Best Practice Guidelines for Climate Data Rescue

Clive Wilkinson[1,2,5], Stefan Brönnimann[3], Sylvie Jourdain[4], Emeline Roucaute[4], Rick Crouthamel & IEDRO Team[5], Philip Brohan[6], Antonia Valente[7], Yuri Brugnara[3], Manola Brunet[2,8], Gilbert P. Compo[9,10]

Delivered 29 March 2019

Summary of Contents

This "Best Practice Guidelines for Climate Data Rescue" aims to facilitate planning and performing of data rescue projects. It covers all important aspects from the design of a project to digitization. A second report will then cover formatting, quality control, and metadata. It points to available tools such as data bases to check whether data have already been imaged or digitized and guides data rescuers through all aspects of work, from archive searching to imaging and digitization. A technical Chapter covers specific aspects such as camera settings, and an Appendix shows a concrete example of a data rescue project. A second Appendix on preserving physical sources provides information for record holders.

Best Practice Guidelines for Climate Data Rescue

Clive Wilkinson[1,2,5], Stefan Brönnimann[3], Sylvie Jourdain[4], Emeline Roucaute[4], Rick Crouthamel & IEDRO Team[5], Philip Brohan[6], Antonia Valente[7], Yuri Brugnara[3], Manola Brunet[2,8], Gilbert P. Compo[9,10]

[1] CSW Associates- Data Services UK, [2] Climatic Research Unit, University of East Anglia, [3] Oeschger Centre for Climate Change Research and Institute of Geography, University of Bern, [4] Météo-France, Direction de la Climatologie et des Service Climatiques, [5] International Environmental Data Rescue Organization (IEDRO), [6] Met Office Hadley Centre UK, [7] University of Lisbon, [8] Centre for Climate Change, University Rovira i Virgili, [9] University of Colorado/CIRES, [10] NOAA/Earth System Research Laboratory Physical Sciences Division

Table of Contents

1. Introduction

1.1. Scope

There is a growing need for weather and climate-related data to support science and decision making, ranging from adaptation planning and climate risk management to model evaluation and basic research. More and longer time series of measurements would ameliorate these needs, but such time series often exist only in hard copy form, perhaps at an unknown location. In recent decades, climate Data Rescue (DARE) to recover these measurements has seen an increase in interest and has become a community effort coordinated by the Atmospheric Circulation Reconstruction over the Earth initiative (ACRE, Allan et al., 2011a; Allan et al., 2011b).

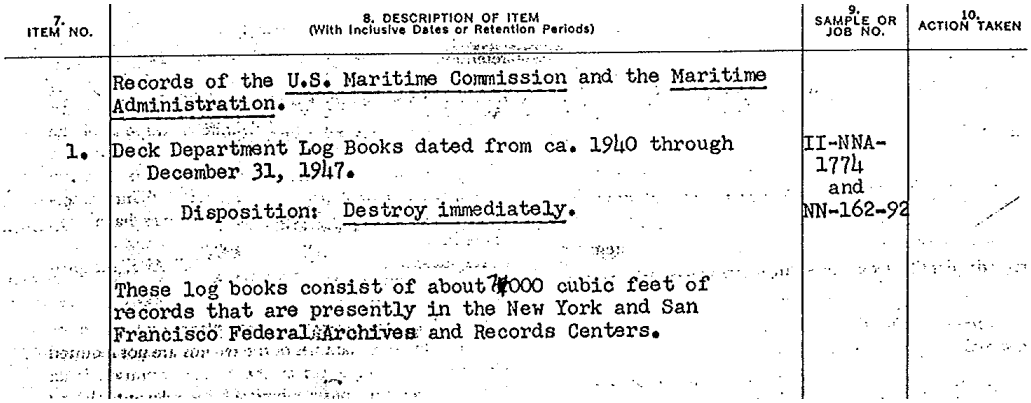
The actual data rescue work is performed by individual groups, projects, and initiatives ranging from small student projects to weather services' operations and citizen science initiatives. This Best Practice Report, which is part of Copernicus Climate Change Services (C3S) Data Rescue Services (DRS) (https://data-rescue.copernicus-climate.eu), is meant to facilitate such work. It is intended for leaders of small or big data rescue projects and guides them through the process of setting up a project, planning and performing archive work, imaging data sheets, and digitizing the data from the imaged sheets. It helps to avoid common pitfalls, allows more efficient planning of projects, and guides the user towards a coordination of their projects with the C3S efforts

The report complements and extends existing best practice reports such as the World Meteorological Organization (WMO) guidelines for best practices in the digitization of climate data (WMO, 2016) or the older paper by Brönnimann et al. (2006) on digitizing techniques. It also links to other resources provided by C3S Data Rescue Services.

1.2. Objectives of Climate Data rescue

The term "data rescue" is often used for the process of finding, imaging, digitizing (see Box 1 for a definition of these terms) and processing non-digitized data, in this case "climate" data (i.e., all weather and climate-related measurements and observations). All archives are unique and at risk of being damaged without any back up of their holdings.

There is past evidence of the wholesale destruction of valuable records containing climate data. On the marine side, the collection of English East India Company logbooks held at the British Library, is only a fraction of what once existed. In 1818 a decision by the Company Directors was taken to destroy the logbooks of chief and second mates, and that in future they would only be retained for twenty years. Apparently, there was also a mass burning of East India Company logbooks (pre-dating 1855), in Calcutta (Foster, 1919). The Chilean Navy disposed of many 19th century logbooks, retaining only those considered of historic value, a rational decision at the time, but one now to be regretted. A vast collection of merchant marine meteorological logbooks in the United States was also destroyed in 1974 (Fig. 1), being deemed of no historic value. It is all too easy to make a judgment about what material is relevant and important today, and thereby make an irreversible decision that will be regretted in the future.



7. ITEM NO.	8. DESCRIPTION OF ITEM (With Inclusive Dates or Retention Periods)	9. SAMPLE OR JOB NO.	10. ACTION TAKEN
	Records of the U.S. Maritime Commission and the Maritime Administration.		
1.	Deck Department Log Books dated from ca. 1940 through December 31, 1947. Disposition: Destroy immediately.	II-NNA-1774 and NN-162-92	
	These log books consist of about 7000 cubic feet of records that are presently in the New York and San Francisco Federal Archives and Records Centers.		

Fig. 1. One of a set of memos describing the destruction of World War II period ship logbooks in 1974.

In the case of meteorological observations taken at land stations, archives of data can be lost when institutions re-locate. Some developing countries have archives in a poor state of preservation or data stored in unsuitable environments (e.g., Fig. 2). Even in developed countries, paper records and other media can suffer from damp, fungal growth, or contamination from substances such as asbestos. The data-rescue community should be aware that even today valuable records are at risk of being lost from poor storage, neglect, ill-informed decision making, or being deliberately destroyed. This is an urgent problem that needs to be addressed by the data-rescue community working closely with the record holders.

Fig. 2. Contrasting storage facilities. On the left, Météo France, on the right, another archive somewhere in the southern hemisphere. (Photos Sylvie Jourdain, Clive Wilkinson & David Nash)

Fortunately, there are groups such as the International Environmental Data Rescue Organization (IEDRO) and others, working around the world to identify records at risk and undertake data rescue. IEDRO has charitable status. It relies on grants, volunteers, and goodwill to carry out this important work. Through the late 1990s and early 2000s, with funding from the US National Oceanic and Atmospheric Administration (NOAA), IEDRO was able to locate, image, and have digitized tens of thousands of upper-air rawinsonde and PIBAL observations from Kenya, Malawi, Mozambique, Niger, Senegal, Tanzania, and Zambia. These digitized measurements are archived at the NOAA National Centers for Environmental Information. Surface observations from these countries are still on paper and need rescue and digitization. Recently, a very large data rescue project was begun in Uzbekistan to image over 17,000,000 pages of hand-written meteorological surface information. The program was funded by the South Korea Meteorological Administration and administered by the WMO with technical assistance from IEDRO. The National Meteorological and Hydrologic Service of Uzbekistan (Uzhydromet) has already imaged over 4,000,000 pages.

Box 1: Terminology

Climate data: All weather, meteorological, oceanographic, cryospheric, terrestrial, hydrological, and climate-related measurements and observations. In a strict sense, measurements record weather conditions, from which climate can then be derived. In this report we use the term "climate data" as an umbrella term.

Weather record: A specific type of climate data in the form of consecutive weather observations or measurements. This is the typical object of data rescue.

Data Rescue: The entire process of archiving, searching, finding, imaging, digitizing, and converting non-digital data to a machine-readable format (note that rescuing digital media also is a problem but is not covered in this report).

Imaging: Making electronic images or scans of the hardcopy or microfilm data, typically into a JPEG format.

Digitizing: Convert the information on the hardcopy data or images to numbers of measurands. This can be achieved by optical character recognition, typing, or speech recognition in the case of written or printed numbers, or by special software in the case of charts.

Archive: [1.] A collection of original documents, fully documented and curated, or the act of creating such a collection. [2.] A generic term used to refer to a repository of historic documents, or the holder of such documents (e.g., The National Archives of the UK)

Repository: The physical location at which an archive is held.

Record holder: The institution responsible for the curation of an archive within its repository.

2. General Guidelines

2.1. Climate Data Sources

2.1.1. Types of Record Holders that archive weather and climate data

Almost any archive, museum, or library in the world has the potential to hold historical documents containing climate data. Material often appears in unlikely or unexpected places. Providing material has been catalogued and listed on a website, it is feasible to conduct detailed on-line searches on a global scale. The U.S. National Archives and Records Administration is one such institution. However, a large amount of material is not yet listed on-line, and many archives and museums hold relevant material that is inadequately catalogued or not catalogued at all. For example, in Switzerland the vast majority of archive holdings of weather-related documents prior to the start of the national weather service have only been catalogued recently (Pfister et al., 2019). Below is a list of the types of archives that will hold either weather and climate observations or original documents that contain such data. The list is not exhaustive.

- **National Meteorological Services (NMS)**

All such institutions will hold data either in digital form or in original paper form. Many will have a dedicated archive holding this material, but the quality of archiving and preservation will vary from country to country. Many NMSs will hold both terrestrial and marine data. Some NMSs have dedicated data centers where data from around the world is collected, for example the United States National Center for Environmental Information, (NCEI), previously known as the National Weather Records Center, in Asheville, North Carolina, is the world's largest active archive of climate data. It hosts the World Data Center (WDC) for Meteorology, Asheville and works in collaboration with the WDC for Meteorology, Obninsk and the WDC for Meteorology, Beijing.

- **Meteorological Institutes**

Meteorological institutes are research organizations distinct from a national meteorological service. They can often be associated with a university and will often have their own archive of original documents or data collected as part of their research activities.

- **Polar Institutes**

Any country active in polar research can be expected to have a polar institute and archive, usually holding station records, ship logbooks, ice core data, and other material. In some cases, this material is centralized in an institution such as The National Snow and Ice Data Center (NSIDC) in the United States, described as an information and referral center in support of polar and cryospheric research.

- **Oceanographic Institutes and Marine Fisheries**

Many countries are active in marine research and will have a dedicated national oceanographic institute as well as similar organizations attached to universities. These can range from departments dedicated to physical oceanography, to marine biology and to fisheries. They are likely to hold ship logbooks, cruise reports, and any other observations taken at sea as part of their research programs.

