. Before the 1850s, and in some cases much later, it should not be assumed that the longitude recorded is based on the Greenwich meridian. A provision should be made for keyers to record which meridian is being used. When keying data, the keyer must be alerted to changes in meridian, and a provision in the spreadsheet should be made for noting the name of the meridian being used. The data to be keyed should therefore include the name of the meridian (or the name of the landmark if a landmark was used as a reference), not its actual longitude. Sometimes, there is no position recorded in the logbook; this is always the case when a ship is in port. In sight of land, and usually in coastal waters that the ship's captain was familiar with, the latitude and longitude were also not recorded in the logbook, simply because there was no need to, which is again a reminder that ships' logs and journals were kept for practical purposes and not for the recording of scientific data.

5.3.3 *Capturing hydrological metadata*

When deciding which metadata to capture, consider who would want to access the data, both at present and in the future, and for what purpose (for example, technicians might want to find sites where certain instruments are deployed, hydrologists might want to find stations with good-quality flood records, and so forth).

Structure data hierarchically, using concepts such as river, catchment, station and time series (measurement), ensuring that all data values can be traced back to the stations and locations where they were observed.

Ensure that good spatial information, for example station locations, is maintained, including vertical heights relative to a datum and details on what this datum is. Catchment areas are often essential for the derivation of standard results such as runoff.

Reduce text descriptions to a series of simple, common elements where possible. For example, instead of storing text descriptions of the way in which data from each station are measured and processed, an alternative could be to list the common elements for all the stations – instruments, measurement frequency, parameters being measured, digitization method, post-processing method, and so forth. These could be described in a series of fields, which would allow the information to be searched and used more readily. However, it is always better to maintain a text description if useful information would be lost without one.

Develop and use controlled lists where possible. For example, make a list of instruments used or gauging station structures. Maintain any useful metadata about the list items or a more detailed description.

Capture the parameters being measured in detail (for example, river flow, water temperature, air temperature), the exact units of measurement, the measurement statistics (instantaneous, mean, total, maximum and so forth), the start date of the record and the end date for completed records.

Store information about the quality of the measurements, where possible, and information about any processes applied to produce a result, for example, flow data derived by rating from stage data manually measured with a stage board.

Capture information about the digitization or data rescue process itself. At some point in the future, it may be important to find out whether a time series was manually keyed in or digitized from a chart, and whether the digitization was undertaken directly from the chart using a tablet or from a scanned image.

When using hydrological data, it is particularly important that information about the quality of readings across the hydrological regime be captured. This could be as simple as categorizing stations, or parts of a record, as good or poor for high or low flows, but it could be more detailed if the information is available.

Link to other references where possible. For example, when describing an instrument used to measure data, a simple link to a scanned copy of the instrument manual could be included.

As the metadata are likely to be stored separately from the data, for example, in the case of a time series of river flows, the metadata should indicate where the data are stored. If the data have not been digitized, the physical location of the paper records should be described.

The following list contains essential basic metadata:

Latitude and longitude of the measurement location, and geographic datum for those values;

Name of the organization that collected the data;

Start and end time of the hydrological record;

Parameters and units being measured;

Method used to measure the parameter (for example, flow calculated from water level through rating) and its accuracy where possible (for example, water level measured from staff gauge to nearest 0.5 cm);

Time interval of the data and the applicable time zone;

Information on what the data value for each time interval represents (for example, mean over the preceding interval, total over the subsequent interval or instantaneous value);

How the record was digitized;

Where the record is stored, when it is not possible to include such metadata with the record.

5.4 Digitizing rating and gauging information

Often, river flows will have been measured at open river sections using the velocity–area method, where gaugings are made on a regular basis, measuring the actual flow and recording it against the water level. These gaugings may be used to create a rating for the station. It will often be the case that ratings have already been made in this way and that the records of these ratings and the gauging from which they are derived will also have been stored. In such instances, it is important that the full gauging history be digitized, together with any comments from the field technicians and hydrometrics officers who made and reviewed the measurements. It is also important that the rating curves and equations be digitized and that this information be kept alongside the digitized chart data.

Gaugings will often consist of manually written records of each flow measurement. These may simply be the date and time of the measurement, the measured flow and the water level, but they will often also include the velocity and cross-sectional area information used to calculate the total flow, as well as comments regarding the conditions of the section at the time of the measurement. Water level and flow information are the most important data for the production and analysis of ratings and should be sufficient for this purpose. The more detailed information should remain accessible as it can be useful should any individual gauging need further assessment. For some gaugings, usually those that capture very high or low flows, the detailed information could be very useful in assessing the quality of the flow measurement and whether it should be included in a rating.

If both gaugings and ratings are available, it may be worthwhile to reassess the quality of the rating. If only ratings are available, they may have to be taken on trust, but any available local expertise should be used when deciding whether this is appropriate. Flow series derived from ratings should be inspected carefully to ensure that no obvious shifts have taken place due to changes in the rating.

6. ARCHIVING DIGITAL MEDIA (IMAGES AND DATA VALUES)

6.1 Cross-checking printed media, images and digital datasets

While the work is underway, it is useful to record the progress of the project. This can be done by cross-checking the number of months of images and digital data created with the number expected based on the number of pages imaged and to be imaged and with the period of record for each station. At the beginning of the data rescue process, it is advisable to create an electronic holdings inventory of the original paper/microfilm/other old media data. Then, at the end of the process, the quantity of digital data created (found in the image and digitization inventories) can be compared to the initial electronic holdings inventory to determine if any data were missed.

6.2 Refreshing the media

Data rescue is just as important for digital data as it is for paper data since the media on which both types of data are stored are not permanent. For example, magnetic tapes lose their magnetism over time, particularly in warm or humid conditions. Paper starts to crumble after centuries, but magnetic tape media and other computer-readable media become unreadable in a matter of decades. Magnetic tapes, optical discs and 8-mm cassette tapes are generally obsolete, and devices to read these media may not be readily available. If digital data are still stored on these media, the national archives or universities where they are held should be contacted and encouraged to move these data to modern long-term digital storage solutions. Equally important is the preservation of format descriptions and of software to read/decode the data. Otherwise, the data might be an unintelligible string of numbers and characters.

As computer technology evolves, computers, computer operating systems, computer languages, and the software used to read the old media also become obsolete. The problem of obsolescence will continue. It is highly recommended that, in the future, both the imaged and digital climate data be migrated to the latest established media (including cloud storage, hard drives, and so forth). CDs and DVDs were widely used until recently, but they may no longer be used in the near future. It is also recommended to keep multiple backups of the data since computers (as well as digital storage devices) fail catastrophically from time to time.

It must be emphasized that the software to record, store, read and interpret the climate data in question must be migrated from older computers and computer technology and updated. If this is done often enough, the data will be rescued before there is a problem and they are lost. The decision to migrate media should be reviewed perhaps every two to five years, and the decision to update computers and computer software perhaps every five to seven years.

Finally, JPEG, TIFF and PNG are considered the preferred formats for storing images, and plain ASCII text data files may be considered the preferred format for the easy reading of digital data and for moving data from one computer to the next. If databases are used, the data must be easily exportable to an ASCII format (such as a .csv file). For the sake of migration, it is also important that the digital code in which the files are stored be documented and straightforward (that is, that it not require decoding).

6.3 Recommendations for archiving digital media

(a) Safeguard both digital image and digital data files. Digital image files may be considered official records and should therefore be protected.

(b) Organize the data for easy future access, enabling retrieval by digital storage device or by directory structure (station, year, month). Retain an inventory of files on each digital storage device and in the master directory.

(c) Ensure that the CDMS is backed up on a daily basis.

(d) Store digital images and data on appropriate media so the data will continue to be available indefinitely, migrating datasets as required to mitigate the risks of obsolescence and decay.

(e) Store digital media off site to preserve electronic images and digital data. To safeguard the data in the event of a man-made or natural disaster, store copies of the images and data in multiple locations, including in a multinational Regional Climate Centre outside the immediate region.

(f) The method by which future users can access the rescued data should be communicated clearly.

(g) Establish and maintain an institutional long-term data management and archival plan.

APPENDIX 1. A GLOBAL ACTIVITY

1. WHO SHOULD UNDERTAKE DATA RESCUE?

Any group or individual who has paper, microfilm/old media, or digital data should attempt to facilitate environmental data rescue. Those who are responsible for the management of a nation's environmental record should have a special role in data rescue since they are in a better position to appreciate and value the data being rescued and to know which are most important. These staff are often located in governmental agencies, including NMHSs, for which data rescue is a component of their custodial responsibilities, as well as in universities and research institutions. However, data rescue proponents can be found in many other institutions, both public (agricultural departments and so forth) and private (plantations, agribusinesses, etc.). In addition, several volunteer organizations, such as the International Environmental Data Rescue Organization (IEDRO) and the Atmospheric Circulation Reconstructions over the Earth (ACRE) initiative have been created to facilitate data rescue. Other individuals may also be interested in assisting with data rescue activities and could play an important role including in crowdsourcing operations.

