

Guidelines for Hydrological Data Rescue



**World
Meteorological
Organization**

Weather · Climate · Water

WMO-No. 1146

Guidelines for Hydrological Data Rescue



**World
Meteorological
Organization**

Weather · Climate · Water

2014

WMO-No. 1146

EDITORIAL NOTE

METEOTERM, the WMO terminology database, may be consulted at http://www.wmo.int/pages/prog/lsp/meteoterm_wmo_en.html. Acronyms may also be found at http://www.wmo.int/pages/themes/acronyms/index_en.html.

WMO-No. 1146

© World Meteorological Organization, 2014

The right of publication in print, electronic and any other form and in any language is reserved by WMO. Short extracts from WMO publications may be reproduced without authorization, provided that the complete source is clearly indicated. Editorial correspondence and requests to publish, reproduce or translate this publication in part or in whole should be addressed to:

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

Tel.: +41 (0) 22 730 84 03
Fax: +41 (0) 22 730 80 40
E-mail: publications@wmo.int

ISBN 978-92-63-11146-3

NOTE

The designations employed in WMO publications and the presentation of material in this publication do not imply the expression of any opinion whatsoever on the part of WMO concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The mention of specific companies or products does not imply that they are endorsed or recommended by WMO in preference to others of a similar nature which are not mentioned or advertised.

The findings, interpretations and conclusions expressed in WMO publications with named authors are those of the authors alone and do not necessarily reflect those of WMO or its Members.

CONTENTS

Page

PREFACE	v
1. WHY DATA RESCUE?	1
1.1 Introduction	1
1.2 Hydrological data rescue	1
1.3 Reasons for hydrological data rescue	2
1.4 Benefits of hydrological data rescue	2
1.5 Hydrological data for monitoring change	3
1.6 Need for engagement between national agencies and the international community	3
2. GENERAL GUIDELINES FOR DATA RESCUE	4
2.1 Definition of aims	4
2.2 Creating an inventory of data holdings and a data rescue plan	4
2.3 Prioritizing data to be rescued	6
2.3.1 Prioritizing river gauging stations	7
2.3.2 Prioritizing surface water records	7
2.3.3 Choosing the format	8
2.3.4 Prioritizing other records	8
3. PRACTICAL TIPS FOR DATA RESCUE	8
3.1 Assembling a team and managing a data rescue project	9
3.2 Improving paper storage conditions	10
3.2.1 Storage environment	10
3.2.2 Archive management	12
3.3 Digitization of records	12
3.3.1 Approaches to digitization	12
3.3.2 General guidance on digitization projects	13
3.4 Creating and storing images of paper documents	13
3.4.1 Low-cost solution for scanning documents	15
3.5 Digitizing microfilm and microfiche	15
3.6 Digitizing charts	16
3.6.1 Before digitizing charts	16
3.6.2 Cataloguing station information	17
3.6.3 Cataloguing available charts	17
3.6.4 Checking the quality of charts	17
3.6.5 Tips for digitizing charts	17
3.6.6 Software for chart digitization	18
3.7 Digitizing rating and gauging information	19
3.8 Keying data from paper documents, images or charts	20
3.9 Recovering data from obsolete digital media and formats	20
3.10 Verification of digitized or recovered data	22
4. GOOD PRACTICE FOR HYDROLOGICAL DATA MANAGEMENT	22
4.1 Hydrological databases and data management systems	22
4.2 Cataloguing metadata from rescued data	22
4.2.1 Metadata standards for hydrological data	22
4.2.2 Capturing metadata	23
4.3 Securing rescued data for the longer term	25
4.4 Sharing and securing data through storage in an international database	26
5. CONCLUSION	26

APPENDIX A. EXAMPLES OF PREVIOUS DATA RESCUE PROJECTS 28

APPENDIX B. HYDROLOGICAL DATA RESCUE: THE CURRENT STATE OF AFFAIRS. 31

REFERENCES 37

PREFACE

The potential loss of historical meteorological, climatological and hydrological data in many parts of the world is an issue that the World Meteorological Organization (WMO) has been addressing for more than three decades. It has become a major concern in hydrology where for more than two decades, hydrological observing networks worldwide have been on the decline. Significantly, most countries have at least some need for data rescue. Although the reasons are varied, most fall within two general categories: record deterioration and risk of catastrophic loss. The World Meteorological Organization, through its Hydrology and Water Resources Programme, has been supporting the efforts of Members in data rescue activities. In addition, in 2006, it sponsored a survey of WMO Members that documented significant international concern about the state of historical data records and the need for data rescue activities in almost all countries. The survey also demonstrated the need among many Members for a set of guidelines specifying procedures for implementing hydrological data rescue. The current publication is designed to meet that need. These guidelines review the rationale for hydrological data rescue, the benefits to be derived therefrom, appropriate rescue methods, sound data management practices as well as data management systems, procedures for securing rescued data far into the future and for safeguarding data through storage in an international database. As with all WMO guides, this one provides a range of options so that Members can choose the most appropriate to their unique circumstances. We hope that the hydrological community will find these guidelines of help in its efforts to secure and preserve its data holdings over the long term.