- **Astronomical, Magnetic, Geophysical, and Marine Observatories**

Records of observations will have been archived, if not at the observatory itself, then possibly at a national or regional archive, at a university, or with a national meteorological service.

- **Hydrographic Services**

Most countries with a naval service and/or merchant marine will have a hydrographic service to support those activities, providing charts and assistance with navigation and pilotage. Many of these services will have an archive containing original documents connected with marine surveying in which both meteorological and oceanographic observations will figure prominently.

- **Coast Guard and Marine Environmental Protection Agencies**

Most countries with an extensive coastline or marine interests will have a coast guard affording protection from smuggling, enforcement of environmental regulations, and to ensure safety for ocean-going users. These agencies might also administer a network of lighthouses. Ship logbooks and lighthouse registers are some of the documents that might be archived. Both can be expected to have meteorological and oceanographic observations. Environmental protection will also extend to terrestrial areas and land-based agencies may keep records of weather conditions.

- **Diplomatic and Consular Services**

These government departments have a long history of collecting weather-related data either on a formal basis or as part of an individual private initiative by a consular official. The British Foreign Office and the UK National Archives at Kew have an archive holding such material, although diaries and other forms of recording might still be found in the countries where the official served. Spanish archives holding sources connected with the administration of the viceroyalties of South and Central America also have weather records; some of these pre-date the general use of meteorological instruments and describe natural disasters and extreme events.

- **Other Government Agencies**

National park services, telecommunication and postal services, airport and aviation services, and others may hold archives of climate data material that has not been transferred to a national archive.

- **Medical Services**

Doctors frequently recorded climate data in the 19th century as they clearly saw that it was related to health, especially in the tropics. Some of these records can now be found in the archives of organizations devoted to the history of medicine and medical research, such as the Wellcome Trust in London.

- **Military Archives**

Military archives are usually separated into those for the army, navy, and air forces. In some countries such as the United Kingdom, the United States, Australia, New Zealand, and Chile, declassified material has been turned over to a national archive or to a museum. In the case of Chile, however, the Naval Museum is still part of the military establishment and the records of the army and air forces are not publicly accessible. Most naval and military records around the world still reside with their services. These records will contain a wealth of weather-related material, including meteorological observations. Access to such material can be problematic when it has been retained by the military service or has not been declassified. However, some naval and military services can and have been persuaded to cooperate with data rescue activities and release climate data. Contact needs to be made at a high level with individuals who understand the security risks of weather and climate extremes.

- **Museums**

Museums come in many types and forms, and a large number of them have libraries and archives where weather-related documents can be found. Typically, the types of museum to hold an archive of such material are naval and maritime museums, military museums, polar museums, and natural history museums. However, useful material can often be found in the most unlikely places, and no museum should be overlooked.

- **Botanical Gardens**

Botanical Gardens can often have a library or archive. The botanical gardens at Kew (London) and Singapore are typical examples.

- **Scientific Societies**

Most scientific societies will have records and some of these can relate to weather and climate. In the UK for instance, the archives of the Royal Society and the Royal Geographical Society hold a wealth of relevant documents including diaries, ship logbooks, and expedition records. There are similar societies in most countries with a history of scientific research.

- **Commercial Company Archives**

Some businesses routinely record meteorological or oceanographic observations as part of their activities. Commercial shipping, aviation, whaling, and fisheries are some of the most obvious examples. Some companies will keep an archive of their business history, although it is more common for these archives to be donated either to an appropriate museum, to a national or regional archive, or to a university.

- **National Archives and National Libraries**

Most developed countries have a national archive, and all will hold records of both terrestrial and marine climate data. However, catalogue descriptions will seldom mention weather observations, so it is essential to first determine the types of documents that will be relevant (see Section 2.1.2). It should be noted that many of the agencies and institutions listed above are likely to have contributed some if not all of their material to a national or regional archive. Likewise, a National Library can be an important holder of environmental data.

- **Regional and Local Archives**

State, regional, county, and city archives have been largely unexplored for weather-related data, but often hold significant quantities of useful material. Like national archives, however, catalogues seldom mention weather, meteorology, or oceanography. These regional and local archives are most likely to hold the business records of local companies. For example, the Vestfold Archive in Sandefjord, Norway, holds the records of many Norwegian whaling companies, with the most extensive collection of Antarctic whaling logs in the world. Local archives are also most likely to have historical documents containing measurements from private observers, such as doctors and naturalists.

- **University Libraries and Archives**

Almost all universities will have an archive, sometimes related to the locality and its history or to the university's research work. In the latter context, individual university departments may also keep an archive of material. Universities will sometimes also hold business and company records. Two typical examples are the University of Edinburgh holding the business records of the Christian Salvesen Company, involved in shipping and Antarctic whaling, and the University of St Andrews holding the archive of the Sea Mammal Research Unit. The latter archive has extensive material relating to sea-ice and meteorology in the Antarctic regions.

- **Electronic Collections**

Many institutions have electronic collections, or publications in digital form. Some can be scanned library holdings such as those found on Internet Archive (https://archive.org), JSTOR (https://www.jstor.org), Google Books (https://books.google.com), the collections of Foreign Climate Data held by the NOAA Central Library (https://library.noaa.gov/Collections/Digital-Documents/Foreign-Climate-Data-Home), or the French National Library-Gallica (https://gallica.bnf.fr/). The digital collections of the Munich Digitisation Centre (https://www.digitale-sammlungen.de/), the Bavarian State Library (https://opacplus.bsb-muenchen.de/), the Presidential Library, St. Petersburg (https://www.prlib.ru/), of Swiss Journals at ETH Zurich (https://www.e-periodica.ch/), or of the Berlin state library (https://digital.staatsbibliothek-berlin.de/) also contain climate data.

2.1.2. Types of documents that contain weather and climate data (and metadata)

- **Manuscript Sources - Terrestrial**
  - Meteorological registers (bound)
  - Meteorological forms, sometimes called a Landform (single sheets)
  - Metadata station forms
  - Rough or draft notebooks
  - Barograms
  - Thermograms
  - Tephigrams
  - Pluviograms (rain gauge charts)
  - 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
  - Pyroheliometer record
  - Meteorological collections (e.g., Monthly Bulletins, Yearbook Reports, Monographs, Observatories' reports)

- **Manuscript Sources – Marine**
  - Ship logbooks
  - Marine meteorological logbooks (also called a weather book, meteorological register, or a meteorological journal, and often including a metadata form)
  - Marine meteorological forms
  - Abstract logbook
  - Register 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 (International Whaling Commission and its predecessors)
- Marine insurance records
- Hydrographic and surveying reports
- Dredging reports
- Tidal records
- Ocean current reports (i.e., United States Hydrographic Office H-19 current report)
- Typhoon and storm reports
- Ice reports
- Ice charts
- Marine radio weather logs and reports
- Harbor master records
- Ships' newspapers

- **Printed Sources**

There is a wealth of climate data to be found in print. Sources range from expedition reports to annual meteorological summaries, articles in scientific and geographical 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; footnotes and bibliographies can often be enlightening. Once located, printed sources are relatively easy to transcribe or can be read electronically. Their chief importance lies in the fact that printed sources can sometimes be the only surviving historical record where the original documents are lost or difficult to locate. Printed sources are however no substitute for the original documents because they are frequently edited versions of the original material. A recent comparison of the published *Report on the Scientific Results of the Voyage of HMS Challenger 1873-1876*, with *Challenger*'s meteorological logbook (held at the UK National Meteorological Archive, Exeter) revealed that detailed ice descriptions had been reduced to the word 'ice' in the published report. See also Electronic Collections above.

- **Strip Charts**

There exist several million analog strip charts detailing many different hydrometeorological parameters such as precipitation (pluviograms), atmospheric pressure (barograms), temperature (thermograms) as well as pyroheliometer records (sunshine), wind speed, relative humidity, etc. While containing very detailed parameter values, most of these paper-based records lie stored on shelves and in boxes since it is very difficult to manually extract the values from these records and place them into a digital data base. In addition to this difficultly, there is possibly another important reason why these records have been largely ignored. Since these parameter values at extremely short time intervals (i.e., minutes) have heretofore not been readily available in digitized format, scientific applications using these high frequency values have not been developed. Our belief is that once pluviogram data are available in digitized form, applications will be developed such as flash flood and mud slide warning models. Other applications may include their use in designing road culverts and storm runoff engineering situations. Barograms depict a continuous record of atmospheric pressure changes could be used in examining, e.g., gravity waves. Pyroheliometer records may contain valuable information for designing solar power arrays, while these strip-chart wind observations could be used in designing and locating wind turbines for power generation.

2.2. Searching and Locating

2.2.1. Checking existing inventories

DARE projects are costly and labor intensive. Duplication of work should be avoided. The important first step of any DARE project is to check whether the climate records of interest have already been inventoried. The inventory then might tell you whether the data have already been imaged or digitized, or it might contain information on the location of the records. The most comprehensive source of such information is the C3S Data Rescue Registry (https://data-rescue.copernicus-climate.eu/projects). This registry provides access to metadata on a series-by-series basis. The registry is searchable and provides detailed metainformation on stations, variables, and the status of data rescue, i.e., whether or not a series has been imaged or digitized. For early instrumental (pre-1850 generally, pre-1890 for Africa and the Arctic) land station data, a recent inventory provides detailed information such as observer names, locations, years, sources, and whether or not the data have been imaged or digitized (Brönnimann et al. in preparation).

If the data of interest cannot be found in these inventories, it is advisable to check the C3S Data Rescue Activity Database (https://data-rescue.copernicus-climate.eu/projects). This database holds information on ongoing projects, i.e., information that may not have made it to the inventories yet. It includes status and contact information for each project. Additionally, C3S publishes annual "State of Data Rescue" assessments based on literature research and questionnaires (Brönnimann et al., 2018). These assessments may also contain useful information on the level of projects.

If the data in question have already been imaged but not digitized, continue with Sect. 2.5. If the inventory lists the source, continue with Sect. 2.4. In any case it is important to update the inventory so that others will not duplicate your work. Also, if the data have not yet been digitized, early in the project it is important to register with the Copernicus Climate Change Services Data Rescue Activities database so that your project will be visible to the community.

For data from the period of NMSs, the first step is to contact the corresponding NMS and check their inventories. Until recently, historical weather and climate records were seldom imaged before they were digitized, usually by the NMS themselves. However, as DARE is more widely known, the imaging of these historic records is now the rule rather than the exception. Also, many countries' judicial systems will not allow digitized records to be entered into evidence without a link to the original observer's record, either the original paper-based or microfilm/microfiche. Without this link, no matter what QC process was involved, there could never be a 100% certification of the digitization accuracy.

This brings up a conundrum which has plagued all DARE efforts since the process began. What should be done if digitized observations for a specific observation site, date, and time are discovered in a WDC, such as the NOAA NCEI, but there is no image of the original data sheets? If the DARE Project has the opportunity to image the original documents, should the data be re-digitized since there would then be a validation of the newly digitized data?