2. WHERE SHOULD DATA RESCUE ACTIVITIES TAKE PLACE?

There are a number of approaches to selecting the location for data rescue. One could be to select a location within the country where the data to be rescued are stored. This may be in the NMHS headquarters and/or in various provincial offices. The best approach may be to gather the data to be rescued in one location or, if the expertise exists or can be developed in each region, to have the imaging and preservation take place in provincial offices. Another possibility is to work in regional, multinational climate centres, such as the WMO Regional Climate Centres. The advantage to this option is that the expertise and equipment in these offices may exceed what individual nations are able to afford. An alternate approach is to have a third party perform the imaging or digitizing outside of the country or region, with all documents returned to the originating country. The best option will depend on individual circumstances. Before the data are transferred to a third party, the WMO Secretariat would be happy to provide input on planning the location of data rescue activities.

The decision about in-house or outsourcing solutions will depend on the project: its duration, the number of images to be scanned, the staff required and staff available, the financial resources, and so forth. Apart from the more general and strategic pros and cons of outsourcing, it is important to take into account the risks attached to the transport of the paper documents (removal from the archive, transport through the building and into a vehicle, storage conditions in the vehicle and at the location of the company/entity doing the imaging, transport back to the archive).

3. THE C3S DATA RESCUE PORTAL

The C3S data rescue portal is a Web-based resource for those interested in data preservation, rescue and digitization. It provides a single point of entry for information on the status of past and present data rescue projects worldwide, on data that needs to be rescued and on the methods and technologies involved. It is a gateway for the exchange of information on all aspects of data rescue, including established and emerging rescue technologies. Such information is also applicable to other scientific disciplines.

Because its goals are to enhance the visibility of existing and to stimulate new data rescue activities, as well as to better coordinate international data rescue efforts, the C3S data rescue portal is a useful communication tool. Each project listed in the portal allows the excellent work of each county to be viewed by others with a similar interest. The C3S data rescue portal focus on informing the community on past and existing data rescue activities will also assist in identifying gaps and opportunities, help prioritize data rescue in regions where it is most needed and aid in attracting funding for projects.

APPENDIX 2. COMMON EXAMPLES OF TYPES OF DOCUMENTS THAT MAY CONTAIN METEOROLOGICAL, MARINE AND HYDROLOGICAL DATA (AND METADATA)

Manuscript sources – terrestrial

- Meteorological registers (bound)
- Meteorological forms, sometimes called landforms (single sheets)
- Metadata station forms
- Rough or draft notebooks
- Barograms
- Thermograms
- Tephigrams
- Pluviograms (rain gauge charts)
- Autographic water level charts
- Daily staff gauge cards
- Hydrometric archives
- Hygrograms
- Solar duration strips
- Synoptic charts
- Weather balloon records
- Anemometer registers
- Hydrological registers
- Records of ice and snow depth
- Lighthouse registers
- Expedition reports
- Weather diaries
- Official reports on natural disasters
- Newspapers (considered here as a primary source)
- Ecclesiastical records
- Missionary records
- Consular records
- Colonial records and government gazettes
- Insurance records
- Agricultural records
- Phenological records
- Pyrheliometer record
- Meteorological and hydrological collections (e.g., monthly bulletins, yearbook reports, monographs, observatory reports)

Manuscript sources – marine

- Ship logbooks
- Marine meteorological logbooks (also called weather books, meteorological registers, or meteorological journals, and often including a metadata form)
- Marine meteorological forms
- Abstract logbooks
- Registers of synchronized weather observations
- Barograms
- Thermograms
- Tephigrams
- Synoptic charts
- Remark books
- Marine expedition reports
- Research ship cruise reports
- Marine cable-laying reports and logs
- Whale and fishery catch books
- Whale and fishery day reports

- Whale and fishery inspectors' diaries, field notes and reports
- Other miscellaneous records connected with whale fishery regulation (the International Whaling Commission and its predecessors)
- Marine insurance records
- Hydrographic and surveying reports
- Dredging reports
- Tidal records
- Ocean current reports
- Typhoon and storm reports
- Ice reports
- Ice charts
- Marine radio weather logs and reports
- Harbor master records
- Ship newspapers

Other sources

There is a wealth of environmental data to be found in print. Sources range from expedition reports to annual environmental summaries to articles in scientific and geography journals and other publications. Contents can range from tabulated data to textual records. Printed sources, such as an account of a voyage or an expedition, can sometimes provide useful metadata on instruments or observing methods. Printed books and articles can often provide evidence of the existence of observations and sometimes where they can be found; the footnotes and bibliographies in these sources can often be enlightening. Printed sources, however, are no substitute for the original documents because they are frequently edited versions of the original material.

APPENDIX 3. DATA RESCUE ASSESSMENT

Table 3 provides a general guide for those involved in data rescue assessments for new data rescue programmes.

Table 3. Checklist for data rescue assessment

1.	Review C3S data rescue portal for technical resources and good practices related to data rescue
2.	Download any known inventories of data from international databases, libraries, literature to assess the country's known environmental data assets
3.	Determine the current status of data rescue in the country
4.	Visit and evaluate existing archive rooms
5.	Talk to director and staff to learn why data rescue is important to the entity and its data rescue priorities
6.	Advise on development of inventory of original sources, data and metadata, if necessary
7.	Advise on paper/old media storage, space and supply needs, and work, health and safety considerations
8.	Advise on methods and equipment needed for imaging
9.	Assess CDMS/HDMS capabilities and infrastructure relevant to data management (age, backup, contents)
10.	Review current digitization and QA processes to see how newly digitized data can fit into the work stream and QC process or if the current digitization and QC process needs revision
11.	Advise on methods and equipment needed for digitizing
12.	Advise on imaging and digitizing inventory
13.	Advise on metadata collection, archival and imaging (see Aguilar et al., 2003)
14.	Advise on key data entry or chart data tracing
15.	Advise on potential use of rescued data to show value to donors/NMHSs
16.	Advise on personnel required for imaging and digitizing steps
17.	Estimate timeline of each data rescue component based on size of task and number of staff
18.	Develop data rescue plan based on priorities of the NMHS
19.	Add data rescue mission information to C3S data rescue portal

APPENDIX 4. INFRASTRUCTURE, EQUIPMENT, SUPPLIES AND PERSONNEL

In all cases, storage, imaging and digitizing should take place in a secure, well-lit, dust-free and climate-controlled setting (the minimum requirements for climate control include air-conditioning and a dehumidifier). Equipment and supply acquisitions will vary according to each country's data rescue needs and the methods selected for use. Table 4 can be used to begin planning the details of a data rescue project.

Choosing the appropriate personnel is critical to the success of a data rescue project. A climatologist is ideal for searching for climate data and making decisions on organizing an archive, while hydrologists and oceanographers are ideal to assist with hydrological and marine data, respectively. Imaging procedures are best established by someone interested in preservation, possibly a person with an interest in photography. Individuals who are careful and pay attention to detail make the best imagers and digitizers, since speed and accuracy are important for these tasks. Several people may alternate between imaging and keying to reduce the boredom of repetitive work and to remove the single point of dependency on an individual. For chart data rescue, consistency in editing is important and may be best done by one individual per meteorological or hydrological station. Once edited, the task of digitizing the charts could be carried out by a team of practitioners.

Both imaging and digitizing proceed more efficiently when the personnel are familiar with computer technology and when the process is done in pairs: one person to image and one to check for imaging quality, and two to enter or trace the same page for digitizing (double-keying). It is strongly recommended that the imaging and digitizing work be done in half a day or four-hour periods as both tasks require undivided concentration.

The number of cameras and type of copy stand required depend on whether the data are presented as bound or unbound books. When photographing one sheet at a time, two cameras are suggested, one to be used only as a backup. A copy stand is used to position both the camera and the lights. To photograph data in a bound volume, however, a camera stand uses two cameras. This system enables the imager to photograph two pages at a time without damaging the binding. Appendix 7 and the C3S data rescue portal provide information on camera stands.