The WMO Commission for Hydrology would like to thank Matt Fry for preparing this publication, and Harry Dixon, Gregorz Słota, Theo Brandsma, Bogdan Ozga-Zieliński, Ross James and Hama Kontongomde for their valuable contribution to the document.

(Harry F. Lins)
President, Commission for Hydrology

GUIDELINES FOR HYDROLOGICAL DATA RESCUE

1. WHY DATA RESCUE?

1.1 Introduction

Historical data records are important for water resources management and modelling, climate change assessment, flood modelling and prediction, and other hydrological activities. Hydrological data are costly, in terms of both effort and resources, to record and collect. Large volumes of data are lost due to inadequate archiving and poor maintenance of data archives.

There are many examples of environmental data rescue projects, particularly in the meteorological and atmospheric communities such as the Atmospheric Circulation Reconstructions over the Earth (ACRE) initiative. Examples of past environmental data rescue projects and the lessons learned from them are described in Appendix A. The World Meteorological Organization (WMO), through its Hydrology and Water Resources Programme, has been supporting the efforts of its Members in data rescue activities (through Voluntary Cooperation Programme (VCP) and World Hydrological Cycle Observing System (WHYCOS) projects) to address issues of data loss and to secure data against future loss.

The results of a survey of WMO Members, carried out in 2006 (Fry, 2010), demonstrated significant international concern about the state of historical data records and a need for data rescue activities in almost all countries. Many of them requested assistance with data rescue activities. These guidelines are an important step towards providing that assistance.

The survey found that the risks to data records differed from one country to another, so there could be no “one size fits all” solution. There are, however, some broad approaches to data rescue that can be applied in most cases.

These guidelines aim to:

- Highlight the importance of hydrological data rescue;
- Provide an overview of hydrological data rescue issues;
- Identify data rescue issues that are specific to hydrological data;
- Provide sufficient information to allow data managers to decide how to go about data rescue.

Each organization will have to find the approach to data rescue that best suits its own needs. Hence, these guidelines aim to provide general guidance and suggestions, with links to help find more specific information.

1.2 Hydrological data rescue

The World Meteorological Organization defines data rescue as the ongoing process of preserving all data at risk of being lost, due to deterioration of the medium, and of digitizing current and past data into computer compatible form for easy access (WMO, 2002).

This definition applies to all types of meteorological and hydrological data.

It should be noted that adequately preserving hydrological data does not refer solely to the storage of the measured values of parameters such as flow, but also to the crucial information that

allows these values to be fully understood, such as where they were measured, the means by which they were measured or derived, the accuracy of the values, any major influencing factors (for example, upstream reservoirs) that could determine further use of the data, and any other metadata related to the quality of the values.

1.3 **Reasons for hydrological data rescue**

There are many reasons why data rescue is required. The WMO data rescue survey (Fry, 2010) identified the following problems:

- Deterioration of records:
 - Paper damage due to humidity;
 - Paper damage due to rodents or other pests;
 - Microfilm/microfiche damage due to heat and humidity;
 - Deterioration of archives due to unmanaged use, resulting in loss of records and inventories;
 - Destruction of archives due to war or civil unrest;
- Risk of catastrophic loss of data due to:
 - Storage of a single set of unique records;
 - Vulnerability to fire, floods, hurricanes or other natural disasters;
 - Storage media becoming unreadable due to changes in technology;
 - Electronic file formats becoming obsolete and unreadable, often due to changes in hardware, software or the loss of technical know-how to read them;
 - Loss of corporate knowledge of the existence of the archives or sets of records, leading to unintentional destruction of archive material or deletion of electronic files;
 - Loss of human knowledge, often through manpower changes or retirement, that could significantly enhance the utility of records and data.

In each of these cases it is possible to stop or reduce the pace of deterioration of archived records, or to mitigate the loss of data or knowledge.