Bearing the above in mind, it is clear that without a global registry of imaged documents, checking to see if a document has been imaged and/or keyed is difficult, and sometimes impossible. But it is not just NMSs that undertake or commission the imaging of documents. The record holders themselves often have active digital programs, and have images available on-line, and it is worth checking at the outset, if any of the records of interest have been imaged. Note however that the term 'digitizing' can mean either imaging or keying or both so it is important that clarity is sought on this point. In rare instances, a record holder may wish to charge you for the images that they hold. In this case it may be more economical to re-image the documents. This brings up a valid point. It is strongly recommended that there should be a convention to distinguish between providing digital images or providing digital data. At present (2019) no such convention has been agreed upon. Perhaps we can use the term "electronic images" for digital images and "keyed data" instead of digitized data (even though in some instances the data is generated by a computer program (i.e., Weather Wizards or Optical Character Recognition (OCR)) and not "keyed" by a typist. This is important as miscommunication could result in project delays.

2.2.2. Methodology for record searches

Locating material held in the archives of NMSs is usually straight forward when the archive is well organized. Records can be found by type, station, or in the case of marine records, the name of the ship or other platform. Many NMS archives are well organized, but others less so, depending on location and attitude to the preservation of historic records. Do not depend on an NMS having an on-line digital catalogue, although many of the well-established NMSs do. Others will often have either hard-copy inventories, or electronic copies of inventories.

Searches in museums, national and regional repositories, and other record holders described in section 2.1.1 require a different approach because the records they hold are more varied and diverse, and seldom specifically catalogued or labeled as environmental. It is necessary first to determine a specific aspect of what you are looking for. This could be a type of record, a location, the name of a ship, the name of an observer, etc. Create a list or inventory, such as a list of vessels in polar regions or voyaging to specific places that are of interest, e.g., a list might include ports in South America, India, or China. Or a list could include known meteorological observers or some other search criteria. The compilation of such inventories is useful even if records are not immediately located as they might be found at a later date.

On-line searching is the quickest and most obvious route to locate material, but the document collections of some record holders will not be found on-line for a variety of reasons. There may for instance, be a digital catalogue that is only accessible on-site. When a record holder does not have a digital catalogue, there will usually be card catalogues or registers. Searching card indices requires specific terms such as names or institutions. Using general terms such as 'meteorology' will prove disappointing. When a record holder has no on-line search facility, direct contact with the holder's staff is recommended, as they will sometimes be able to produce a list or inventory of their relevant holdings or advise you on material that may be useful.

When an on-line search is available, the choice of on-line search terms is critical, and some imagination needs to be exercised in second-guessing the tags and terms used by archivists when producing the catalogue. General searches need to use a large number of potential search terms. Results are almost always productive. It is a good strategy to compile a list of search terms relating to meteorology, climatology, document names, and regions and places, and apply these systematically to holdings in different archives. Terms entered in the search engine of an archive in a non-English speaking country need to be in the language of that country. This may seem obvious but can be a stumbling block to native English speakers and others accustomed to global searches in English. For instance, the search term 'ships logbook' will produce poor or nil results, whereas 'journal de bord' (France), skips dagbok (Norway), bitacora del mar (Chile), etc. will produce positive results in archives of those countries. Similarly, general meteorological and other search terms mentioned above need to be translated for non-English on-line searches. Note that early scientific observations may have been recorded with documentation in such languages as German, French, Japanese, or Latin. Early colonial records might be found in French, Spanish, or Portuguese. A list of search terms and their translations will be collected and continuously built-up at the C3S Website.

It should be noted that museums and other record holders are always receiving new collections but often have limited resources for cataloguing. Therefore, it can be expected that all of these institutions will hold uncatalogued material. Collections of tabulated data of interest to the rescuer of weather and climate observations might be considered low priority where resources for cataloguing are limited. In addition, collections are sometimes transferred between archives, for instance an NMS might transfer records to a national archive. When this happens collections can apparently disappear, sometimes for a period of years, and in some instances languish in a storeroom, unnoticed and forgotten. If you are aware of a significant collection having been sent to an institution that denies having received it, it is essential that you insist that its location be found. The original institution can often provide evidence of the transfer. Be aware, however, that an archive may deny having a collection because it is uncatalogued, and producing it for research is inconvenient. This can be overcome by displaying patience and understanding, and by developing a close working relationship with the staff of the, museum or other record holder. It is also well worth repeating on-line searches from time to time as you will discover that new material has been recently catalogued and new sets of collections are now represented on-line.

2.2.3. Advice for working with record holders

Experience has shown that record holders, and museums in particular, are very interested in the work being undertaken to rescue historical weather, climate, and other environmental data. They not only understand how important this work is but also see it as an opportunity to justify the significance of their collections to the funding bodies that support them. We should not underestimate the importance of this. They are therefore often enthusiastic to engage with the science community and to help support the work of data rescue. There will be exceptions however, so all approaches should stress the mutual benefits of working in cooperation.

From the outset, it is well worth the effort to build a good working relationship with record holder's staff. This is generally easier in smaller repositories and museums, less so with larger national archives, where public traffic is much higher and a more personal service to the rescuer is often difficult. Nevertheless, there are some basic guidelines that can be followed in all circumstances.

When first communicating with a record holder you should say at the outset that you will wish to photograph some of their documents. You should also be clear about the intentions you have with these images. It is important to inquire about the archiving policy, and about the rules concerning consultation and reproduction at the repository. Give them full information on the project and your work, with links to websites. Usually, the bigger and more professional the project looks, the more interest a record holder will take in it. At this stage you should already have a list of documents for imaging and this should be sent to the record holder.

In the first communication, it is important to state clearly that you intend to offer copies of the images back to the record holder. Many archives will welcome having their collections digitally imaged at no cost to themselves. However, some archives have been known to refuse, or to demand substantial payment for the privilege of imaging their archive. Other archives have been known to charge a substantial amount to one agency, and then to charge another agency for the same work. Good communication between data rescue practitioners should help to avoid these situations, but extreme caution should be exercised if it appears that an archive or museum is trying to profit from data rescue activities

Some archives may express a concern about how the images will be used and may place restrictions on having them published or put on-line (e.g., they may require images to be non-downloadable), or they may require to use their URL (for a digitally available document) rather than duplicating the image. It is worthwhile discussing any potential restrictions on image use with the archive as a matter of courtesy and to establish trust, even though many archives will place few, if any, restrictions on the use of the images of their material. A notable exception is military archives. For example, the naval archives of Argentina and Chile, where ship logbooks have recently been imaged, have had the remarks sections masked or redacted, leaving only the environmental and positional data visible. Such security concerns should always be respected and discussed fully when using military archives.

Another offer that can be made to an archive or museum is to give a presentation to the staff, or the public, on how the data in their collections will be used by climate scientists. Such offers are always appreciated and sometimes very well attended. Once you find that archive staff are more aware of what you are doing and what your requirements are, they will often look for, or produce material, that you might not have been aware of.

Before starting work in an archive, especially at the imaging stage, it is essential to organize the work beforehand. There are several practical but impactful issues that need to be addressed. You should establish the most effective way of having the documents produced and where you will work. Some smaller archives have very limited space for readers, requiring you to book or reserve a place in advance. Work is likely to be undertaken in a public area, and your activities should not disturb other users (e.g., cameras should be silenced). Discuss with the archive staff in advance, what is allowed to be used, for example a tripod, and the lighting conditions where you will work with the documents. Lighting conditions are important for choosing the right equipment and location to undertake the work. Always ask for a separate room to work, if this is possible.

Ideally a large number of documents on a cart can be delivered, and in small repositories and museums this is often the case, but in larger national archives with more formal procedures, there can be restrictions on the number of documents issued at any one time. In some repositories, such as the UK National Archive at Kew, London, it is possible to bulk-order in advance, but otherwise where order restrictions are adhered to, additional orders for documents need to be placed while work is in hand, so as to avoid a long wait for additional documents to be procured. Rarely, a repository will insist on producing documents one at a time. This severely impacts the rate of generating images, so it is useful to know this in advance and plan accordingly. This is why it is essential not to just turn up at a repository to work, but to organize the work beforehand.

Once you have established a good working relationship with the record holder, you should discuss with them the subject of uncatalogued material. All repositories and museums have uncatalogued collections for the reasons already mentioned above. You might be permitted to look at these items and possibly even image them, although usually uncatalogued material will not be released. You should however be permitted to document the collection in a superficial form for future reference. You should also ask the archivists if they are aware of other relevant material in the collections of other record holders. Many NMSs, repositories, and museums are connected through a professional network. They can help you find other data collections, provide contact details and various other forms of assistance. Experience has shown that significant new data sources can be uncovered through this type of assistance. Therefore, you should always work *with* an archive not just *in* an archive.

The advice outlined above assumes that the record holder is co-operative, if not positively enthusiastic to work with the DARE community. Sadly, this is not always the case. Record holders may not be archivists at all but the administrative head of the facility like the Director of the NMS. Occasionally, these officials are not scientists and have limited scientific training. They may not know why their organization has held on to these old data nor appreciate their current value. It is important that such officials are provided with information so that they understand the data in its present form is practically worthless and must be imaged and then digitized before any value will be derived. Some record holders believe their data is extremely valuable in its present form and are reluctant to allow its imaging or digitizing without payment of substantial funds. In these instances, we must justify the cost in time and funding required to transform these paper/microfiche/microfilm data into useful electronic formats. We also need to ensure that the record holder realizes that if they allow the imaging and digitizing of their data, some WDCs, such as hosted by NCEI, will be willing to archive the images and digitized parameter values with restrictions as to who will have access and under what conditions.

2.3. Inventorying Physical Data Sources

It is important to be aware of the logic by which items are arranged in an archive. Weather records are not understood on their own as individual items. By maintaining the original order, custodians can protect the authenticity of the records and provide essential information as to how they were created, kept, and used. Weather records are often arranged by provenance, by type of material, by territory, by form, by station, or by ship (see also Appendix 2 on preserving physical sources).

There are several types of inventories/catalogues depending on the material and the objectives. Archivists produce finding aids containing the inventory of the series and the folders. These finding aids include a wide range of formats, including card catalogues, calendars, guides, inventories, shelf and container lists, and registers.

The data elements essential to finding aids are defined by the International Council on Archives in the General International Standard Archival Description ISAD(G). Many finding aids are encoded (marked up) in XML; in such cases, the Encoded Archival Description standard can be used.

The guidelines on Best practices for Climate data Rescue (WMO 2016) recommend to develop an electronic inventory of material including station, year and box/file drawer location, medium type, form type including variable types and missing periods. When developing this inventory, it is important to keep in mind that this information should eventually constitute an update of the existing inventories (see Sect. 2.2.1), so that others will be able to work with your information.

2.4. Imaging Documents – General Guidelines

2.4.1. Planning an imaging project

- **Small Projects**

We assume here that a small project will involve only one or two persons and no more than three. Small projects will usually image a selective group of documents based on specific data requirements such as data needed for a region or time period, or a specific data type (e.g., upper air observations or barometric pressure), or particular type of platform. This sort of work is often outsourced to an expert individual or can be performed in-house by an NMS. In the case of an NMS organizing a small/medium sized project within its own records, there may be more than three persons involved, but the same advice will apply. The plan below is an outline of the tasks that need to be carried out. Where tasks are in italic, these will be of particular use to NMS expert staff.