Table 4. Planning considerations for data rescue missions

	Organizing	*Imaging*	*Digitizing*
Infrastructure	Well-lit room, with climate control and security	Room with sufficient electrical sockets and lighting, as well as sufficient tables for sorting paper forms	Room with sufficient electrical sockets and lighting
Equipment	Acid-free archival boxes Double-space shelves (2-m wide, 2.5-m high shelf) Stepladder	Two or more cameras (see Appendix 7 for features and accessories) Copy stand Computer Book cradle Desk Technical manuals	CDMS/HDMS, digitizing board (as specified by chart digitizing software)
Supplies	Gloves, overalls, facemasks, dust cloths	Digital storage devices/cloud storage space USB cable or wireless communication (WIFI)	
Data rescue personnel	Two to clean documents, organize and develop archive and develop inventories	Two to photograph, quality check, store and update inventory	Two to digitize and update inventory
	Training in work, health and safety management in case of contaminated records		
Personnel for follow-up	Skilled person to update inventories and to monitor room conditions	Climatologist/hydrologist/marine/environmental expert to oversee new data and metadata ingest	Climatologist/hydrologist/marine/environmental expert to quality control and develop products

APPENDIX 5. GUIDANCE ON MAINTAINING PAPER ARCHIVES

Paper archives should be kept in a dedicated storage room, preferably not in a basement or attic. The room should be kept clean of dust, dirt, mould and pests. Sturdy shelving (preferably metal shelves) and acid-free archive boxes should be used to keep documents off the floor. The room should be well-lit, and boxes should be labelled and easily accessible. The humidity and temperature should ideally be regulated to keep a favourable room climate and provide good storage conditions. A dedicated person should regularly check the condition of the archive room.

Keep all paper documents at least until (a) imaging has been completed; (b) all images have been checked for quality and completeness; and (c) all images have been duplicated. It is important to make multiple electronic copies of the inventories and entire image archive and to store these in several different off-site locations.

It is preferable to always keep the paper copies for a number of reasons:

(a) Paper copies may have indications of measurement or instrument errors (pencilled or penned slashes) that are not readable on the electronic copies;

(b) The original copies with correction marks and notes are important for the verification of values;

(c) Electronic copies can be lost due to computer failure;

(d) Paper copies thus far have had a longer lifespan than digital and electronic media.

Maintaining the paper archive as a backup is prudent.

If there are no space limitations, the paper archive should remain at the NMHS. If resources do not permit the continued storage of paper archives, national or university archives and libraries should be approached. A country’s national archives may set legally binding standards about how long data records (both paper and electronic) should be kept. If these national institutions are not interested in storing the paper archive, international archives should be contacted.

APPENDIX 6. PRIORITIZING CONSIDERATIONS FOR IMAGING AND DIGITIZING DATA

While it might be ideal to image and key all data at once, the reality is that the archival process will take time and the process must begin somewhere. A start might be made by filling in gaps in existing time series. The International Data Rescue Meeting (Geneva, 11–13 September 2001) established that the highest priority for rescue of both historic and current climate data should be that data:

(a) Are of high quality;

(b) Are of importance nationally, regionally or globally and are at risk of loss;

(c) Would fill in gaps in established datasets.

1. IMAGING

When imaging paper documents, it seems logical to begin with the oldest, rarest or most fragile documents, one station at a time. Since paper and ink quality has declined in recent years, however, newer documents may need to be imaged sooner. When prioritizing the imaging of charts, one might begin with the charts of most importance to the NMHS (perhaps rainfall, pressure, wind or sunshine charts). All pages in a book (for example, a monthly register) should be imaged to ensure that all metadata available (including site and data descriptions and photographs, instrument information, calibrations, observing practices, and quality-control procedures) are rescued.

2. DIGITIZING

As digitizing data is time-consuming, prioritization is important. The factors below are to be considered when digitizing surface observations:

(a) To show proof of concept to potential funders, it may be advantageous to select the station with the longest continuous records or to concentrate on an unusual year that exhibits a particularly topical environmental feature (such as a drought or a flood) as a way to begin;

(b) It might be prudent first to key stations that have some data already digitized, that are still operational, or that have the longest period of record;

(c) Fewer data are available worldwide in the pre-1950s period. In individual countries, however, any data prior to 1990 might be crucial;

(d) Location is another factor to consider. It might be beneficial to select stations that represent different climatic regions within a country or socioeconomically significant locations;

(e) Data availability might be another consideration. A station with the most reliable data and least amount of missing data might be selected first;

(f) Neighbouring stations, however – even those with a short duration – are important to corroborate measurements;

(g) For current or historical data, the temporal resolution of the data (monthly, daily, sub-daily), or the need for particular parameters could determine the priority for digitizing data;

(h) If, for example, a climatology of diurnal changes in temperature, precipitation or wind is desirable, then chart data or hourly (sub-daily) observations are crucial;

(i) For climate data rescue to examine essential climate variables, select meteorological variables that can be converted into monthly values such as temperature and precipitation (Bojinski et al., 2014);

(j) For areas where there is a spatial gap in the data, individual stations might be selected;

(k) For climatologies of thunderstorms, hail, tornadoes, dust storms, fog and ice storms, non-instrumental visual weather observations are required.

Surface observations are usually considered for imaging and digitizing, but upper-air soundings, surface and upper-air charts and marine observations are also of great importance.

(a) Individual soundings provide valuable information on changes in temperature, pressure and moisture with height;

(b) As individual sounding data may be lost, such information might only be derived from upper-air charts;

(c) Surface charts can also be quite important as original forms have sometimes been lost after surface station data were plotted.

These are the general guidelines and considerations for setting data rescue priorities. Each country or centre must decide on its own priorities based on the primary needs of the country and the condition of the climate data records. Once a station and year have been selected, it is wise to test the imaging and digitizing procedures. This will help determine staffing requirements and how to streamline the process.

A final reminder on digitizing forms: it is best to digitize all data on a given page, including metadata (since it is unlikely that there will be another opportunity to return to that page and record any omitted data). This can best be done by creating a template for digitizing the various form types. In all cases, the metadata, the description of the data and the observing practices (observer name, observation time), location (latitude, longitude) and site characteristics must be rescued and digitized together with the imaged or digital values.

3. SPECIFICITIES OF HYDROLOGICAL DATA RESCUE

Prioritizing river gauging stations

When rescuing data for surface water gauging stations, it is very important to ensure that data from the most important stations are selected. The prioritization and selection of gauging stations is not straightforward as the resulting data will be used for a number of different purposes, which may require information from different types of stations. Priorities are likely to vary significantly depending on local circumstances and will be informed by an understanding of the current availability of data and the most pressing needs for further information.

Below is a list of criteria, in a suggested order of priority (most important first), that should be considered when selecting stations for data rescue:

1. Length of record;

2. National operational importance of river/watercourse (for water resources, flood risk and so forth);

3. Ability to provide information on hydrological change (for example, natural flow regime);

4. Ability to provide information to improve hydrological understanding, for example, of the representativeness of a hydrological regime within the country including how to assess this representativeness;

5. Perceived quality of gauging station (data);

6. Perceived future potential in respect of 2, 3 and 4 above.

These criteria are obviously not entirely independent of each other; for example, a very low-quality record under point 5 may not be of great significance despite its length or the importance of the river.

Prioritizing surface water records

There will probably be several different types of records relating to river flows at a given gauging station, and there will again be a need to prioritize what to rescue when resources are limited.

The list below suggests priorities (most important first) in selecting these records:

1. Processed river flows (for example, daily mean flows);

2. Summarized river flows (for example, monthly mean flows and annual maxima) and previous reports and analyses of hydrological data;

3. Key gauging station metadata and additional metadata concerning data quality, such as a description of hydraulic structures in place, if any, and major influences within the catchment that facilitate the interpretation of data;

4. Stage records;

5. Instantaneous discharge measurements (gaugings);

6. Stage-discharge relationships (ratings);

7. River cross-sections;

8. Other gauging station metadata such as site surveys and photographs.

These priorities may change due to specific local concerns, type of station and quality of documents.

Prioritizing other records

The rescue of other datasets will depend on the availability of such data and on other aspects such as the quality of records. However, the following points should be considered:

1. Rainfall: Data are often widely available and rescue of a long record of rainfall data would be very important. A scarcity of rainfall records would make data rescue a priority. Confirmatory records in data-sparse areas or periods also have great value;

2. Snow: Records of snowfall and snow depth are of crucial importance in some regions and are often very hard to come by;

3. Evaporation: Measurements of actual evaporation could be extremely useful to future hydrological studies;

4. Automatic Weather Station (AWS): A full and good-quality meteorological dataset from an AWS can allow the estimation of potential evaporation and would be considered a useful record should similar datasets not be widely available;

5. Meteorological station metadata: Additional information about widely used meteorological stations provides clearer insights into the quality of the data. These metadata may include exact location and any changes over time, period of record and variables measured, and also information on site exposure, types of instruments used, procedures for reading instruments, and so on;

6. Thematic data: Maps or spatial data can enable hydrological processes to be better understood and are therefore important in hydrological studies. Thematic data could include soil maps, geological maps, land cover information or even cartographic maps containing the elevation and locations of rivers. NMHSs may store such maps, which are no longer available elsewhere, and these would be a high priority for data rescue.