1.4 **Benefits of hydrological data rescue**

Data rescue can require significant effort and resources, but the benefits often outweigh the costs:

- Longer records can provide more precise and accurate assessments of water resources and flood variability, hence, potentially, more cost-effective or robust water supply and flood protection infrastructure;
- A more dense observation network, or a network monitoring a greater range of catchment types, can provide a better understanding of hydrological processes and improve the results of regionalization methods that underpin many hydrological models;

- Through data rescue, hydrological records can be extended at a fraction of the cost of ongoing field measurements;
- Instigating improvements in data management and archiving practice can bring organizational benefits both through efficiencies and through improved access to information;
- Rescuing and making use of quality metadata can provide enhanced understanding of, hence accurate use of, hydrological information;
- A comprehensive inventory of data holdings can be useful for demonstrating the value of a data archive and justifying the ongoing expense of running such an archive.

1.5 **Hydrological data for monitoring change**

Hydrological monitoring activities have always considered the need for long hydrological records for the purposes of water resources planning, flood estimation and to maximize the understanding of the range of conditions. Recently, however, an increasing need has emerged for long-term datasets to understand how hydrological regimes are responding to climatic variations and anthropogenic influences. Whilst global climate models can inform us about expected impacts of change, the validation of these models requires real data. More importantly, society needs to know the impacts of change at the national and catchment levels and to identify emerging trends or changes in hydrological regimes at these scales. This can only be done by assessing long-term records that capture the natural variability.

There are local and regional initiatives to establish networks for monitoring trends in river flows and the impacts of climate change (Whitfield, 2012; Bradford, 2003), but there is an urgent need for long-term records globally and for information about the impact of human activities on these flows. Data rescue provides an opportunity to identify important records that previously may have been considered unimportant (for example, data on small rural catchments). The relevance of such records for national water resource planning and for global understanding of the impacts of climate change should be a major driver for data rescue activities.

1.6 **Need for engagement between national agencies and the international community**

Hydrological data are important at the national level for operational monitoring and planning, but there is also an increasing demand internationally for hydrological information, both for operational and research purposes. Recognizing this, the Thirteenth World Meteorological Congress, through its Resolution 25, adopted “a stand of committing to broadening and enhancing, whenever possible, the free and unrestricted international exchange of hydrological data and products (...)” (WMO, 1999). This resolution sets out more explicitly the following aims:

(1) Members shall provide on a free and unrestricted basis those hydrological data and products which are necessary for the provision of services in support of the protection of life and property and for the well-being of all peoples;

[...]

(3) Members should provide to the research and education communities, for their non-commercial activities, free and unrestricted access to all hydrological data and products exchanged under the auspices of WMO.

All data rescue activities should aspire to additionally meet these aims in order to maximize the benefit of hydrological data through increased accessibility.

2. GENERAL GUIDELINES FOR DATA RESCUE

2.1 Definition of aims

Before embarking on a data rescue project, one should carefully consider and define its aims. These may evolve throughout the project as more information about the size and extent of the task becomes available; so it is important to keep these aims under review to ensure they are realistic and, if not, to revise them. The aims of a data rescue project will generally, but not always or exclusively, include:

- Creation of an inventory of data holdings;
- Rationalization and indexing of physical records/data holdings;
- Identification of data holdings of genuine importance;
- Improving storage conditions for physical data holdings and storage media;
- Digitization of time series data;
- Extraction of data from obsolete storage media (physical and digital);
- Storage of digital data in future-proof formats and storage media;
- Storage of time series data within a hydrological database;
- Establishing procedures for the ongoing use of data archives;
- Establishing procedures for the ongoing use of database systems;
- Establishing procedures for ongoing security of digital archives such as electronic backups.

2.2 Creating an inventory of data holdings and a data rescue plan

The first, and perhaps most important, step towards data rescue is the drafting of an inventory. This will help to identify the data at risk and prioritize the data rescue effort. The inventory will also help estimate the resources required to carry out the project and their cost, and formulate a plan for what needs to be rescued.

Table 1 below lists some possible steps in setting up an inventory where there is a large amount of data. Two examples, next to each step, describe what this might entail for (a) paper and (b) digital media of obsolete format. In some cases the action required may be the same for both paper and digital data rescue.