- Archive searches [See Section 2.2]
- Organizing a pre-imaging inventory either for an individual archive or group of archives. It will save time if the pre-imaging inventory is based on the format of the image inventory that you intend to produce (i.e., an inventory that describes the contents of folders or of images with multiple pages; see below).
- Seek permission from the archive(s) to image the documents
- Produce a full cost estimate based on the amount of work indicated by the pre-imaging inventory (estimates are discussed separately)
- Once funding is secured, contact the data holder to arrange a visit and discuss working arrangements. In particular you should find out whether they already have camera stands or if you will need to use your own tripod or other stand. In small repositories, space can be restricted. In some instances a workspace needs to be booked, so these arrangements should never be left until the last minute. Some archives, (usually the larger national archives) will require an identification document and proof of address and possibly a letter from your institution. In certain countries such as the United Kingdom, archives connected to the military establishment will require low-level security clearances. All of these issues need to be handled well in advance of your visit.
- Think of a format for collecting metadata (see upcoming Best Practice Guidelines on Formatting, Quality Assurance, and Metadata)
- Before a document is imaged, its suitability should be checked. This takes only a few moments, and avoids spending time imaging a document that cannot be used. Sometimes there are duplicate documents, and if this is noticed then only the better of the two needs to be imaged. Beware, however, that a copy of an original might appear neater but may be edited and not faithful to the original. If in doubt, image both. If a document is not imaged because it is unsuitable, the reason for not imaging should be noted in the image Inventory. This will prevent future projects from thinking that the material had been overlooked.
- After each day's imaging, the images produced must be downloaded from the camera, and a backup produced. If more than one person is producing images, the downloads must be kept separate to avoid images from different documents being intermixed.
- Images will need to be sorted and organized into folders corresponding with the document archive reference and any other pertinent information that allows you to easily identify a group of images. This task can either be performed daily or after the imaging process is completed for each week. **Do not** wait until the entire imaging process is completed because the sorting of images is also an opportunity to check image quality and to arrange to re-image a document if required. Once the images are sorted, a backup must be produced.
- The images are delivered to the NMS or other commissioning institution.
- An image inventory or inventory should be compiled and can be based on the pre-imaging inventory. The image inventory will describe the contents of each folder, for instance a ship logbook, or each image if, for example, you are imaging meteorological forms (1 to 4 pages) and need to know precisely the date and the station. See WMO guidelines (WMO 2016).

- The image inventory must include the archive reference, a brief description of the document, the type and frequency of data recorded, any metadata concerning instruments and their exposure, the location of the observations or in the case of a marine platform, its movements, and any other relevant or unusual observations, such as severe weather etc. Land and marine inventory templates are available.
- It is recommended that after an imaging project is completed, a report should be produced. The report should include details of how, why, and where the imaging was undertaken. It should include descriptions of the format of some of the more important documents and be illustrated with example images. It should draw attention to any notable or unusual observations and any potential problems that might be encountered when either transcribing the contents of the images, or processing the data. General notes about metadata, especially non-English, or antiquated scales or terms are useful even if attention is only drawn to them. Most importantly, however, the report should outline precisely what work was carried out, and also what work was not undertaken, so that any future project can be confidently organized.

- **Large Projects**

Planning for large projects is very similar to small projects. The differences are mainly in the scale of human labor and financial costs. Small projects usually have a finite set of data that needs to be imaged and keyed. A large project can focus on a region or country and attempt to consolidate all the historical observations from a variety of "data owners". This could be the NMSs, museums, universities, agricultural organizations, etc. Many times, the data owners themselves do not realize that others in their region have additional observations. Gaining the cooperation of all data owners can lead to a more successful project. Also, if the data owners know others in their area also have data, there is more of a chance they will share theirs rather than considering their data as proprietary.

2.4.2. Scanning or photography?

Currently, a scanner produces a better-quality image than a camera. That is virtually its only advantage over conventional photography. The chief disadvantage to scanning is that the process is slow, meaning far fewer images will be produced per day, with a resultant increase in staffing costs. Scanning is also not suitable for certain types of documents, especially bound volumes, and is often not permitted. A photographic station (see Sec. 3.2) using a camera stand or a tripod (+ boom) that allows photography in a vertical axis, can produce many more images than a scanner. Depending on the condition of the documents, speed of presentation, size, format binding, and other handling issues, the number of images produced in six to seven hours can range from 800 to 1500 images. This includes a brief pre-examination of the material to be imaged. In many cases, the photographic image will be nearly as good as the scanned images, but there may be a slight fall-off of image quality towards the edges of the document. It is also easier to image bound documents, and as the documents are not pressed flat as they would be in a scanner, there is no risk of damaging or straining any binding. Image files produced by scanners are also very large (a TIFF file), resulting in storage issues. In the context of partnerships with data holders, TIFF files and JPEG files are usually requested.

2.4.3. Outsourcing?

Why outsource document imaging? The decision about in-house or outsourcing depends on the project: its duration, number of images to be scanned, staff required and staff available, financial resources, etc.

Imaging is a very labor-intensive process. Using an external agency to image allows you the advantage of specialized equipment that may be too expensive for you to own to get the work done by staff who have expertise in imaging. The skill of the operator (scan and quality-check) will have a strong impact on the quality and consistency of the images, especially if digital cameras are necessary for the imaging project. Imaging agencies use professional quality checked software. Some imaging agencies use scanners with a book scanner system adapted to thick bounded books. Outsourcing imaging allows you and your local team to concentrate on project

management and the selection and organization of the documents to be imaged. The outsourcing work can be done in the archive or outside at the private agency.

It is easier to organize outsourcing if you are the record owner, because it is necessary to prepare the documents (pulling out staples, repairing damaged pages, file reassembly, and re-filling). It is sometimes possible to organize outsourcing in the context of partnership with record holders. In this context, the latter often want to prepare the documents, give technical specifications about image technical characteristics, file renaming, and image file inventory spreadsheets, and in this case, outsourcing is the best way.

Usually, you have to develop and to provide a pre-image file inventory spreadsheet to the private agency. You can ask for a specific renaming of image files. You will receive renamed image files and the completed image file inventory spreadsheet.

Advices for outsourcing: There are plenty of digitization/imaging agencies. Prices can be very different according to the countries and the agencies. You have to be certain that the private agency you choose has experience doing the work. It is useful to seek the advice of local or national repositories, and to compare quotes of several private agencies. You need to check references and ask for equipment and quality-check procedures. You can save a significant amount of time and money by properly examining and preparing documents beforehand.

2.4.4. Estimating costs for an imaging project

- **Small Projects**

Estimating the costs of a small imaging project should take place once an inventory of material has been produced, and data holders have consented to have their documents imaged. If the types of documents to be imaged are familiar and have been imaged before, it is possible to make a reasonably accurate estimate of the number of images that can be expected. The number of images that can be produced in a day depends on how the documents are presented (e.g., in bulk or individually) and other handling issues connected with the size, format, and condition of the documents. For instance, individual documents such as meteorological record forms will take longer, per image, than bound documents. Very old or delicate documents require additional care in handling and unpacking, if boxed.

As noted above, you can expect to take 800-1500 images and, on average, 1200 images per person per day. If you have calculated the number of expected images from your pre-image inventory, then it is possible to determine the number of days the project will require for the imaging stage. The number of images taken per day will depend on a number of document-related factors, from the production of the material by the archive to the size, format, and condition of the documents themselves. Different working methods such as an individual working alone or two people working together, handling, inspecting and photographing, needs to be considered. Also photographing double pages or single pages will impact considerably the number of images produced per person per day. This is discussed further below.

Very often however, it is not possible to determine how many images will be needed to complete the task, and it is therefore necessary to base the number of days spent on imaging on a best guess. Fortunately, such guesswork is not often required when a fixed sum or budget is allocated for the work. Cost estimates can then be based on the expected production of 1200 images per day per person (modified by the document-related factors), and fitted into the pre-arranged budget allocated for the work.

Time and labor costs are not just confined to the imaging part of the project. The images will need to be inventoried. If five *person days* are spent imaging, at least ten person days and possibly more, will be needed for sorting and inventorying the images. If more than 1200 images are produced per person per day, then the time and cost of inventorying will be higher, so high productivity on the imaging side has cost implications on the inventorying part of the project. Nevertheless, high productivity should never be discouraged as long as the quality of the images is not compromised. As a general rule, double the number of person days spent imaging to calculate the number of person days needed to produce the inventory. An additional three or four days should be

added to write a report. If notes on the documents are kept during the imaging process, the report can be partly written up from these, and four days should be adequate for most small projects.

The imaging project cost template (Table 1) assumes that photographic and other equipment is already provided for. It also assumes that the necessary pre-imaging archive searches have already been performed and have been funded separately. It is strongly suggested that costs are re-examined and revised if funding is not forthcoming after three or four months have elapsed from the original estimate, as large seasonal price increases can affect travel and accommodation costs. The project should ensure that when for instance employing a student assistant or other person, all employment laws and regulations are complied with and any attendant costs accounted for.

- **Large Projects**

There are many similar considerations in estimating the costs of large projects as those in small projects. Again, the main difference is in labor costs and equipment, travel, and training.

With large projects, one initial requirement is that the data owner(s) consolidate their data to be rescued and digitized in one secure place. Once that is accomplished, a complete site survey usually involving at least two DARE experts working for 1 to 2 weeks at the data site is required. These experts evaluate the type, condition, and amount of records to be rescued and digitized, as well as the suitability of the facility (power, heat/air conditioning, environmental protections, security, furniture, space, etc.) to handle the DARE work. Additionally, the team examines the number and capabilities of the personnel who will be actively performing the DARE on-site work as well as the management of the staff. The team, with the help of the data owner, will identify a DARE on-site focal point who has the responsibility, and authority, to manage the project from the beginning to the conclusion. These site survey costs can be substantial considering airfare, accommodation and (unless volunteers are used) salaries. However, a site survey is critical for the project to have any chance for ultimate success. It is far better to expend a few thousand dollars on a site survey to find that the project is not viable than wasting hundreds of thousands of dollars when the program fails.

Table 1. Estimating the cost of an imaging project

Item	**Explanation**
Labor cost – person 1 (Lead)	An agreed daily rate that should include a reduced rate to cover non-working days spent on travel.
Labor cost – person 2	An agreed daily rate that should include a reduced rate to cover non-working days spent on travel.
Labor cost - person 3	An agreed daily rate that should include a reduced rate to cover non-working days spent on travel.
Expenses	
Travel, Air, Rail, Bus Taxi	All travel expenses
Insurance (if required)	Travel insurance, or general insurance if for instance a student is employed to assist.
Accommodation	Hotel expenses
Subsistence	Food and refreshments. An allowance of GB£ 30-35 per person per day is adequate. (2018)
Fees	Some archives charge a token fee per person to perform photography
Sundries	For example the cost of an external hard drive to store to deliver the images
Postage	Cost of posting the images to the commissioning institution
Expenses- sub total	
Contingencies @ 15% of expenses (sub-total)	This should be included to account for unexpected expenses, unforeseen price increases, and foreign currency fluctuations.
Expenses - Total	
Labor and Expenses - Total	

2.4.5. Equipment required

- **Small Projects**

Each person should be allocated the following equipment. If the choice is to use photography, the following equipment is required:

- Laptop computer
- Camera
- Remote switch
- Spare batteries and charger
- Tripod with any attendant accessories
- Notebook and pencils
- Spare data cards

In addition, the lead person should have at least one and maybe two external hard drives to store and back-up the images and an additional camera to cover equipment failure.