APPENDIX 7. ELECTRONIC IMAGING TECHNIQUES

Creating an electronic image is highly desirable because it reduces handling of the paper or other old medium. It also allows the image to be resized and enhanced to make the writing more legible when key entering the data. Typically, JPEG images are suitable, and uncompressed image formats (for example, uncompressed TIFF images) have the advantage of being at least partially readable even if some data are corrupted. Depending on the size of the original document, some sources suggest that an image size of 1 600 x 1 200 pixels is sufficient.

1. SCANNERS

Scanners are often considered for imaging paper documents. The two general types are flatbed scanners (frequently used for books) and feeding scanners (often used for individual pages, such as charts). Scanners are an attractive alternative to cameras as they may reduce the amount of equipment required for imaging. There are a number of drawbacks to scanning, however:

(a) A flatbed scanner can damage book bindings or may occasionally result in the data closest to the binding being out of focus or not imaged. This can be avoided with a camera set-up. Note that dedicated book scanners exist;

(b) Feeding scanners via an automatic document feeder risks causing single sheets to get crumpled or even destroyed when caught in the scanner (paper jam);

(c) Microfilm and microfiche require specially designed scanners;

(d) Scanners have many movable parts and usually require maintenance and a license;

(e) Specialized training is often required with the more complex scanning machines. Because of the training issue, changes in personnel have brought scanning to a stop in some projects. In other cases, inadequate training and licensing issues have prevented the use of scanners for imaging;

(f) Scanners are generally slow and cumbersome, except for very expensive systems.

2. CAMERAS

Cameras are the tool most often used for creating digital images. They are effective with both bound and loose paper copies, as well as for oversized documents, such as maps. Digital cameras are commonplace, and even automatic cameras can provide reliable images, especially with a constant source of good lighting. With the use of a stationary camera stand and lighting system, a remote control for taking pictures and an alternating current (AC) adapter, little needs to be done except to change pages, making this process quick and efficient. In addition, inexpensive software is now available to retitle digital images and move images to a computer. The archiving of images is also very rapid.

3. CAMERA FEATURES AND IMAGING WORKPLACE EQUIPMENT

The equipment needed to set up a photographic imaging workstation is summarized in Table 5. First, it is necessary to provide a table with a computer and the camera set-up (including a tripod or camera stand). An adequate number of grounded electrical sockets is crucial. Although natural lighting or sunlight is ideal for imaging – taking care that the light is diffused and shadows are avoided – a room with natural lighting is not always available. Thus, a camera set-up often includes two or four lights. A tripod or camera stand placed in conjunction with the lights allows the camera to be kept stationary so that multiple pictures can be taken quickly. It is important to test the lighting frequently since some paper media are glossy and reflect light back into the camera lens, obliterating some of the imaged data. A remote-control feature allows a picture to be taken without the photographer being positioned alongside the camera and also reduces the chance of jarring the camera body. A computer to hold the digital images is crucial.

Table 5. Basic camera infrastructure requirements

1	Table
2.	Electrical sockets
3.	Camera stand/tripod
4.	Lights
5.	Camera and accessories (including a replacement camera)
6.	Computer

Cameras generally drain batteries in a relatively short time. An AC adapter allows an unlimited number of pictures to be taken and saves considerable time and expense in purchasing and replacing batteries. If an AC adapter cannot be used with the available camera, then spare, rechargeable batteries and a battery charger are recommended. Importantly, if an AC adapter is not available and batteries must be used, the camera design should allow the batteries to be replaced without removing the camera from the camera stand. Most digital cameras have plastic threads to hold the camera on the camera stand. These threads wear out after a few dozen camera removals for battery replacement.

Software and a USB cable or access to WIFI is needed to automatically transmit the image from the camera to the computer. If such software allows the mouse to engage the camera shutter, a remote-control cable may not be necessary. If a USB-port or WIFI is not available, the largest memory card obtainable should be used. Digital cameras often include software to generate filenames for each image and transmit photographs directly to a computer. Such software saves the photographer a considerable amount of time. Depending on the camera available, a zoom lens and the ability to change camera settings is useful. Table 6 provides a summary of the desirable camera accessories that allow for an efficient process.

Table 6. Camera accessories to facilitate an efficient process

1.	A copy stand that holds a camera and two to four lights
2.	An AC adapter to power the camera
3.	A remote control for taking pictures or camera software such as in item 4
4. 5.	Camera software to take and transfer photographs directly to a computer via a USB cable or WIFI Alternatively, the largest memory card available
6.	Software that allows the image files to be renamed
7.	A zoom lens if readily available and a macro-lens for close-ups
8.	Hardware or software that allows the camera settings to be changed (via camera or computer)
9.	Glare-free glass to gently flatten bound records

Consider the size of the image, particularly when using a memory card. Depending on the quality of the image, it is likely that the largest possible image is not required and that a medium-sized image may be sufficient. It has been found that even a 4-megapixel camera is enough to produce JPEG files with enough resolution to result in readable alphanumeric parameter values. Currently, even mobile phone cameras have about 12 megapixels of resolution. It is important to note that excess fine resolution, while not needed for alphanumeric data recovery, may be needed for use with automatic computer digitizing programs. Typically, photographs are taken in colour, not black and white. Colour allows corrections and other notations found in original notebooks to be read more easily.

Other common camera features to consider are the ability to cancel the flash, a macro for close-up mode, the ability to adjust the film speed sensitivity, and the ability to adjust the white balance. Table 7 summarizes common camera features to ensure picture quality.

Table 7. Common camera features to ensure picture quality

1.	Ability to select the size of the image
2.	Ability to cancel the flash
3.	Macro close-up mode
4.	Ability to adjust the film speed sensitivity
5.	Ability to adjust the white balance

4. THE DOS AND DON'TS OF IMAGING

– Never take images using a flash.

– Always use a camera stand or tripod.

– Always use a remote switch to operate the camera. This is much quicker than pressing the shutter button and avoids causing the camera to move unnecessarily. Some camera software allows the camera control to be handled by the computer.

– Make sure that the camera sounds are muted to avoid disturbing others.

– Ensure sufficient and steady lighting and avoid shadows on the document.

– Do not fill the entire frame with the document; instead, leave a small border around all the edges.

– With bound documents, take double pages with each shot, providing the volume can be laid reasonably flat. The size and format may dictate otherwise.

– Always photograph the entire document, including the cover (it may have the archive reference fixed on it), the title page, any pages containing instrument metadata, observing instructions and notes on the observations.

– Use colour photography – it is not unusual for colours to have a specific meaning (for example, temperatures in red may indicate negative temperatures even if the negative sign is not written).

– Image all pages of a collection (this may include blank pages and covers). Some notations may hold across pages (for example, sometimes negative/positive signs are only indicated when the sign changes).

APPENDIX 8. OTHER DIGITIZING METHODS

The following is a general discussion of digitization methods other than keying by selected individuals. All methods require quality control since errors can occur with any transcription.

1. OPTICAL CHARACTER RECOGNITION, INTELLIGENT CHARACTER RECOGNITION AND INTELLIGENT WORD RECOGNITION

Optical character recognition is a commonly available software technology used to digitize typewritten text or printed tables; however, it has shown mixed success for data rescue projects relating to handwritten observations (Wilkinson et al., 2019; Lorrey et al., 2022). Intelligent character recognition (ICR) targets individual printed handwritten characters and allows for different fonts and styles of printing; it works best for reading structured forms. Intelligent word recognition (IWR), on the other hand, translates entire handwritten words, either printed or cursive. Historical forms are generally not ICR/OCR-compliant and do not work with current ICR/OCR technology. ICR and OCR use machine-learning technology to update recognition databases automatically for new handwriting patterns.

At the time of writing, OCR, ICR and IWR were under development and are likely to become more reliable over time. At present, digitization by keying is used most often in the belief that the human mind can best translate handwritten materials.

2. CROWDSOURCING

Introduction

A crowdsourcing approach to digitization involves using online volunteers, often referred to as "citizen scientists", to process information.

Even though in this case, the keying is done by volunteers, there are still costs. Time, talent and funding resources are required to develop and maintain the infrastructure for Web-based data ingest, database development, quality control, to develop training instructions for keyers/observers and for outreach to attract potential volunteers.

Major advantages to using Web-based citizen science for data rescue efforts are that human resources can be drawn from all over the world, volunteer time is free, and progress is made more or less continuously. In addition, citizen science data rescue provides an opportunity to engage and educate the general public about the importance of long-term meteorological observations and climate change. It also improves the transparency of nation- and archive-specific data holdings.