Table 1. Steps to set up an inventory of large amounts of data

<i>Process</i>	<i>Example (a): large number of paper records stored in cabinet drawers</i>	<i>Example (b): obsolete formats of digital media (e.g. magnetic tape stored in boxes)</i>
Identify existing institutional knowledge	First of all, try to identify people or documents that could describe the data holdings, how they are currently organized and their contents. There may be an existing inventory, or someone may be able to summarize what it contains.	Try to identify people or documents that could describe the data holdings, or someone who might remember the hardware or programs used to read and write the data in the past and could thus assist with formats.
Estimate approximate number of records	Estimate number of cabinets, number of drawers and approximate number of records per drawer.	Estimate number of boxes and tapes; try to gauge the volume of data per tape and translate this into some sort of number, which may represent the number of digital records or the volume of data.
Inspect samples of the total content	Select a document from each drawer or cabinet, as appropriate, and note the quality of the document and the information it contains; take a copy where possible. Ensure each sample is returned to the original location.	Sample the boxes of tapes in order to check the different types and sizes of media stored; if possible, estimate the sample quality (for example, how much it has physically deteriorated). As it might be difficult to extract the data from the media, note carefully any written information accompanying the sample, which could shed light on its contents, arrangement (for example, tape blocking factor) or extraction method.
Collect examples of different types of record and range of quality	Collect examples of records such as site descriptions, gaugings/discharge measurements, cross-sections, logger charts, gauge recording books, ratings, processed data sheets and publications. Collect documents showing range of quality: at least worst, best and indicative typical quality.	Collect an example of each storage medium for further investigation. Ideally, choose an example of good quality media that could be read, and several, lower-quality examples to determine the extent to which data have already been lost.
Identify content of records at the appropriate level	For each type of record, identify a list of the data held, for example, a list of monitoring sites and estimated periods of record.	Attempt to read contents of sample media. If this is not possible using existing hardware, a specialist company, or another national institution, may have to be used (these should be carefully sourced). Resulting information may well need further processing to be meaningful.
Refine estimate of total number of records by type and contents	Produce a more accurate estimate of the number of records of each type and the number of different monitoring sites.	If information can be extracted from the sample media, try to scale up the volume of data on one tape to give an estimate of the total volume of data and the total number of stations involved.
Consider priorities (see section 2.3)	Consider the different types of data and different stations in the light of priorities such as those outlined in section 2.3.	
List the processes which the data need to undergo for rescue	Considering the aim of the project, for each type of document, list the steps which need to be undertaken, for example, photographic scanning and archiving of channel cross-section records, digitization and entry of water level records in a database, and securing storage for station records.	Consider the different steps required to extract the information from the media so that it can be used.

<i>Process</i>	<i>Example (a): large number of paper records stored in cabinet drawers</i>	<i>Example (b): obsolete formats of digital media (e.g. magnetic tape stored in boxes)</i>
Estimate effort required per unit, for example, per document or station	Attempt to assess the time and financial resources required for each process.	
Estimate total effort	Scale up the effort per document to gauge the total per station and/or per document type.	Scale up this estimate to gauge the total cost and effort required per station or storage medium.
Break down total effort into smaller components and define priorities	Group stations and types of document, for example, as high-, medium- or low-priority, and estimate resources for each.	Group stations and tapes, for example, as high-, medium- or low-priority, and estimate resources for each.
Define scope of project based on available resources	Consider available resources and try to define what could be included in or excluded from the data rescue project.	
Revise scope to ensure data rescue aims will be met	Reconsider what could be included/excluded in terms of the aim of the exercise.	
Develop a data rescue plan	Prepare a detailed plan, listing what will be rescued, when and by whom, and include plans for ongoing storage of data and records that are not going to be rescued.	

2.3 **Prioritizing data to be rescued**

Most data rescue projects will not have sufficient resources to move all data at risk to an optimal storage system. Often the resources available will be a small fraction of those required. Hence, prioritization of the data to be rescued will be a major consideration of each project. When prioritizing data to be rescued, the following aspects should be considered:

- Institutional, national and international importance of data;
- Operational importance (for example, for use in the operation of a reservoir or key flood monitoring site);
- Scientific importance;
- Uniqueness of records;
- Likelihood of data being used;
- Risk of physical loss of records;
- Risk of loss of knowledge about the data or their accuracy;
- Impact of loss of records.

When assessing which data should be rescued, it is important to consider the main priorities in terms of:

- Monitoring sites;
- Hydrological variables;
- Data formats at risk.

The sections below provide examples of how this might be done for surface water discharge data.