- **Large Projects**

In small projects, many times it is assumed that the labor would be accomplished by DARE experts rather than the data owner's staff. On Large Projects, the equipment is slightly different. Large projects need several DARE work stations. These include:

- **Computers** - office computers are often preferred over laptops as laptops can more easily be removed by unauthorized persons. The office computers are more likely to remain operational since they are seldom moved once the project begins. In some instances they are less expensive than laptops.
- **Digital cameras** – since these are the central focus of imaging, the best quality is most desirable. It is important to look at their longevity even more than their ability to record at super-high resolution which is not absolutely necessary for most data rescue needs. Also, it is important to have spare cameras available for the workstation to ensure the imaging process isn't held up by the failure of one.
- **Remote switches –** Although remote switches are preferred over hand activation of the cameras, for large projects it is preferred that the digital cameras be hooked up directly with the computer allowing the computer software to activate the camera.
- **Spare batteries and charger –** these are acceptable for small projects but for large projects that involve imaging tens of thousands of pages, the frequent battery replacement almost always results in the camera mounting threads or battery compartment covers to be damaged beyond repair. Therefore, cameras relying on an A/C power supply are preferred.
- **Tripod –** in large projects a camera stand works significantly better than a tripod. Once set up, a fixed camera stand can be configured to always hold the sheet of paper being photographed in the exact same place from sheet to sheet which will speed up and enhance the quality of the images. Additionally, if the data being imaged is in a bound volume, an apparatus such as a "Book Liberator" (see Appendix 1) should be used. The Book Liberator enables two pages of a bound volume to be recorded simultaneously without having to separate the pages from the binding. Also, devices similar to the Book Liberator enable the pages in the bound volume to be held flat enabling good resolution of the data close to the binding.
- **Notebook and pencils -** are fine for small projects but having a separate computer dedicated for the data rescue focal point to maintain an inventory of images taken and other records such as imaging technician work records, equipment failures, and replacements, etc. is important.
- **Storage –** this is one of the most critical parts of the imaging process, and the most expensive. Storage of a few hundred thousand images can be handled by a few portable hard drives but care should be made to use the highest quality available, not those which retail for less than USD$100. Inexpensive drives are prone to failure in time. However, large projects dealing with millions of images require substantial storage devices. These can cost over USA$40,000. Redundant storage devices are needed so that imaging

can also be stored at a physical location distant from the main storage device. Also, portable devices are acceptable for temporary storage but long-term storage requires a proven IT configuration that is often beyond the ability of the local data owner (i.e., NMS). In such cases, we recommend the data owner reach an agreement with an organization, such as the country's military to provide storage and retrieval of their imaged and digitized data. This is usually possible if there are good working arrangements between the NMSs and the military. All military organizations are very interested in weather data for strategic purposes and would regret if the data held at the NMS was lost. Many military organizations have far better computer facilities than any other government agencies. Other possible data backup archives could be the country's Ministry of Health/Medicine, Natural Resources, etc.

2.4.6. Handling manuscript material for imaging[*]

Historical documents are irreplaceable, and must be treated with care in all instances. Repositories will have firm rules about the handling of documents. You should never use an ink pen in an archive, lean on a document when writing, or take any other action that would compromise the integrity of the document. In particular you should

- Make yourself familiar with the repository's rules on document handling and comply with them.
- Not attempt to handle or image documents (especially bound documents) that are too fragile, are damp or have fungal growth, or are otherwise in need of conservation. Such items should be reported to the archivist who may not be aware of their condition.
- Never bend or strain a closely bound document in order to image that "hard to get data" right in the middle. The document is more important than the data, although if you have sufficient time it might be possible to manually transcribe the data that cannot be picked up by imaging.

2.4.7. Cameras and imaging - general advice[†]

- The cardinal rule is *never* to take images using flash photography. This is the quickest way to get yourself removed from the building without a fire alarm.
- 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 unnecessary movement to the camera. Some camera software allows for the camera control to be handled by the computer initiating the photographs and then storing the images with unique file names.
- Make sure that the camera sounds are muted to avoid disturbing others.
- Ensure sufficient lighting where possible and avoid shadows on the document if possible.
- Do not fill the entire frame with the document, instead, leave a border around all of the edges. This is because image quality deteriorates towards the edges of an image due to the characteristics of the camera lens. However, do not use a very wide border around the edge since the blank space is also stored as part of the image. When photographing thousands of images, the wasted space will limit the storage capacity available.
- Always photograph the front cover, title page, box or envelope of a document so that the images can be correctly assigned afterwards. The first image(s) should include the archive reference, and if this information is not present, then place a piece of paper on the cover with the reference written on it. This will prevent confusion concerning the identity of the images. This process also helps to identify where a new document begins, when sorting images into folders.
- With bound documents, take double pages with each shot, providing the volume can be laid reasonably flat. The size and format may dictate otherwise. Nevertheless, photographing double pages will capture more data in one image and save time. In some instances, photographing double pages can produce a

[*] Specific advice on handling some of the more challenging documents can be found in Sect. 3.
[†] Specific advice on cameras and camera settings will be found in Sect. 3.2 and 3.3 of the technical guidelines below.

minor blurring either in the middle of the image or at the edges. Note that blurring may be due to other factors such as pencil notation that is not picked up as well as ink or print. One image per page may produce better resolution, and this is a decision that needs to be made at the time depending on the document and its content (for example tiny handwriting). However, photographing all documents as single rather than double pages has cost and productivity implications. You need to determine whether your priority for the day is 1200 high quality and perfectly readable 'working documents' or 500 pin-sharp and perfect archive quality images. Another possibility is the use of a device called a "Book Liberator" (see Appendix 1). This device costs about USD$400 and employs the use of two digital cameras and enables the imaging of two facing pages in a bound volume at the same time without requiring the breaking of the binding.

- 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. Do not image blank pages beyond the first blank page if a bound volume is not completely full. This will indicate that there were no further observations. It is essential however to check all blank pages in case observations resume. In particular, notes, seals, signatures, and other notation may be written at the very end of a bound volume. On observation sheets and other handwritten information, be sure to check the back of the paper/page. Occasionally a weather observer makes additional comments on the reverse side of the document.
- Use color photography – it is not unusual that color means something (e.g., temperatures in red may be negative temperatures, though the sign +/- is not written.
- Image all pages of a collection. Some notation may hold across pages (e.g., sometimes negative/positive signs are only indicated when the sign changes).

2.4.8. Types of image file

Most cameras will take two types of image file, a JPEG and/or a RAW file. Both types can be produced simultaneously. The JPEG is the best file to produce for DARE purposes and is usually 2-5Mb in size depending on the camera and settings. A RAW file is much larger usually 10-15Mb. RAW files can be converted into 'archive quality' TIFF files. An archive might request that their copies are archive quality, and this means a TIFF file. However, a TIFF file can be 100-200Mb, so TIFF files produced and stored as RAW files take up much less storage space. Handling and sorting both JPEG and RAW files will require additional time and resources, and it is not recommended that you produce additional RAW files unless the archive is insistent. There is more on image files in the technical guidelines (Sect. 3.4).

2.4.9. Organizing the image files

If your camera has wireless settings, you can download images directly to a folder on your computer as the images are taken. A folder should be produced for each separate document or bound volume. The folder name should include the archive reference, and other information so that the set of images within the folder can be quickly identified. You will save time in the archive if you have created and named the folders from the pre-imaging inventory before you start work.

Alternatively, if you have no wireless capability on your camera, a batch of images can be downloaded to your computer at the end of the day simply by inserting the data card into your computer and transferring the images. Be careful not to mix images from separate data cards, if more than one person is working on the imaging. After downloading the images, they should be copied to a backup immediately. Note that if imaging is done over the period of a week or more, the image file numbers are likely to exceed 9999, and therefore the image file numbers will revert to 0001 and continue from there. Potential problems can therefore be avoided by using a separate folder for each day's downloaded images.

The next task is to sort and transfer the images into separate document folders. This can be done daily or at the end of a week. It should *not* be left until the imaging work is completed. The folder name should include the archive reference, and other information so that the set of images within the folder can be quickly identified. For instance, a meteorological register might be prefixed MR, followed by the number assigned by the repository, and then the name of the location and a year. (MR_2990_Aberdeen_1903). In the case of a ships logbook we could use the catalogue number, ship name and year (ADM 53/5743_President_1853_1854) Imaging the front cover or using some other visual clue to indicate the first image of a document is useful. Sorting images into their folders will be made much easier when you can visually identify the first and last image in a set from the images displayed as (enlarged) tiles. Use the view settings on your computer to choose the best method to identify sets of images. Then select the required images and drag them to the corresponding folder. Once this process is complete, copy the folders to a backup. This is also the time to check on the quality of the images and ensure that duplicate images are deleted and missing images noted and captured at the next imaging session. This is why you should not leave such checks until the end of the imaging stage.

There is presently (2019) no standard convention for file naming. This may change in the future. The current best practice is to use unique filenames for traceability (e.g., such that filenames stored next to a variable value in a file allow tracing back that value to an image). Filenames can include the project number, source, content of source, imaging person, date, and a running number. It is advisable to keep the file naming simple but remember that the filename should be easily traceable. It is a good idea to consult with others about their experience and any recommendations that they might have.

2.4.10. Image file storage and security

The need for multiple backups of images cannot be over emphasized. Backups of the images should be made on a daily basis while the work is undertaken. Once completed, the individual or organization undertaking the imaging should be encouraged to keep copies, as should the repository holding the original documents. The commissioning institution should keep multiple copies and ideally copies should be stored by more than one institution. It is unwise to depend on one institution to hold images securely. Access to images can be restricted or lost entirely when institutions face budgetary restrictions or scientific programs are terminated for economic or political reasons. Some important climate data are known to have been lost (see Section 1.2).

Unfortunately, at the present time (2019), there is no best practice for the archiving of images. However please ensure images have multiple backups. Please also remember to register the imaging work you did in the C3S Data rescue activities data base (at the level of the project) and in the C3S inventory (at the level of individual station records) (https://data-rescue.copernicus-climate.eu).

2.5. Digitizing Data

2.5.1. Approaches to the transcription of weather observations

Datasets of historical weather observations are vital to our understanding of climate change and variability, and improving those datasets means transcribing millions of observations - converting paper records into a digital form. Here we review a variety of methods for doing this transcription.

It is challenging to intercompare transcription projects, as the difficulty of transcription varies a great deal depending on the difficulty of the source documents: hard projects using older documents can take 10-times as long as easy projects using modern documents. The project funding structure makes much less difference: citizen science projects perform about as well as large commercial projects, while being much cheaper. Indeed, citizen science has been a success in this field - with several projects covering a range of different source documents.