The following key recommendations and lessons learned regarding crowdsourcing are taken from the Southern Weather Discovery citizen science project (for more details, see Lorrey et al., 2022).

Pre-condition

The data to be rescued have been identified, and good-quality electronic images are available.

Overall approach: Define a project and a workplan

Set up an internal project team, create a citizen science project identity, project description and website:

– Consider two basic website components: a front end component, which the general public can see and work with, and a back end component, which contains the design and content elements required to organize workflows and create data entry fields;

– Humanize the project by providing a face behind the science as well as key points of contact;

– Include the ability for citizen scientists to contact the internal project team for any questions and to engender a better connection with the public engaged in data transcription. Allocate sufficient time for the project team to respond to questions from volunteers, particularly in the early stages following the initial project launch;

– Define the architecture of a basic workplan to determine how the division of labour should proceed.

– Project documentation: Make sure that all project-relevant original observations/logbooks, their digital copies (images), forms used, software and code versions, process descriptions, and so forth are kept after the end of the project.

Communication aspects

Promote the project, engaging with the media and ensuring that the public is strongly connected to the project's overall success. Consider the following points relating to project communication:

– Define and implement a strong communication strategy with a ''hook'' to get people involved;

– Ensure that the project team is willing to engage with the media and the project participants;

– Promote the project on multiple social media platforms;

– Make repeated contact with the citizen science community using emails and updates as tasks progress;

– Set up a multiphase communication plan to introduce and promote the project, including video and print media campaigns to garner participation and maintain interest;

– Consider identifying an "attracting" meteorological event to start the project (for example, a storm event, snowfall event, heatwave event, etc.).

Technical aspects

– Consider specialized workflows, including workflows for replicated keying, to enable volunteers to participate in transcribing data;

– Workflows should be designed to be as simple as possible and should utilize the logbook clips discussed below rather than displaying a whole logbook page;

– Testing and retesting of workflows is required to ensure that they are simple to follow (heeding participant feedback);

– Evaluate initial inputs and data retrievals with a small dataset before launching a full data rescue campaign;

– Prepare enough material in advance to ensure that momentum can be continually maintained.

Replicated keying of logbook segments is designed to provide a majority consensus (and a measure of confidence through replication) that defines what numeric value exists in each table cell. Based on tests, a compromise can be reached between time spent keying by citizen science volunteers and achieving completeness and accuracy of a transcribed dataset when, for example, three or eight or any other number of replicate entries are employed. It is also worth noting that these results are dependent upon the nature of the data being transcribed. Integer values with no decimal points are the most straightforward to key and require little repetition to ensure a correct consensus value. Conversely, alphanumeric values and values with many significant figures introduce more complexity or opportunity for variation among responses from the citizen scientists (for example, representing a decimal with a period, a comma, a space, or ignoring the decimal altogether). Eight replicate data entries were used in the Southern Weather Discovery project to obtain consensus with minimal errors.

Logbook clips: Individual meteorological register pages may be subdivided into small parts to provide a segment, rather than a whole page, for an individual volunteer to transcribe. This approach helps to (i) ensure that data keying contributions can be completed in short bursts rather than lengthy intervals of time, (ii) minimize the risk of volunteers abandoning a data entry form before submitting a full transcription, (iii) reduce mistakes that are associated with transcribing the wrong column or row, and (iv) decrease the probability of widespread errors propagating across an entire log if, for example, a specific volunteer has a difficult time deciphering the handwriting on a specific page.

It is also possible to integrate data rescue by citizen scientists into the classroom. This provides students with multiple learning benefits while adding resource capacity to the data rescue project. Data rescue activities may be incorporated as in-school activities (for example, as a partnership between the agency leading the data rescue and a secondary school) or as an accredited assignment for tertiary students at university (again requiring a partnership between educators and data rescue practitioners). Surveys from recent citizen science projects in Ireland demonstrate that students felt they had gained an appreciation for how data are created, why they are important and what they can be used for, and could see the value of their contribution. Further insights on the use of data rescue in the classroom and guidance on classroom procedures can be found in Ryan et al. (2018) and Mateus et al. (2021).

3. CHART DIGITIZATION

Examples of chart data include stripcharts and autographic charts that record continuous measurements of meteorological and hydrological variables such as temperature, wind speed and direction, atmospheric pressure, precipitation, and water levels. Digitization may involve (i) the manual transcription of values from the chart into tabular form, or (ii) the use of specialist digitization software.

Manual digitization is a tedious task, especially when charts have rapid fluctuations or there are multiple lines on a single chart. While fine time resolutions can be read from a chart, peak or daily values are often the only values reported. In contrast, chart digitization software more easily enables the higher temporal resolution of data values from minutes to hours.

Various digitizing software programmes exist to allow trained practitioners to manually trace a chart line and convert points along that line to digital numbers, often using specific tracing boards (hardware). One potential issue is that this software may only be compatible with certain computers and operating systems. The compatibility of the digitized data (file type) with the relevant CDMS and HDMS is critical for and may limit the choice of digitization software.

Although it may be possible to obtain the source code of the digitizing software to adjust it for specific output formats, this code may need to be maintained (updated for newer programming languages and computer technologies) and adapted by a software technician.

Several digitizing software programs have been developed and are in use. Features of these programs include curve extraction from images of the raw chart data, the ability to read a variety of chart types, and the ability to handle multiple lines on a single chart. Documentation on this software can be found on the C3S data rescue portal. Some of these programs are available for use at no cost, with the stipulation that the data be placed in a free and open database available to all.

Converting chart lines to digital values is only part of the chart digitization process. The other part is known as "editing" and is a necessary quality control process to mitigate human and equipment errors from being transferred to the digital output. For example, editing should confirm that the line as it appears on the paper chart is correct relative to the axes for the start and finish time, date, and measurement value.

The editing process makes use of information such as calibration checks and other annotations to provide details of required corrections where the chart line is believed to require adjustment as it is converted to digital format. For example, if a chart line began at the wrong measurement value because of human error, the trace line will require correction in the digitizing software so that the digital numbers are accurate. Establishing editing procedures at the start of a chart data rescue project is therefore critically important to ensure consistency between practitioners and the production of precise and accurate digital records (see, for example, de Smeth et al. (2024)).

The sequencing of editing and digitizing is at the discretion of the NMHS and may be influenced by time constraints and data rescue priorities. Editing generally requires more time than digitization and should be undertaken by those with specialist knowledge (for example, hydrologists or hydrometric officers). Due to time and resource constraints, an NHMS may decide to digitize raw chart records and then edit only specific flow events within its HDMS. This may occur, for example, if the purpose of data rescue is to assist with understanding flood risk. Alternatively, an NHMS may decide to edit a set of records (for instance, by calendar or hydrologic water year) and then to digitize them in annual batches.

Editing introduces additional record keeping requirements for the data rescue of chart records:

(1) The unedited and edited chart data must be stored and accessible to users of the CDMS/HDMS so that they can cross-check between the original and produced records;

(2) Contextual information, such as calibration checks and annotations, should be recorded alongside the edited data;

(3) Editing decisions should be recorded:

   a. Within the CDMS/HDMS that holds the digitized and edited output as annotations on the produced records;

   b. As a detailed account of chart-by-chart decisions for internal reference by the NHMS; and

   c. As an aggregated monthly/yearly summary for reference by future data users to support their understanding of historical measurement or equipment issues and data quality history.

This information is an important source of data quality metadata and should be stored in the same location as the station metadata and made available to future data users. These metadata essentially reflect measurement issues identified during the data rescue process and the solutions applied to resolve them. Making this information available to future data users provides them with information on measurement quality that may be important for their intended use of the data.

Editing decisions should be also reflected in the quality coding approach taken for the produced digital data. See Appendix 9 for more information.

APPENDIX 9. DIGITIZING ERRORS AND QUALITY ASSURANCE PROCESS

1. QUALITY CONTROL FOR KEYING ERRORS

The following is a brief overview of the quality control of digitization errors specifically relating to keying. Some or all of these quality-control features may be available on the data management systems (for example, CDMS, HDMS, etc.) of Members' agencies. If this is the case, as data are keyed into the data management system, errors are flagged and later corrected. If the data management system does not incorporate these quality-control features, it is strongly suggested that they be considered. The initial quality control of newly key-entered digital data is centred on keying, instrument, observer and metadata errors, not on the more refined analyses performed during homogenization, sometimes called "delayed mode" quality control. The keyed data should be considered (and flagged) as provisional until they are verified through quality-control tests.