2.3.1 ***Prioritizing river gauging stations***

When rescuing data for surface water gauging stations, it is very important to ensure that data from the most important stations are selected. The prioritization and selection of gauging stations is not straightforward, as the resulting data will be used for a number of different purposes, which may require information from different types of station. Priorities are likely to vary significantly depending on local circumstances, and will be informed by an understanding of the current availability of data and the most pressing needs for further information. Below is a list of criteria, in a suggested order of priority (most important first), that should be considered when selecting stations for data rescue:

1. Length of record;
2. National operational importance of river/watercourse (for water resources, flood risk and so forth);
3. Ability to provide information on hydrological change (for example, natural flow regime);
4. Ability to provide information to improve hydrological understanding, for example, on the representativeness of a hydrological regime within the country including how to assess this representativeness;
5. Perceived quality of gauging station (data);
6. Perceived future potential in respect of 2, 3 and 4 above.

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

2.3.2 ***Prioritizing surface water records***

There will probably be several different types of record relating to river flows at a given gauging station, and there will again be a need to prioritize what to rescue when resources are limited. The list below suggests priorities (most important first) in selecting these records:

1. Processed river flows (for example, daily mean flows);
2. Summarized river flows (for example, monthly mean flows and annual maxima) and previous reports and analyses of hydrological data;
3. Key gauging station metadata and additional metadata concerning data quality, such as a description of hydraulic structures in place, if any, and major influences within the catchment that facilitate the interpretation of data;
4. Stage records;
5. Instantaneous discharge measurements (gaugings);
6. Stage-discharge relationships (ratings);
7. River cross-sections;
8. Other gauging station metadata such as site surveys and photographs.

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

2.3.3 **Choosing the format**

The specific priorities for data rescue will be dictated by local conditions, but the choice of format should be determined by the risk of losing the data:

1. Deteriorating paper records → possible action: improved storage;
2. Obsolete electronic formats → possible action: converting data to readable electronic formats, which may require migration to a different physical medium, such as tape or disk, and/or storage format;
3. Other paper records → possible action: digitization to image or, where possible, to electronic values;
4. Electronic data (for example, spread sheets) → possible action: storage in a hydrological database.

2.3.4 **Prioritizing other records**

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

- Rainfall: Data is often widely available and rescue of a long record of rainfall data would be very important. A scarcity of rainfall records would make data rescue a priority. Confirmatory records in data sparse areas or periods also have great value;
- Snow: Records of snowfall and snow depth are of crucial importance in some regions and are often very hard to come by;
- Evaporation: Measurements of actual evaporation could be extremely useful to future hydrological studies;
- Automatic Weather Station (AWS): A full and good-quality meteorological dataset from an AWS can allow the estimation of potential evaporation and would be considered a useful record should similar datasets not be widely available;
- Meteorological station metadata: Additional information about widely used meteorological stations helps us understand the quality of the data. These metadata may include exact location and any changes over time, period of record and variables measured, but also information on site exposure, types of instruments used, procedures for reading instruments, and so on;
- Thematic data: Maps or spatial data can inform understanding of hydrological processes and are therefore important in hydrological studies. Thematic data could include soil maps, geological maps, land cover information or even cartographic maps containing the elevation and locations of rivers. It is possible that National Hydrological Services store such maps, which are no longer available elsewhere, and these would be a high priority for data rescue.

3. **PRACTICAL TIPS FOR DATA RESCUE**

Practical data rescue advice will change with changing technologies. This section provides tips that focus on different types of media and data rescue techniques. Much of the advice is gathered from international standards and from freely accessible sources from a number of countries. It is likely that other useful advice is available in a variety of languages. Links are provided throughout this publication and are listed in the reference section. It is recommended that the websites of the

organizations providing advice in this area be regularly checked for updates. Where possible, documents have been named so that, in the event that the links become broken, it should be possible to search for them.

It should be noted that data, documents and digital media could be damaged in the process of data rescue itself. This risk should be carefully assessed prior to commencement of a data rescue project and expert advice sought where there is a potential for damage.

3.1 **Assembling a team and managing a data rescue project**

A key factor in the success of past data rescue projects (WMO 2002, 2006, 2008c) has been shown to be the inclusion of an appropriate mix of technical skills; managerial skills and responsibilities corporate, scientific and technical understanding; and dedication to hard work with attention to detail. This should be carefully considered when embarking on a data rescue project. The size of the team and the level of planning and project management required will depend on the scale of the project. In general, a successful approach to a data rescue project will include:

- Describing and communicating the wider context and long-term benefits of data rescue to those commissioning and undertaking the project;
- Identifying similar projects – in scale, type of data and locality – and using them to understand potential issues and approaches;
- Setting clear, measurable goals;
- Planning the project thoroughly, being prepared to alter plans based on progress and findings;
- Assessing the skills required to undertake the project and ensuring these skills are available; undertaking training where necessary;
- Assigning responsibilities for the overall project, but also for tasks within the project, and making sure individual responsibilities are clear;
- Setting out how the project will be managed, including how issues arising in the course of the project will be raised, discussed and dealt with;
- Defining tasks and working practices in detail, and documenting these in notes that can be referred to by those undertaking the work;
- Defining in advance how the results will be reviewed and assessed, for example, by describing steps for quality control of the outputs of a digitization process;
- Setting up mechanisms for reviewing progress throughout the project;
- At the end of the project, reviewing the results against the original goals, and communicating the successes, lessons learned and problems encountered.