The major remaining difficulty is speed. There are so many weather observations requiring transcription that we need a much faster method than anything that has been used so far. Transcription is fundamentally slow - speeds in

these projects vary from 6 observations/minute to 1 observation every 3 minutes. It is also a lot of work - from 0.2 to 2.2 person-minutes for each observation. This is the major current limitation: the number of observations remaining to be transcribed is unknown, but 1 billion (1,000,000,000) is a reasonable planning number. At typical rates shown above, this will take of order 100 years elapsed, (and 500 person-years effort).

2.5.2. Optical character recognition

Transcription is the process of extracting digital data from printed/typed/handwritten documents. At least for some of the source documents described in Section 2.1.2, this ought to be possible, especially given the rapid pace of modern research in machine vision and image analysis (benchmarks on the Modified National Institute of Standards and Technology database of handwritten digits have reached well over 99% accuracy). However, OCR of tabular weather data is a different problem from OCR of prose: the analysis of the layout of each page is essential, i.e., finding the locations of data values on the page to be transcribed, and there is less opportunity to guess character values from preceding and following characters.

In a recent trial (2018) to test the power of the current OCR system to transcribe weather records, the OCR system was about 50% successful in reading barometric pressure observations from the documents. This is an insufficient success rate (manual methods are about 99% successful), but the test did identify opportunities for improvement. In particular, the system was almost as good at reading handwriting as print, but varied considerably in the success rate between pages from the same source. This suggests that systematic experimentation in image pre-processing (de-skewing, de-speckling, contrast enhancement, etc.) might generate large improvements given only modest effort.

In the ERA-CLIM project, OCR was used successfully for many of the sources (Stickler et al. 2014). The success depended very much on the quality of the paper and print and the layout and organization of the material. Successful OCR required pre-processing such as removing lines from table, deskewing, contrast enhancement, and despeckling.

Document transcription is not a problem specific to climate data - it is valuable to many fields, so the absence of much serious research in the area is strange. We should encourage research in this area, and in particular we should provide some benchmark datasets: collections of log page images with known transcriptions - so people writing auto-transcription tools can test their accuracy (benchmark datasets like MINST or ImageNet have been very successful in driving development). Our existing data rescue projects are obvious sources for these, and we have already made one such dataset.

Lessons learned

- OCR has been successfully applied in climate data rescue projects, but only worked for some of the sources (typically neat, printed tables).
- OCR is not yet accurate enough to be used for handwritten data.
- OCR is fast and cheap enough to eventually remove transcription as a problem - other parts of the data rescue process (imaging and analysis) would become the rate and resource limiting steps.
- Reading the text is not the main problem - handwritten documents do not do much worse than printed. The core difficulty is the layout analysis finding the text to be read on the page.
- Success rates of transcription are very variable from page to page. Researching the cause of this variation seems likely to produce large improvements in overall accuracy.

2.5.3. Manual keying – logistics and costs

Before documents are distributed for keying, they should be inspected in depth. This includes identification of station, parameters, units, time reference. The layout of the sheets may change frequently, or observers start to use an obsolete column for recording something else. Therefore, each sheet should be inspected. Based on this work, the parameters to be digitized can then be selected.

For larger projects, it may be advisable to set up an online platform for job management. Digitizers can then log in, download a package (consisting of images and templates) and upload the digitized data.

Providing good templates and instructions to the digitizers is fundamental to avoid data quality problems, particularly if the digitizers are not experts. Historical instructions for observers should be collected and examined before writing instructions to digitizers. The templates should closely mirror the structure of the data sheet in the source, so that they can be aligned on a screen (or on two screens). This is not just helpful for keying, but also for the first round of data checks, when errors will still be frequent and they will have to be checked on the original sheet. On the other hand, one should keep in mind that eventually the templates will have to be converted into a machine-readable format (typically ASCII). Therefore, it is important that a template has a consistent structure (in particular, one column contains always the same variable). If the data structure in the source changes frequently, it is much better to create many different templates for each structure than having to deal with changing structure within a template. A new template should also be created when a variable is reported in different units. It is better to create one file for each template instead of using different sheets in the same file, because additional sheets can be easily overlooked in later inspections. File names should have a standard structure that contains the most relevant information such as station name, source, period, and time resolution. In large projects, the templates should have a standard header containing additional metadata.

The instructions should be as detailed as possible. The following issues need to be addressed:

- Typographical errors in the source. It is good practice to transcribe such errors as they are, even if obviously wrong, and correct or flag them in the quality control process (exceptions can be made for straightforward mistakes such as column shifts).
- Missing data. There should be a consistent way to indicate missing data. In many cases one can simply leave empty cells, but for some variables this can be ambiguous. For instance, an empty field commonly indicates zero precipitation and should not be confused with a missing observation.
- Special symbols. The keyers should be instructed to reproduce every symbol as it is in the source and not to interpret them. For instance, even though the symbol " usually indicates repetition from the previous row, in some sources it could have a different meaning. Special cases are when a symbol is unambiguously used to represent missing values or zeroes (for example -.5 for 0.5); these should be translated by the keyers (specific instructions must be given, possibly with some example).
- Negative temperatures. One of the most common mistakes when digitizing temperature data is the loss of the negative sign. This kind of error is very difficult to correct and sometime even to detect, therefore it is very important to check how negative temperatures are represented and instruct the digitizers accordingly. For instance, it is common in cold climates to indicate the sign only when it changes from the previous observation; in this case, there should be an additional column in the template for the sign.
- Non-decimal sub-units. When digitizing early instrumental data, one has often to deal with non-decimal sub-units. For instance, French inches were divided into 12 "lines"; in this case, a pressure observation of 26.11 translates to 26.917 inches in decimal units. The most sensible thing to do is to divide each observation into two or more columns in the template (depending on how many sub-units are used). It is also good to inform the digitizers on the possible range that each column can have (the "lines" column, for example, can only have values smaller than 12).
- Visual aids. In the templates, it is useful to add borders and/or different fill colors to separate columns and in general to highlight the data structure. The digitizers should also be encouraged to use these to facilitate their work or to highlight problems.

- Disable Auto complete. By default most spreadsheet programs suggest what to write in a cell based on what was written in the previous ones. 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, units change, etc.) 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. Sometime 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 that in the template. If possible, the columns with the dates should be already filled in the template that is given to the digitizers.
- Monthly sums. When available, monthly means/sums or similar can be useful for quality control. They should be typed in a way so 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.

Ideally, the digitizer should receive personal instruction about each digitization work. However, in practice this is not possible in large projects. In those cases, it is important to provide instructions that are at the same time detailed and comprehensible. Complex issues might be easier to explain with a figure.

Even though most of the instructions can apply to any template, a specific subset of instructions is usually necessary for each template. Template-specific instructions should be given particular prominence (the digitizers might tend to ignore the instructions after completing several templates). Moreover, it is extremely helpful to provide an example of typed data in the templates (this can be the first row of data and/or some potentially problematic entries).

2.5.4. Crowd-sourcing – set up and management

Citizen science works as well as large-scale commercial efforts and has a much lower financial cost. For instance, the costs of the Citizen science projects “Royal Navy World War 2” and “East India Company” were about 1 million GB£ each, with 8 million and 0.6 million observations being rescued (see brohan.org/transcription_methods_review/). The projects “East India Company” and “oldWeather3” cost only 100 k GB£ and 110 k GB£, respectively, and rescued 7 million and 1.3 million observations, respectively. It’s also encouraging that citizen science has been successful not only with ship’s logbooks (thanks to plenty of human interest) but also with the intrinsically less appealing data tables used for the UK DWR stations. A lot of the credit for the success of citizen science in this field is due to Zooniverse.

Participation rates in citizen science projects are sensitive to the difficulty of the exact task requested of the volunteers. Making the requested unit of work as small as possible, for instance, by presenting only a fragment of a page to be transcribed, rather than the entire page, can increase participation. The speed, cost, and efficiency of transcription depends most on the difficulty of the task: observations in hard-to-read older documents take several times more time and effort to read as those in easier, more modern documents. This is why the time per observation rescued in the above-mentioned projects varies between 0.15 and 3 minutes per observation (see brohan.org/transcription_methods_review/).

2.5.5. Automatic curve extractors - present status of technology

Programs exist, such as IEDRO's WeatherWizards.org crowd-sourcing website, to determine the values on a pixel by pixel basis of traces on analog charts such as pluviograms, barograms, and thermograms (see video http://weatherwizards.org/learn-more/). In these processes, digital images of specific strip charts are input into the program. The software automatically places a dot along the original trace at 15 minute intervals (or less). Based on the baseline data of the particular chart, parameter values are produced in a comma-delimited format that can easily be transformed into a spreadsheet such as Excel. The program is quite accurate, however crowd-sourced volunteers correct the occasional misplaced computer-generated dot before the data are archived. Although the initial cost of configuring the program for each specific chart from a variety of manufacturers can run USD$1000-$1500, the subsequent costs for volunteers to image thousands of imaged charts is negligible.

There are also more simple techniques. Sometimes charts can be halfway properly aligned in an image processing software and then imported into a graphics software such as CorelDraw. Here, the curve can be redrawn manually by clicking on the image – each click will give a node of a curve. Do not forget to also digitize four known points of the coordinate system. Removing the image, saving just the curve as a postscript file and opening that file in a text editor lets you extract the coordinates of the nodes easily. These coordinates, together with the four points of the coordinate system, lets you convert the node coordinates into the units of the underlying coordinates.

2.5.6. Speech recognition

Another way to digitize imaged data is speech recognition. After some training, the software will recognize your voice well, and software such as Dragon Naturally Speaking can be constrained to understand only numbers. The number can be spoken in any combination of cyphers, e.g., 142 can be spoken as "one-four-two", "fourteen-two", "one forty-two", or "one hundred and forty two", and the software will produce 142 in all of these cases. Language commands can be used to control the software ("next line", "save" etc.). A trained speaker can perform speech recognition at high speed for one, possibly two hours but not longer in many tests conducted. The speed is that of speaking and can reach 1-2 s per number. Speech recognition does not work equally well for all people, and some easily get a sore throat, which then hampers further speech recognition work.

2.5.7. Keying errors and quality control

If redundant information is available (e.g., multiple levels in upper-air data, several times daily pressure and temperature), applying QC allows reaching error rates below 2% without double keying. Double keying is independent keying of the same data. If redundancy is low, which is typically the case for precipitation, QC is more difficult, and double keying may be necessary. C3S offer a suite of QC tools, which are described in separate Best Practice Guidelines.

2.5.8. Further steps

Digitized data may be in any format, often in a spread sheet that mirrors the layout of the paper source. For further processing a conversion into a common format is required. Specific Best Practice Guidelines on Formatting, Metadata and Quality Control are forthcoming. The most important aspect to keep in mind here is to keep raw data raw. The original spread sheets (or whatever file format it is) should be kept unchanged. When submitting data to the Data Holding for land stations of the Copernicus Climate Change Services, these original files should always be submitted and constitute the fundamental source of the data.

Again, as mentioned at the end of the imaging section, it is important to update the C3S inventory (https://data-rescue.copernicus-climate.eu) to let the community know that the data have been digitized. This helps avoiding duplicate work.