The use of double-keying, templates, and one-to-one keying minimizes keying errors in individual values, but keying errors still occur. Problems due to metadata issues (largely related to station naming and identification number), scanned image quality, instrument malfunctions and non-standard observation practices remain challenges. In order to ensure the most faithful digital representation of the data, quality-control tests should be used to flag suspicious values of both the metadata (for example, station name or identification number, year or month, element type) and the meteorological variables. Suspect values are flagged for manual examination, with flags added indicating the type of error and the verification outcome.

Common keying errors include transposed numbers, missing or misplaced decimal points, missing digits (for example, a zero), incorrect accumulation periods, and missing values (relevant items for recommended instructions for digitizers are listed below). Other common errors include: observer errors (for example, non-conventional units); instrument malfunction; metadata problems (station identification or naming convention inconsistency, incorrect year or month); transposed or duplication of a particular element for a given month and station; and unclear forms. These errors could indicate formatting issues and systematic data quality issues.

Flags are used to identify the reason for each error. Once verification of the flagged error value is complete, the assessor selects the appropriate verification type for the data. Values may be treated as: correct, matching the form, reasonable, and left unchanged; an error, flagged but with no change to the data; an error, but with a replacement value specified; non-matching, corrected to match the original form; an error and set to missing; or needing to be keyed, and then keyed. Usually, when a flagged value is ambiguous and could be correct, it is not changed.

2. QUALITY CONTROL TESTS

Quality-control tests and procedures assure users that the digitized data accurately represent the observations recorded on the original documents. If tests are not run automatically by the data management system, a quality-control inventory spreadsheet should be developed to track each test for each station. The outliers from each test need to be checked manually and any corrections applied at the conclusion of each test. After each quality-control step has been performed on a station, the quality-control accounting spreadsheet is updated.

Quality-control tests include range checks (on daily values, monthly totals, and means), internal consistency checks (comparing measurements among the data types) and extreme value checks (made on individual values). However, sometimes data quality issues may only be detected when the output of a data rescue project is appended to the existing available data for that particular measurement station (if any exists). An example of this is absolute water level data that may be correctly digitized from the raw paper or chart format, but which plots too high or too low because the incorrect datum was applied. In this instance datum levels in the metadata may have been misread, but the error is only detectable when plotted in context with the rest of the record.

Table 8 below provides examples of quality control checks that may be completed on rescued data as part of quality control procedures specified at the outset of the project. Further information can be found in the *Guidelines on Surface Station Data Quality Control and Quality Assurance for Climate Applications* (WMO-No. 1269), in the *Guide to Climatological Practices* (WMO-No. 100), section 3.3 and in Westcott et al. (2011) and other sources on the C3S data rescue portal.

Table 8. Examples of quality control checks on rescued data

Quality control test	*Description*
Date sequence	Check for missing dates and duplicate dates. Be aware of leap years.
Means	Calculate mean values per time period and plot these to identify any outliers that may indicate errors in data rescue.
Aggregated totals	Sum data by the most relevant time period (for example, daily for sub-daily data, monthly for daily data), and plot these to identify any outliers that may indicate errors in data rescue.
Range of values	Identify the minimum, maximum and range of values per time period (for example, daily, monthly, yearly), and plot these to identify any outliers that may indicate errors in data rescue.
Duplicate values	Compare the length of value entries with the length of date entries to identify whether duplicates exist. For daily data, flag repeat consecutive daily entries (for example, more than three days with the same data value) to check against the raw data.
Outlier values	Define an interval (for example, two standard deviations on either side of the mean) and search for individual outlier values to check against the raw data.
Missing values	Check for missing entries (null values) or values denoted by 'NA' or '-9999' or similar. Ensure consistent expression of missing values.
Data gaps	Either by calculating the number of missing entries per time period, or by visually inspecting the rescued data, identify data gaps to check against the raw data. Obtain expert advice before applying interpolation or gap-filling methodologies, and only as appropriate.
Step changes	Using statistical tests, by visual assessment, or by defining a threshold change (for example, greater than a specific number of standard deviations from the mean), identify step changes and investigate whether these are data rescue errors, whether they potentially relate to other quality issues in the dataset, or whether they are genuine value changes.
Tie-in between datasets	Append the rescued data to any existing data for that station and inspect the tie-in period. This includes looking for potential step changes in the data, noting the shape/smoothness of continuous data and looking at the range of values in both datasets.

3. INSTRUCTIONS FOR DIGITIZERS

Items to be highlighted in instructions for digitizers include:

– Discussion of typographical errors in the source;

– Missing data. There should be a consistent way to indicate missing data;

– Special symbols. The digitizers should be instructed to reproduce each symbol as it appears in the source without interpreting its meaning;

- Negative temperatures. One of the most common mistakes when digitizing temperature data is the loss of the negative sign;
- Non-decimal sub-units. When digitizing early instrumental data, it is often necessary to deal with non-decimal sub-units;
- Visual aids. In the keying templates, it is useful to add borders and/or different fill colours to separate columns and in general to highlight the data structure;
- Disable autocomplete. By default, most spreadsheet programs suggest what to write in a cell based on what was written in the previous cells. This function can lead to errors and must be deactivated before starting the digitization;
- Comments. Even with a careful inspection of the sources, many details can remain unnoticed. The digitizers are much more likely to see them and should be instructed to report any relevant metadata (notes, unknown symbols, unit changes, and so forth) in a comments column. Moreover, the comments column can be used to report problems (for example, a typo in the source, or an unreadable value) and actions taken;
- Handwriting. When the source is handwritten, it is important that the digitizers dedicate a few minutes to get familiar with the handwriting before starting the digitization;
- Dates. Sometimes, pages can be missing, duplicated, or in the wrong order. The digitizers should always check that the date (particularly the month) indicated in the document corresponds to the date in the template;
- Monthly sums. When available, monthly means/sums or similar can be useful for quality control. They should be typed in such a way that they do not interfere with data conversion (for example, in a dedicated column or row);
- Self-check. Many digitization mistakes can be easily detected in a spreadsheet, for example, by checking the maxima and minima in a column or by plotting each column. The digitizers should be encouraged to perform these basic checks and correct their own typos.

Even though most of the instructions can apply to any template, a specific subset of instructions is usually necessary for each template. It is also helpful to provide an example of typed data in the templates.

4. QUALITY CODING APPROACHES FOR ENVIRONMENTAL DATA

Coding of rescued data may be required to convey measurement quality to future users. For example, it is a common hydrometric practice to quality code hydrological data such as water levels, rating curves and river flow records (see *Guide to Hydrological Practices* (WMO-No. 168) for detailed guidance). However, this practice can be extended to any chart data that have undergone an editing process as part of the data rescue. The assignment of quality codes to rescued data may be challenging where there is a lack of detailed information or metadata to provide important context for interpreting the records, especially if historical charts display evidence of some malfunction or poor data collection practices.

Quality codes applied to rescued historical data should complement the modern-day quality coding practices of the NHMS. This is particularly important where data rescue extends the record at a station or fills a gap. A consistent approach to the definition and application of quality codes between existing and rescued data enables measurement quality to be assessed over time. It may be necessary to redefine the numeric or alphanumeric levels of a modern coding approach so that it can be applied to rescued data (see de Smeth et al. (2024) for a worked example from Ireland).

APPENDIX 10. LESSONS LEARNED

Note: This appendix builds mainly on climate data rescue experiences. The lessons learned from these experiences can be generalized and applied to hydrological, marine and other environmental data rescue activities.

1. ENCOURAGE GREATER COOPERATION BETWEEN THE FORECAST AND CLIMATE DIVISIONS OF AN NMHS

The Climate Division of an NMHS often oversees field operators at meteorological sites, quality-controls national climate data (hourly, daily tables) and keys in a portion of the data into its own CDMS. Typically, coded SYNOP data are sent by weather observers using the WMO Global Telecommunication System, but these digital data are often available only to the Forecast Division of the NMHS, and often only in real time. In addition, frequently, only the handwritten hourly and daily data tables, not the coded digital data, are sent to the Climate Division. Furthermore, if the Forecast Division archives the decoded weather data from the SYNOP stations or from the automated stations, they are often not available to the Climate Division. As a result, the Climate Division must obtain data for its country from international databases. It would be ideal if the transmitted digital data were archived in a central database accessible to both forecasters and climatologists and if this central database were used for all products. Better coordination between these groups would allow the Climate Division to have a more up-to-date and, more importantly, a more complete dataset for use in developing climatological analyses. It would also allow the dataset to be compared with the printed reports for extreme value validation.