Technical skills can be critical to these types of project, whether in handling old materials, operating digitizing equipment or managing databases and file systems. It is very important that someone with appropriate technical skills and understanding of quality issues has oversight of, and detailed involvement in, all stages of a data rescue project. There are many organizations that may be able to advise on skills requirements and options for training.

3.2 Improving paper storage conditions

Many valuable archives still exist in paper format only. The fact that they have endured is in many ways a testament to an innate security of paper archives. However, there remain many risks and the likely absence of a copy of paper records increases the risk of catastrophic loss. It is strongly recommended that a digitization project be considered for all important, unique paper-based data, and some guidelines for doing this are outlined below. However, it may not be possible to undertake such a digitization project and, even if such a project is feasible, it is advisable to retain the paper archive. This section suggests ways to improve storage conditions for paper-based records.

The risks to paper records are many, including disasters such as fire and flooding, vermin and insects, and deterioration due to atmospheric conditions. In addition, simply lack of security or poor management procedures can cause damage to paper archives.

3.2.1 Storage environment

Basic Preservation, a document published by The British Library Preservation Advisory Centre, sets out general issues to do with the protection of paper documents (http://www.bl.uk/aboutus/stratpolprog/collectioncare/publications/booklets/basic_preservation.pdf).

The US-based Northeast Document Conservation Center (NEDCC) has a series of helpful leaflets on archiving techniques, including one entitled: Temperature, Relative Humidity, Light, and Air Quality: Basic Guidelines for Preservation (<http://www.nedcc.org/free-resources/preservation-leaflets/2.-the-environment/2.1-temperature,-relative-humidity,-light,-and-air-quality-basic-guidelines-for-preservation>).

There are many national and international standards on appropriate conditions for storage of paper documents, for example:

- *ISO 11799:2003 Information and documentation – Document storage requirements for archive and library materials*, which can be bought online from the International Organization for Standardization (ISO) (http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38536);
- *BS 5454:2000 Recommendations for the storage and exhibition of archival documents*, which can be bought from The British Standards Institution (<http://shop.bsigroup.com/en/ProductDetail/?pid=000000000019996363>).

A summary of the British Standard can be found in Environmental management, a document from The National Archives (United Kingdom), available at <http://www.nationalarchives.gov.uk/documents/information-management/environmental-management.pdf>.

BS 5454 sets out that “unsuitable environments damage documents more extensively than any other single factor” and recommends a temperature range of 16–19 °C and a relative humidity level of 45–60%; within these ranges conditions should be stable and fluctuation limited to +/-1 °C in temperature and +/-5% humidity. Poor air quality, lack of air movement and light can also seriously damage paper documents.

This available guidance on tolerances for temperature and relative humidity often defines standards that are unobtainable by National Hydrological Services (NHSs). However, the storage of documents outside of these ranges will contribute to the rapid deterioration of records. If the conditions described above are unobtainable, a digitization programme is recommended.

If records have been stored in an environment that does not meet those conditions, the state of the documents throughout the depository should be thoroughly assessed. An estimate of the

lifetime of the documents, based on the current rate of deterioration (since being deposited) should be made. Specialist input may often be required, and some simple pieces of professional advice could potentially add years to the lifetime of an archive, or save substantial amounts of money when improving archives. The British Library Preservation Advice Centre has a limited list of professional bodies, organizations and other available resources (<http://www.bl.uk/blpac/links.html>). However, expertise in this area is available in many countries. National institutions may find that national libraries and archives can provide free advice and assistance when nationally important records are at stake. Regional bodies with this type of expertise may also exist.

It is, however, possible to improve conditions by taking some simple measures. The UK National Archives document mentioned above sets out some cost-effective approaches that do not require specialist expertise:

The most effective way to achieve a stable environment is to ensure that the storage room, and ideally the building as a whole, promotes stability. Thick walls, few or no windows in the repository, and a solid roof will go some way towards insulating the storage environment from changes in the outside environment. Do not leave doors or windows open, and switch off lights when they are not in use.