2.6. Implementation of a Large Data Rescue Project

The C3S catalogue of Data Rescue Projects (https://data-rescue.copernicus-climate.eu/projects) is a good place to start to determine if any DARE activities have occurred with the data owner or his/her organization in the past. Using that information, and before contacting a data owner, the project manager must

- find out who the "data owner" really is. Who has the authority over the disposition of the data? It is usually the Director of the NMS, or the head of the NMS climate archives. He/she might be the Minister which has authority over the NMS.
- find out the data owner's attitude on DARE by contacting all parties which have been involved with the data owner such as the WMO, UNDP, World Bank, USAID, professional colleagues to find out as much as possible about the data owner, and the organization. Without the enthusiasm/passion for the DARE, the project will fail
- find out his/her native language and what is his/her English capability after which the project manager must try to communicate both orally and in writing in the owner's primary language. When writing, it also is a good idea to provide the same text in English in case he/she needs to share the contents with another.
- In developing countries realize many data owners were political appointees with no hydrometeorology/ climate background. Many data owners may not have a university education and, a few will have limited high school education. Knowing this will enable the project manager to tailor the "marketing plan" so that the director of the NHMS can understand the importance of DARE and agree to the project.
- Find out how long he/she has been in their current position? The longer he/she has been in the position, the more informed they are regarding the capability of their service. Through the data owner and the staff, the project manager needs to find out as much information as possible regarding the data to be rescued, the availability of local NMS resources (staff, equipment, etc.) so that a reasonable cost estimate could be presented to the potential funding organization.
- Naming of an on-site DARE Project Manager. Once the data owner is identified, an enthusiastic and capable DARE **Focal Point** must be appointed by the data owner, other than the data owner him/herself. This focal point should be able to communicate directly and effectively with both the DARE organization assisting.
- Once we have some basic information about the data owner, we need to realize <u>their</u> specific needs and wants and the specific needs and wants of the funding organization regarding the data to be rescued. These requirements must match for the project to be a success. To accomplish this, we need to know the following
  - type of data (surface, upper-air observations)
  - data locations (regions, countries, ship data)
  - data years (oldest data, data about a specific event)
  - condition of the data and its durability (paper, microfiche, microfilm, mag tape, glass slides)
  - availability of data already digitized (at a world data center such as NCEI)
- In order to get a commitment from the data owner to fully support and be involved in data rescue, the project manager must do the following:
  -The owner must be given realistic options with projected costs in staff time and funding from unbiased sources that have no profit motive which must include changes in facilities, workspaces, training of personnel, understanding local wages and restrictions, etc.
  - The data owner must be educated that the paper-based (or microfiched) data is <u>absolutely worthless</u> in its present form, that any analysis requiring dozens of years of data will be too costly in staff time to complete, that no potential "buyer" of these data could afford to wait the weeks for a technician to arrive at an answer, that no potential buyer of these data would undertake the task him/herself and that the data medium is deteriorating and data will soon be lost.
- The data owner must be informed that the data must be imaged for legal purposes in many countries. Digital photography meets this requirement.

- The data owner must be shown that even the smallest NMS has spare technician time to image the data in preparation for digitization. Once the imaging is done, crowd-sourcing efforts may be the only economical way to accomplish the digitization of the imaged data. Consider the keying of only an NMS's aviation observations over 50 years from 40 observation sites at USD$0.02/keystroke by a commercial organization. This will cost a small country over USD$17,000,000!
- The data owner must be convinced to work with non-profit organizations into placing images onto a citizen science / crowd-sourcing website for accurate, timely, and inexpensive digitization.
- The owner must be shown realistic options with projected costs by placing the images his/her staff has taken onto a crowd-sourcing / citizen scientist website.
- Once the initial cost of modifying the ingest end of the crowd-sourcing website to handle a specific chart or form is made, the actual digitizing of the thousands of sheets of weather and climate data is a small fraction of the cost of manual digitization.
- Data Rescue Reports are essential to record various steps of the entire DARE process noting work flow, bottlenecks, documenting successes as well as failures. All DARE projects should be given a post mortem to evaluate the process and discover ways the entire work process can be improved.

3. Technical Guidelines

3.1. Scanners

In deciding whether to use optical scanners or digital cameras to image historic hydrometeorological data, several factors must be considered. Low cost and speed is usually on the digital camera side when the technician is imaging single pages of letter-sized documents or bound pages. Also when single sheets of paper are significantly large as when imaging maps or charts, cameras can be adjusted to image the entire paper by raising the camera on the camera stand increasing its field of view. However, when imaging strip charts such as pluviograms, thermograms or barograms, a scanner with an automatic feed is preferred. Whichever method is used, it is most desirable to have the imaging device activated by the computer which automatically stores the image on the computer's hard drive.

3.2. Cameras

3.2.1. Types of camera

Most modern cameras produced today are very good, and there is a wide choice available. This makes choosing a suitable camera more difficult; there is almost too much choice. In the end the choice may be dictated by what is considered a set of essential features and functions, balanced against an equipment budget. Past and on-going data imaging projects have made use of cameras produced by Canon and Sony. A bridge camera, or camera without interchangeable lenses, is fine for photographing documents although some of the smaller but very sophisticated compact cameras are more portable and easier to transport. This is an important consideration, as even when only one person is working, a spare backup camera should be carried in the event of equipment failure. A typical compact camera such as the Canon G5X (current at 2019) has been used successfully in recent and on-going imaging projects.

3.2.2. Essential features

Features considered either essential or optional will be a matter of project requirements and personal choice, and are selected here on the basis of experience.

- Fully-articulated viewing screen. This is a primary consideration as the photographer needs to be seated and comfortable while working, and should be able to adjust the screen in order to view the image screen from any angle.
- An option to have either fully automated settings or custom settings.
- An image stabilizer – useful for sample images where no tripod is used.

3.2.3. Optional features

- Wi-fi functionality

3.2.4. Ancillary equipment

- Tripod – this should come with the option of removing the stem (to which the camera is fixed) and inverting it so that the camera hangs downwards between the three legs of the tripod.
- A boom or extension arm attached to the top of the tripod will avoid the necessity of having the tripod directly above the document, if this is preferred.
- Remote switch.
- Power supply or power bars to charge the camera (or spare batteries)
- An additional lighting array, for working in archives with inadequate lighting.

Fig. 3. Camera setup (Photos Clive Wilkinson & Tomoshige Inoue)

3.3. Handling and Imaging of documents

3.3.1. Physical condition of documents

Most historic documents are written on good quality paper that is usually superior to modern paper. This is one reason they have survived for centuries. However even the best quality paper will not withstand damp and other adverse environmental factors. If you find yourself handling documents that are damp or have fungal growth, **stop immediately**. There may be serious health implications, especially if the documents have originated from or been stored in tropical regions. Likewise, if the paper begins to disintegrate, do not attempt to handle it. Any document clearly in need of conservation or in a fragile condition should be drawn to the attention of the archivist or other staff.

Bindings are usually far inferior to the papers they contain, and it is most likely that you will find the bindings are the major difficulty. Always use the book rests or cushions provided by the archive in order not to strain the binding of documents. Nevertheless, when using old leather bound documents, 19thC ship logbooks for instance, there will always be small fragments of old leather in the box containing the document, and on the table once you have completed examining and imaging the document. This is normal and unless the deterioration is excessive, need not be drawn to the attention of the archive staff. If old leather dust is bothersome, you may wish to consider using a mask.

Never wear gloves when handling documents. You may have seen this on the television, but it is nothing more than a dramatic artifice used to emphasize the importance of old documents. All archives have a supply of thin white cotton gloves, but these are only produced when film crews attend. Gloves will slow down handling and cause more damage to the documents. However, try to ensure that you do not rub your eyes, and always make sure you wash your hands after each session in the archive.

3.3.2. Problem documents – format, faded ink and pencil

Most documents are easy to image, but others present some challenges. Large, thick, tightly-bound items are the most common problem. Often these are not the original bindings but are re-bound documents. Nevertheless, do not strain the binding. Useful data is sometimes too close to the center binding and is impossible to image clearly, even when photographing one page at a time. If it is important, then manually transcribe the information as well as imaging the page(s).

Occasionally you will find a document where the ink has apparently faded. Ink will not fade if the document is stored in unlit conditions. Faded ink in hand-written documents is more often the result of the writer running short of ink and watering it down. You will notice an immediate improvement once the writer has a fresh supply.

Imaging faint ink is a challenge. Experiment with your camera settings (white balance, ISO, aperture) and lighting to see if you can get a better resolution. Also try the automatic settings, as the many cameras are sufficiently sophisticated to work out the best settings for themselves. If all else fails, direct manual transcription may be the only solution, but keep a copy of the image you have directly transcribed for verification, no matter how bad the quality.

Pencil notation on the paper records is a particular problem. Printed parts of a document and notion in ink will always look sharp on an image. Pencil tends to have fuzzy edges, and once imaged tends to have poor resolution. There is nothing you can do about this except ensure that you have the best lighting and camera settings.

3.3.3. Positioning of materials to be imaged

Position materials so that there is a small border around the document. Image quality will always deteriorate towards the edge of the image due to the characteristics of the camera lens. This is inevitable, so leaving a small border will mitigate the problem. Very large documents, such as ice charts will require some thought. It might be possible to put a chart on the floor, with the archivist's permission, and photograph it hand-held. An image stabilizer on the camera then becomes essential. Alternatively, you can image the chart in sections and either have a collection of images for the one document or stitch them together using either software on the camera itself or on your computer.

3.3.4. How good are automatic camera settings?

The answer is very good. Recent trials using different camera settings demonstrated that the automatic settings were as good and sometimes even better than manual settings, of shutter speed, aperture, ISO and lighting type. Much will depend on the quality of the camera and the conditions in the archive. It is worth the time to experiment with your camera.

3.3.5. Lighting

Experience has demonstrated that the best image quality results are often with a mixture of natural and artificial lighting, for example positioning yourself near a window. Cloudy days are often best as strong sunlight, especially if intermittent, can cause unwanted shadows, and cause you to shift position as the day progresses. Fluorescent lighting seems to be the best artificial light. Reading lamps often produce a yellowish cast to the image. Your camera should have settings for different lighting conditions, usually daytime, cloudy, fluorescent, and tungsten, but if you use the automatic setting, it will choose the best lighting for you. Be aware, however, that if you take your camera off automatic settings in order to take advantage of other settings such as aperture priority, you may need to set the lighting preference manually.

3.3.6. Advanced camera settings

Your camera should have a 'program' as well as other settings when automatic settings are not selected. Settings can be made for different lighting conditions, for aperture or shutter priority, image file size, image type (JPEG or RAW or both) and ISO. There may also be settings for color enhancement.