2. REVIEW CURRENT PRACTICES OF DIGITIZATION, METADATA COLLECTION AND QUALITY CONTROL; TRAINING REQUIREMENTS

The Climate Division of an NMHS often develops its own method of data keying, chart tracing and quality control, which evolves according to the available staffing, communication, and computer resources and the number/type of meteorological stations within the country. It is advisable to regularly review current practices and to incorporate relevant WMO guidelines into the structure for inventorying, keying, tracing and quality control of current data. See the good practices documented in the C3S data rescue portal.

It is advisable to review any insights gained during the execution of data rescue processes that might improve the data rescue or current data ingest process or that might lead to locating other sources of data in need of being rescued. Climate data are only as good as the observers and analysts.

There will be new, unfamiliar equipment and software for the individuals who undertake data rescue work. Training and experience will be needed in the use of digital cameras, scanners, and digitizing software; in accessing databases; and in performing quality control. As people transition from job to job, multiple persons should be trained with “how to” manuals as they are learning the process to ensure that there is no single point of failure. These manuals must be revised as programs and users change.

3. DEMONSTRATE THE BENEFITS OF DATA RESCUE

The reorganizing, imaging, digitizing and quality-control steps of rescuing data can be expensive and time-consuming. This is probably one of the reasons why data rescue has not yet been implemented in many NMHSs. Beginning and sustaining a data rescue programme begins with the NHMS itself, which needs to clearly articulate why data rescue is needed. This could include identifying practical questions that can be answered by improved climate data.

For instance, historical climate data would allow the NHMS to answer a minister's hydrometeorological questions accurately and quickly regarding the frequency of droughts or heavy rainfall events, or to provide the direction of the strongest wind over the past 100 years so that a new USD 10 000 000 airport runway is oriented in the right direction. Such information provides a "value-added" product that NMHS managers could charge for, bringing in additional revenue to the NMHS.

In addition, organizing climate data could result in reclaiming hundreds of square metres of NMHS floor space without discarding valuable historic records. The increased floor space could be converted to additional offices or training rooms. Finally, if the NMHS is willing to share the rescued and digitized data, it would be recognized internationally as a leader in the field.

Outside support over a number of years is often required for data rescue activities. Appendix 11, which concerns applications, may aid in providing further justification for data rescue. Individuals and groups that have already undertaken data rescue can be advocates and may be able to provide advice and argue for funding data rescue efforts.

Consultations with others interested in data rescue can provide technical information and possible solutions to unexpected problems along the way. The C3S data rescue portal provides some technical information and possible contacts. NMHSs may also contact the WMO Secretariat for assistance in procuring aid from other Members.

4. EXAMPLE RESOURCE ESTIMATES AND COSTS (KEYING METHODOLOGY)

Data rescue cost estimates for different scenarios are provided in *A Proposed Long Term Resource Plan for Climate Data Management and Rescue Activities*. In summary, a three-year national data rescue project may need between EUR 30 000 and EUR 80 000 to deliver sustained results. While it will be hard to achieve anything with less money (unless basic data rescue is already part of the operational activities of an institution), there is space for using more money to fully digitize national environmental records.

Informal estimates suggest the following:

- For imaging, one can expect to take 800–1 500 images, and on average, 1 200 images per person per day.

- Thirty to fifty years of climate documents can be digitized (keying methodology) per year by one person working full time at a rate of 10 meteorological parameters and three sub-daily values per day (photographing, keying and quality control included).

APPENDIX 11. APPLICATIONS TO SHOW VALUE DATA

1. VALIDATING EXTREME VALUES OF METEOROLOGICAL PARAMETERS AND PROXY DATA

It is thought that a changing and more variable climate will result in more extreme values in temperature, precipitation, pressure and wind in the future. These future values cannot be validated as "record-breaking" without knowing the extremes that have occurred in the past. Preserving original documentation and digitizing old weather records will allow current and future extreme values to be easily verified. With verifiable weather parameter values, more will be known about the overall distribution of extreme values and whether the distribution of values is changing. Past climate studies in periods or regions with little or no available meteorological data have been based on proxy data (tree ring data, ice core data). Newly rescued data are extremely valuable in verifying the validity of these methods.

2. HYDROLOGICAL DATA FOR MONITORING CHANGE

Hydrological monitoring activities have always emphasized the need for long hydrological records for water resources planning, flood estimation and to maximize the understanding of various conditions. Recently, an additional need for having long-term datasets has emerged: to understand how hydrological regimes are responding to climatic variations and anthropogenic influences.

While global climate models can provide information about the expected impacts of environmental changes, the validation of these models requires real data. More importantly, society needs to understand the impacts of these changes at the national and catchment levels and to identify the corresponding emerging trends or changes in hydrological regimes. This can only be done by assessing long-term records that reflect natural variability. Data rescue provides an opportunity to identify key records that may previously have been considered inconsequential (for example, data on small rural catchments). The relevance of such records for national water resource planning and for a global understanding of the impacts of climate change should be a major driver for data rescue activities.

Additional reasons for the need for hydrological data rescue include the following:

- Longer records can provide more precise and accurate assessments of water resources and flood variability, potentially leading to a more cost-effective or robust water supply and flood protection infrastructure;
- A more dense observation network, or a network monitoring a greater range of catchment types, can provide a better understanding of hydrological processes and improve the results of regionalization methods that underpin many hydrological models;
- Through data rescue, hydrological records can be extended at a fraction of the cost of ongoing field measurements;
- Instigating improvements in data management and archiving practice can bring organizational benefits both through efficiencies and through improved access to information;
- Rescuing and making use of quality metadata can lead to an enhanced understanding of, and thus a more effective use of, hydrological information;
- A comprehensive inventory of data holdings can be useful for demonstrating the value of a data archive and for justifying the ongoing expense of running such an archive.

3. INCORPORATE IN STUDIES, LINK TO RESEARCH PROJECTS

The best way to show the value of recovering historical climate data is by using them in research studies or for others to do so. Long time series allow for the identification and examination of the severity and frequency of past events, such as droughts and floods, heatwaves and cold spells, tornadoes, tropical cyclones, and hailstorms and dust storms. While meteorologists and climatologists are greatly interested in examining the causes of weather phenomena, others are interested in what impact these phenomena have had on society and even on the direction of a country's history as the impacts and human reactions to these past events can aid in the planning for future extreme environmental events. In addition, using historical data to research past environmental events could greatly engender enthusiasm in potential funders, especially if these same phenomena are currently happening or are expected to occur in the future. Being able to show historical and contemporary statistics relating to these phenomena could also help predict their future probability. Complete data time series also help tune and verify Earth system models.

4. MAKING DATA AVAILABLE TO USERS

It is important to make environmental data available and easily accessible to users. Providing reliable historical data allows scientific researchers to engage in scientific exchange and may provide educational opportunities for students. Agriculture and water managers would be good allies in promoting the need for data rescue, and they may have records or resources of their own to contribute. These managers would likely find rescued data essential in assessing the sensitivity of natural and man-made systems to environmental events and extremes. Furthermore, the development of products that would be of general interest to the public (such as the climatology of weather on national holidays) might engender interest in weather and help develop groups of citizen scientists.

APPENDIX 12. CHECKLIST FOR ENVIRONMENTAL DATA RESCUE

1. Search and locate historical datasets: paper, microfilm/microfiche or other old media and digital data in local, regional, national, international libraries and climate, hydrological and marine centres.

2. Organize, clean and store paper, microfilm/microfiche/other media and place them in safe, temperature- and humidity-controlled rooms in labelled archival boxes on shelves or in filing cabinets.

3. Create an electronic inventory of paper/microfilmed/otherwise stored holdings, including data on the station, year, month, media type, form type and box or file drawer location.

4. Create an electronic image inventory spreadsheet of what has been imaged and not imaged, validated and not validated.

5. Create a digital data electronic inventory spreadsheet of what has been digitized and not digitized, quality-controlled and not quality-controlled.

6. Prioritize the data to be imaged and the data to be digitized.

7. Review current digitization and quality-control steps to see how the new digitized data stream can fit into the NMHS's workflow and whether current digitizing and quality-control steps need revision.

8. Determine hardware, software, personnel, space, imaging, digitizing needs.

9. Develop a workplan for imaging and digitizing.

10. Obtain required equipment and hire personnel, if necessary.

11. Develop an imaging process that includes imaging and archiving the data using inventory spreadsheets and quality control of the images and metadata.

12. Image all copies of the data and metadata and archive the images appropriately.

13. Include the image inventory in directories and on digital storage devices.

14. Duplicate the image archive and store it in an off-site location.

15. Retain original paper/microfilm/other media copies.

16. Develop a digitizing process that includes template and database development, an updated inventory, and quality control of the data and metadata.

17. Review imaging and digitization processes as work progresses to determine if there are ways to improve them.

18. Cross-check the inventory of imaged and digitized data with the period of record of the stations to ensure completeness of the process.