Lighting should be positioned at least 50 cm above shelving to avoid localised heating and drying of documents.

Use archival-quality boxes to store documents. Storing documents in boxes not only reduces their exposure to light, but can also provide some protection against sudden fluctuations in humidity or temperature.

Avoid storing documents on the floor or against an external wall because these areas may be damp. Open shelving is preferable because it allows air to circulate.

The Northeast Document Conservation Center also provides advice for improving environments in libraries, archives and similar institutions (see Low Cost/No Cost Improvements in Climate Control, available at <http://www.nedcc.org/free-resources/preservation-leaflets/2.-the-environment/2.6-low-costno-cost-improvements-in-climate-control>).

Additionally, the choice of room can help to mitigate disasters and gradual deterioration. Choose rooms that are naturally shaded, well above ground level, away from building drainage systems and bathrooms or other sources of water (consider also rooms and pipework above). Remove carpets and other materials that can harbour pests. Do not store documents directly on shelves made of uncovered wood as they contain some of the acids found in the paper itself, which can contribute to deterioration.

Mechanical methods of environment control can be used, including fans, heaters, air conditioning and dehumidification systems. All of these could be useful, but cost money to purchase, run and maintain. Whether or not such systems are used, the temperature and humidity (and the fluctuations in these variables) should be monitored carefully so that there is a full understanding of the environment and the potential for damage.

There are many cost-effective options for box storage, including convenient box drawers, which can help to protect documents from the atmosphere. Use of desiccants to reduce humidity may be helpful on a temporary basis, but will not provide long-term



Shelving with document folders

protection without regular intervention, and should be used following expert advice, as the types and volumes used will depend upon conditions and archive requirements.

ISO 16245:2009 Information and documentation – Boxes, file covers and other enclosures, made from cellulosic materials, for storage of paper and parchment documents provides strict and detailed guidelines on storage options. It can be bought online from ISO at http://www.iso.org/iso/catalogue_detail.htm?csnumber=45988.

3.2.2 **Archive management**

The way a paper archive is managed is also important. Restricting or controlling the use of an archive can help to provide stable temperatures and humidity, for example by avoiding leaving doors open and lights on when it is not necessary. In addition, monitoring which documents are taken out and by whom will help to avoid document loss and misplacement. An inventory (even at the simplest level), coupled with a standard and consistent system for ordering stored records, will help find documents and identify any problems with missing records. Ensure that someone has ultimate responsibility for the archive, and that regular checks on its maintenance and proper use are made. Ensure that the importance of the archive and its maintenance is widely recognized and understood.

If improved conditions cannot be provided, or if the deterioration of documents is inevitable despite remediation, a realistic assessment must be made of the impact of the loss of the documents. Even when a paper archive is considered to be the best means for long-term storage, there will be benefits in digitizing all or some of the documents.

Once again, there is a wealth of knowledge on the subject such as the records management information from the UK National Archives (<http://www.nationalarchives.gov.uk/information-management/projects-and-work/records-management-guidance.htm>).

3.3 **Digitization of records**

3.3.1 **Approaches to digitization**

Paper records can be digitized in a number of ways:

- Producing and storing images of the original documents. This can be rapid and, when carefully done, can improve accessibility and security of information, but has a major drawback: it cannot produce computer readable information;
- Producing computer readable documents. This can hugely increase the accessibility and utility of an archive by allowing records to be searched for specific information, and to be used quantitatively;
- Setting up databases. Databases can provide access to data through the use of a structured query language and allow users to retrieve information for a variety of purposes such as bespoke analysis and reporting;
- Storing data with specialized data management and analysis tools. For hydrological data, this method can support quality control processes; temporal aggregations and specific analyses; production of graphs; tools to assist in the production of rating curves; tools for planning fieldwork; and preparation of high-quality reports.

The level at which digitization is performed should match the required usage of the data and should be balanced against the effort required to perform the digitization. The benefits of modern hydrological software in improving access to and understanding of data generally

suggest that hydrological digitization projects should be aiming to digitize at least the highest priority data for storage in a hydrological database system.

3.3.2 **General guidance on digitization projects**

There are many documents that provide guidance on digital data management and digitization projects.

The Guidelines for Digitization Projects, drafted in 2002 by the International Federation of Library Associations and Institutions (IFLA) and the International Council on Archives (ICA) at the request of the United Nations Educational, Scientific and Cultural Organization (UNESCO), highlights all of the important areas of consideration in the digitization process though it lacks specific technical advice. This document does contain, however, useful information for budgeting and managing a digitization project: <http://archive.ifla.org/VII/s19/pubs/digit-guide.pdf>.