- **Shutter speed, aperture, and ISO – image sharpness and depth of field**

Shutter speed is the length of time that the shutter is open when taking an image. A lower shutter speed is needed in poor light conditions. The aperture is the size of the hole through which the image is taken when the shutter opens. The third property is ISO (sensor brightness). For a given ISO, the narrower the aperture, the sharper the image. There is a trade-off between the shutter speed and the aperture. In any given set of lighting conditions, a slow shutter speed will allow a narrower aperture, and a higher shutter speed will require a larger aperture. A higher ISO also will give you a brighter picture in low light conditions, but at the cost of a grainy, noisy

image. Low shutter speeds require a tripod or camera stand, although image stabilizer software in cameras is now quite sophisticated. For best results and sharply defined images, the aperture needs to be as narrow as possible. Good lighting is therefore essential. The camera can be set for aperture priority, which will allow you to preset the aperture size that you want. This will produce a narrow aperture setting at the expense of a low shutter speed. If the camera is in a fixed position, the low shutter speed will not be a problem. The narrow aperture will also improve the depth of field. Depth of field means that the camera is focused on a range of distances rather than a single point, meaning that where a document is not entirely flat and therefore not at a fixed distance from the lens, more of the document is in focus. This effect is less pronounced with the short distances involved in imaging documents, but may slightly improve the image quality where tightly bound documents cannot be laid flat. One should experiment with automatic and aperture priority settings so see which results are best.

3.4. Handling of Image Files

- **Advanced file numbering and naming**

There is software available (e.g., MongoDB) that will batch-number sets of image files according to a pre-arranged numbering system. There is presently (2019) no standard for image file naming, although this may change in future, and advice should be sought.

- **Type and size of portable storage hardware**

This will depend on the number and size of the images you expect to capture. Larger storage hardware is more cost effective, and at least 2TB is recommended. Always make sure that the storage device is formatted for both PC and Mac.

- **Transferring images to other users**

The organization providing the keying of the imaged data will require the images. This aspect is absolutely required since the only economical way to get the data on the images keyed is by crowd-sourcing which requires the images to be transferred to the digitization organization via the Internet for entering onto the website.

Transfer of images by a storage device delivered in person (at a meeting) or by postal service is always safe and effective. Once the images are on the storage device, make sure that a random selection of these can be opened on both a PC and a Mac. Online transfer through any of the file sharing websites (the Cloud) such as Dropbox or WeTransfer is also a possibility. Be aware however that exceptionally large data transfers may not be cost-free, and will also be time consuming if internet speed is slow (upload and download). Experience has also shown that files may arrive with the recipient in a corrupted form. Images will need to be checked by the recipient, to ensure none have been corrupted. This can be done visually. Alternatively, checksum tools such as MD5 can be used to validate the files and for deduplication of large image collections.

- **Security and backup of images**

There is an immediate need for a backup of all images taken to be provided at locations other than the data imaging area so that in case of fire or other disaster, backup copies of the images are located elsewhere. It should be noted that although relatively inexpensive portable hard drives are readily available, they are not dependable for long term storage. IT experts should be contacted to evaluate and design a robust data storage system.

References

Allan, R., P. Brohan, G. P. Compo, R. Stone, J. Luterbacher, and S. Brönnimann (2011a) The International Atmospheric Circulation Reconstructions over the Earth (ACRE) Initiative. *Bull. Amer. Meteorol. Soc.,* **92,** 1421-1425.

Allan, R., G. P. Compo, and J. Carton (2011b) Recovery of Global Surface Weather Observations for Historical Reanalyses and International Users. *Eos Trans. AGU,* **92(18),** 154.

Brönnimann, S., J. Annis, W. Dann, T. Ewen, A. N. Grant, T. Griesser, S. Krähenmann, C. Mohr, M. Scherer, and C. Vogler (2006) A guide for digitising manuscript climate data. *Climate of the Past,* **2,** 137-144.

Brönnimann, S., Y. Brugnara, R. J. Allan, P. Brohan, M. Brunet, G. P. Compo, R. I. Crouthamel, P. D. Jones, S. Jourdain, J. Luterbacher, P. Siegmund, M. A. Valente, and C. W. Wilkinson (2018) A roadmap to climate data rescue services. *Geoscience Data Journal*, **5,** 8–39.

Foster, W. (1919) *A Guide to the India Office Records 1600-1858*.

PIAF (2019) Notions générales d'archivistique, sur le Portail international archivistique francophone: http://www.piaf-archives.org/se-former/module-2-notions-generales-archivistique

Pfister, L., F. Hupfer, Y. Brugnara, L. Munz, L. Villiger, L. Meyer, M. Schwander, F. A. Isotta, C. Rohr and S. Brönnimann (2019) Swiss Early Instrumental Meteorological Measurements. *Climate of the Past* (submitted).

Stickler, A., S. Brönnimann, M. A. Valente, J. Bethke, A. Sterin, S. Jourdain, E. Roucaute, M. V. Vasquez, D. A. Reyes, R. Allan, and D. Dee (2014) ERA-CLIM: Historical Surface and Upper-Air Data for Future Reanalyses. *B. Amer. Meteorol. Soc.,* **95,** 1419–1430.

The National Archives (2016) Archive principles and practice: an introduction to archives for non-archivists, 24pp., available on website: http://www.nationalarchives.gov.uk/documents/archives/archive-principles-and-practice-an-introduction-to-archives-for-non-archivists.pdf

WMO (2016) Guidelines on Best Practices for Climate Data Rescue. WMO-No. 1182, 30pp. WMO: Geneva, Switzerland.

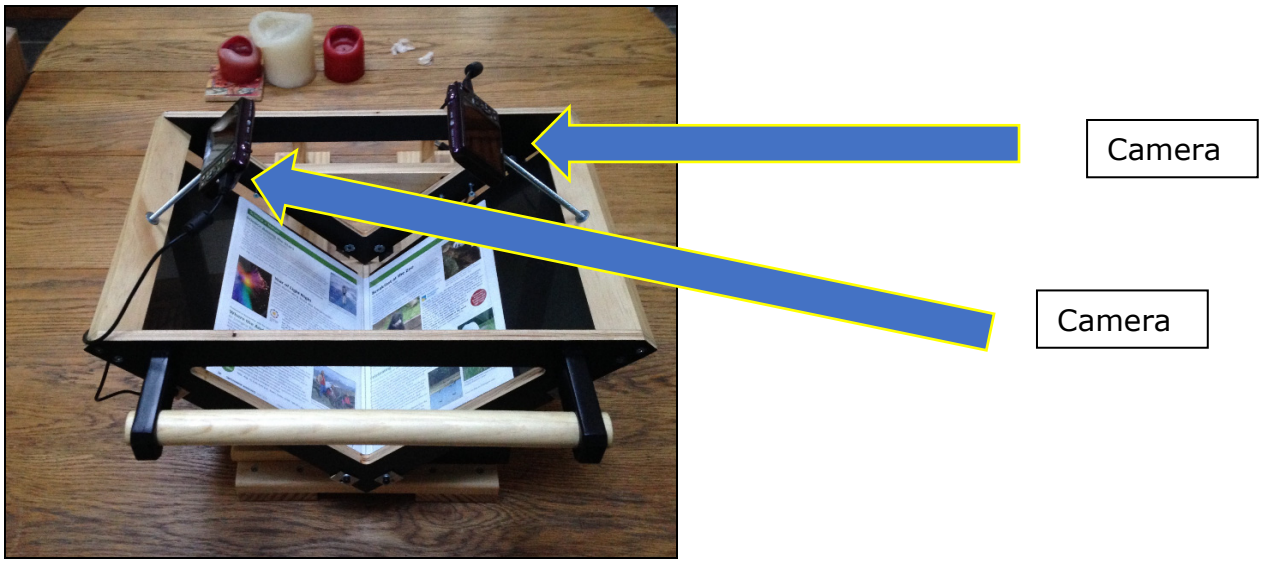
Acronyms

ACRE	Atmospheric Circulation Reconstructions over the Earth initiative
C3S	Copernicus Climate Change Services
DARE	Data rescue
IEDRO	International Environmental Data Rescue Organization
NMS	National Meteorological Services
OCR	Optical Character Recognition
WDC	World Data Center
WMO	World Meteorological Organization

Appendix 1: Example of a Large Project Imaging Setup: Uzbekistan Data Rescue Camera Setup

A dual-camera system (controlled by software on a PC workstation) attached to a custom-built camera stand (”Book Liberator”) was used in Uzbekistan to convert paper records to digital images. This system allowed high-quality images to be obtained without breaking or destroying the book bindings or harming the original paper records. Once the images are obtained, the workers perform post-processing tasks, such as rotating and cropping images, confirming adequate image resolution, and assuring that quality control standards are met.

- **Example of Camera-“Book Liberator” Setup**



- **Dual Digital Cameras**

Each Imaging Workstation has two Nikon Coolpix Digital Cameras (with cables) mounted on a fixed camera stand, called a “Book Liberator”, to image the paper data. IEDRO uses a double imaging technique that enables an operator to photograph two pages at once (assuming the pages are printed on both sides). This significantly increases efficiency and decreases the overall imaging time. The cameras are hard-wired to desktop computers.

- **Camera Stands / “Book Liberators”**

This custom-built innovative camera stand holds two digital cameras so that an operator can photograph two facing pages of a bound document almost simultaneously without needing to pull the bound volume apart.

Post Project Comments / Lessons Learned

- **Need for Better Cameras**

  Cameras used for the Uzbekistan Data Rescue project were sufficient for the task, however much higher quality cameras would be purchased for future data rescue efforts. Lesson learned: Always purchase the highest quality cameras that the budget permits.

- **Light Source**

  The workroom was not sufficiently lit for data rescue and an external light source was attached to each camera stand to insure capture of high quality image. A higher quality camera may not have needed the additional lighting however the lesson learned was to establish a workroom of sufficient light or invest in better cameras

- **Camera Stand / Book Liberator**

  This device was a great success with the only need for modification to the project was creation of more of the larger stands because Uzhydromet desired larger document rescued first.

Appendix 2: Preserving Physical Sources

Preventive preservation seeks to reduce risks of damage and to slow down the rate of deterioration. These aims are usually accomplished by selecting good quality materials and by providing suitable storage environments and safe handling procedures. There are many different reasons why records and archives deteriorate.

- Archives are always in danger because events happen which can damage buildings and the archives stored in them.
- Fire prevention is the first line of defence in preventing destruction or irreversible damage to archives. Water will cause major damage to archives. Flooding can be caused by water coming into the building from outside or by water leaking from tanks or pipe work inside the building.
- Protecting archives from theft, deliberate damage or disorder ensures that they remain complete, intact and usable. Storage rooms, areas and cupboards must be lockable. Only the people responsible for caring for the archives should access these areas.
- The impact of risks can be reduced by disaster management planning, for example arrangements for salvaging the archive if a major incident such as a flood, a fire or a break-in happens.
- Archives need to be stored in conditions which are clean, cool, dry and seasonally stable, with minimum exposure to natural or artificial light and protection from pests, pollution and access by unauthorised people.
- The archive store should include a strong, load bearing floor to take the weight of the archival collections, the packaging and shelving.
- No item should be moved or removed without the permission of those responsible for their care. If an item is moved or removed, a note must be left with details of where it is and who has it.
- Archival processing includes basic preservation practices such as removing staples and paper clips, placing materials in acid-free folders and boxes, isolating acidic materials to avoid acid migration, photocopying damaged or acidic documents, and unfolding papers. Packaging is an extra layer of protection for archives. It should protect the contents from light and pollution. It also provides some protection from damage, pests, and changes in light or humidity.
- It is important to label boxes clearly, including covering dates, so that you know what they contain

The UK National Archives (2016) and the International Francophone Archival Portal (PIAF, 2019) explain what you can do. It is possible to contact national, regional or international conservation associations, to obtain information about conservators and preservation management.