19. On a daily basis, back up the digital data in the computer.

20. Update media, inventory and digital data inventories regularly with current data and metadata.

21. Create multiple copies of the digital data and place them in multiple locations.

22. Develop a technology migration plan and ensure regular migration of all electronic data (imaged and digital) to new storage media.

23. Generate products from quality-controlled data.

24. Make data available to users in a convenient, accessible form.

APPENDIX 13. GLOSSARY

ASCII. American Standard Code for Information Interchange, the most common format for text files in computers

Autographic charts. Charts recorded mostly by a clockwork drum automatically producing continuous measurements of environmental variables such as temperature, wind direction and speed, atmospheric pressure, and precipitation or hydrological variables such as water level (same as stripcharts)

CDMS. Climate Data Management System

Copy stand. Structure that holds cameras and lights in position for photographing documents

DARE. Data rescue, often used when referring to WMO activities in climate data rescue

Digital data. Data that have been keyed into a computer and stored as numbers. These numbers can either be in plain text forms or organized into a relational database such as Microsoft ACCESS or ORACLE or provided by a CDMS.

Digital storage devices. Digital storage solutions encompassing a variety of storage devices, including CDs/DVDs, hard disk drives, solid state drives, and so forth

Digitization. A process to transcribe analogue data into digital form for processing by a computer, usually done by keying text data or by tracing points from a chart

Electronic inventory. Spreadsheet with information on data archives

HDMS. Hydrometric Data Management System

Image. A picture form of the climate data document captured either by a digital camera or a scanner

Keying. Data entry or typing data into a computer via a keyboard

Media. Paper, microfilm, microfiche, CDs/DVDs, tapes, hard drives and so forth or computers upon which data are written

Metadata. A set of attributes or elements necessary to describe a resource. In data rescue, metadata refers to both station information and climate data inventory information.

Microfilm. A method of reproducing images of records at greatly reduced size on photographic film

Microfiche. Rectangular sheet form of microfilm

Migration. A means of overcoming technological obsolescence in hardware and software by moving data from one computer medium to another to enable the preservation of the intellectual content of the digital object

NMHS. National Meteorological and Hydrological Service

Old or outdated media. This phrase is used as a synonym for previously used storage media including microfilm, microfiche, magnetic tapes, diskettes, WORM (write once, read many), CD-ROMs, and so forth.

Preservation. Keeping the data safe from harm or loss

Reformatting. Copying information from one storage medium to another or converting from one file format to another

Refreshing. Copying information from one storage medium to the same or another storage medium without changing the format

Scan. To produce an image file from an analogue record or hard copy of a digital record with a scanner

Scanner. A device used to scan documents into a digital file format

Spreadsheet. A type of format used especially in developing inventories.

Stripcharts. Charts recorded mostly by a clockwork drum automatically producing continuous measurements of environmental variables such as temperature, wind direction and speed, atmospheric pressure, and precipitation or hydrological variables such as water level (same as autographic charts)

APPENDIX 14. REFERENCES

Cited references

Aguilar, E.; Auer, I.; Brunet, M. et al. *Guidelines on Climate Metadata and Homogenization* (WMO/TD-No. 1186). WCDMP-No. 53; World Meteorological Organization (WMO): Geneva, 2003.

Bojinski, S.; Verstraete, M.; Peterson, T. C. et al. The Concept of Essential Climate Variables in Support of Climate Research, Applications, and Policy. *Bulletin of the American Meteorological Society* **2014**, *95* (9), 1431–1443. https://doi.org/10.1175/BAMS-D-13-00047.1.

Brunet, M.; Brugnara, Y.; Noono, S. et al., 2020: *Best Practice Guidelines for Climate Data and Metadata Formatting, Quality Control and Submission of the Copernicus Climate Change Service Data Rescue Service*; Copernicus Climate Change Service (C3S), 2020. https://climate.copernicus.eu/sites/default/files/2021-05/C3S_DC3S311a_Lot1.3.4.2_2020_BestPracticeGuidelines_Part2.pdf.

Climatic Data Rescue (CLIMDRE): www.climdre.com.

de Smeth, K.; Comer, J.; Murphy, C. Hydrometric Data Rescue and Extension of River Flow Records: Method Development and Application to Catchments Modified by Arterial Drainage. *Geoscience Data Journal* **2024**, *11* (2), 176–196. https://doi.org/10.1002/gdj3.206.

Lorrey, A. M.; Pearce, P. R.; Allan, R. et al. Meteorological Data Rescue: Citizen Science Lessons Learned from Southern Weather Discovery. *PATTERNS* **2022**, *3* (6). https://doi.org/10.1016/j.patter.2022.100495.

Mateus, C.; Potito, A.; Curley, M. Engaging Secondary School Students in Climate Data Rescue through Service-learning Partnerships. *Weather* **2021**, *76* (4), 113–118. https://doi.org/10.1002/wea.3841.

Ryan, C.; Duffy, C.; Broderick, C. et al. Integrating Data Rescue into the Classroom. *Bulletin of the American Meteorological Society* **2018**, *99* (9), 1757–1764. https://doi.org/10.1175/BAMS-D-17-0147.1.

Tan, L. S.; Burton. S.; Crouthamel, R. et al. *Guidelines on Climate Data Rescue* (WMO/TD-No. 1210), WCDMP-No. 55; World Meteorological Organization (WMO): Geneva., 2004.

Westcott, N.; Andsager, K.; Stoecker, L. et al. *Quality Control of 19th Century Weather Data*. Contract Report 2011-02; Midwestern Regional Climate Center, Illinois State Water Survey Institute of Natural Resource Sustainability, University of Illinois at Urbana-Champaign: Champaign, Illinois, USA, 2011. https://core.ac.uk/download/pdf/4827106.pdf.

Wilkinson, C.; Brönnimann, S.; Jourdain, S. et al. *Best Practice Guidelines for Climate Data Rescue v1, of the Copernicus Climate Change Service Data Rescue Service*; Copernicus Climate Change Service (C3S), 2019. https://doi.org/10.24381/x9rn-mp92.

Wilkinson, C.; Freeman, E. *Best Practice Guidelines for Keying Data from Historic Marine Documents*; Copernicus Climate Change Service (C3S), 2021. https://doi.org/10.24381/68fy-2307.

World Meteorological Organization (WMO). *Guide to Hydrological Practices* (WMO-No. 168), Volume I. Geneva, 2008, updated in 2020.

World Meteorological Organization (WMO). *Climate Data Management System Specifications* (WMO-No. 1131). Geneva, 2014.

World Meteorological Organization (WMO). *Guidelines for Hydrological Data Rescue* (WMO-No. 1146). Geneva, 2014.

World Meteorological Organization (WMO). *Guidelines on Best Practices for Climate Data Rescue* (WMO-No. 1182). Geneva, 2016.

World Meteorological Organization (WMO). *A Proposed Long Term Resource Plan for Climate Data Management and Rescue Activities*; WMO: Geneva, 2017.

World Meteorological Organization (WMO). *Guidelines on Surface Station Data Quality Control and Quality Assurance for Climate Applications* (WMO-No. 1269). Geneva, 2021.

World Meteorological Organization (WMO). *Guide to Climatological Practices* (WMO-No. 100). Geneva, 2023.

Further reading

Brönnimann, S.; Brugnara, Y.; Allan, R. J. et al. A Roadmap to Climate Data Rescue Services. *Geoscience Data Journal* **2018**, *5* (1), 28–39. https://doi.org/10.1002/gdj3.56.

Brunet, M.; Jones, P. Data Rescue Initiatives: Bringing Historical Climate Data into the 21st Century. *Climate Research* **2011**, *47* (1), 29–40. https://doi.org/10.3354/cr00960.

Poli, P.; Dee, D. P.; Saunders, R. et al. Recent Advances in Satellite Data Rescue. *Bulletin of the American Meteorological Society* **2017**, *98* (7), 1471–1484. https://doi.org/10.1175/BAMS-D-15-00194.1.

World Meteorological Organization (WMO). *Guidelines on Analysis of Extremes in a Changing Climate in Support of Informed Decisions for Adaptation* (WMO/TD-No. 1500). WCDMP-No. 72. Geneva, 2009.

World Meteorological Organization (WMO). *Guide to Instruments and Methods of Observation* (WMO-No. 8), Volume I. Geneva, 2023.

For more information, please contact:

World Meteorological Organization

7 bis, avenue de la Paix – P.O. Box 2300 – CH 1211 Geneva 2 – Switzerland

Strategic Communications Office

Tel.: +41 (0) 22 730 83 14

Email: cpa@wmo.int

wmo.int

JN 24375