JISC Digital Media (United Kingdom) provides a huge range of information on digital data management and usage: <http://www.jiscdigitalmedia.ac.uk/advice/>.

The National Preservation Office of the British Library provides a general advice document entitled Managing the Digitisation of Library, Archive and Museum Materials: <http://www.collectiontrust.org.uk/media/documents/c1/a105/f6/000505.pdf>.

The website of The National Archives (United Kingdom) contains a section called Best practice in records management and transfer, and information reuse: <http://www.nationalarchives.gov.uk/information-management/>.

The Northeast Document Conservation Center provides a series of web-based leaflets on all aspects of archiving and digital information management: <http://www.nedcc.org/free-resources/preservation-leaflets/overview>.

3.4 **Creating and storing images of paper documents**

Between 2010 and 2014, Australia's Queensland State Archives published a comprehensive set of documents to guide digitization projects. This material is set out at a very general level and includes suggested technical specifications (resolutions, bit depths, file formats and compression methods) for digitizing different types of original document. The recommendations are necessarily cautious as the guidelines aim to ensure adequate capture of documents for use in legal processes following destruction of the originals. These documents are an excellent source of advice when planning a digitization project:

Digitisation Disposal Policy Toolkit – Planning for Digitisation

<http://www.archives.qld.gov.au/Recordkeeping/GRKDownloads/Documents/Digitisation-Disposal-Policy-Toolkit-Planning-for-digitisation.pdf>;

Digitisation Disposal Policy Toolkit – Technical Specifications

<http://www.archives.qld.gov.au/Recordkeeping/GRKDownloads/Documents/Digitisation-Disposal-Policy-Toolkit-Technical-Specifications.pdf>;

Digitisation Disposal Policy Toolkit – Quality Assurance Guidance, including checklists and advice on how to check quality

<http://www.archives.qld.gov.au/Recordkeeping/GRKDownloads/Documents/Digitisation-Disposal-Policy-Toolkit-Quality-assurance-guidance.pdf>;

Digitisation Disposal Policy Toolkit – Glossary of digitisation terms

<http://www.archives.qld.gov.au/Recordkeeping/GRKDownloads/Documents/Digitisation-Disposal-Policy-Toolkit-Glossary-of-Digitisation-Terms.pdf>;

Digitisation Disposal Policy Toolkit – Further resources

<http://www.archives.qld.gov.au/Recordkeeping/GRKDownloads/Documents/Digitisation-Disposal-Policy-Toolkit-Further-resources.pdf>.

All of these guides stress some central principles:

- Carefully define the purpose of the exercise and how the digitized records will be used, and ensure the resulting digital media are appropriate for this;
- Plan how the quality of the resulting documents will be assured and set up appropriate procedures. Redigitizing documents due to low-quality results is not a good use of resources;
- Consider the cost savings of purchasing a scanner and optical character recognition software, which avoids the effort of keying in information from images, or having access only to images rather than usable data;
- Research procedures before starting. Test equipment thoroughly and explore all equipment settings to produce the best results. Calibrate equipment against standard colour targets. Record all settings used for digitization so that the process can be reproduced;
- Document progress carefully, including which documents have been digitized, the filenames and locations of the digitized versions, and when the documents were digitized and by whom; this will help to correct any problems that are found later;
- Improve the storage of paper documents as they are digitized. Storing documents in a simple system and adding simple labels will help to identify records when they are subsequently needed;
- TIFF images are widely recommended as a standard open uncompressed file format, though others also could be used. Carefully choosing and sticking to a file name convention is an essential step towards a digital filing system;
- Test all digitized files in the way that they will eventually be used (for example, printed, viewed on screen, and images of charts subsequently converted to data values).

Carefully planning how a digital collection will be managed will help to inform the digitization process. Again, there are several guides on this topic:

Systems for Managing Digital Media Collection

<http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/systems-for-managing-digital-media-collections>;

Choosing a File Format for Digital Still Images

<http://www.jiscdigitalmedia.ac.uk/stillimages/advice/choosing-a-file-format-for-digital-still-images>;

Choosing a Filename

<http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/choosing-a-file-name/>;

Optical Character Recognition

<http://www.jiscdigitalmedia.ac.uk/stillimages/advice/optical-character-recognition-ocr>.

3.4.1 **Low-cost solution for scanning documents**

Copy stands with digital cameras, rather than scanners, can be used to digitize documents by simply holding the camera steady at a fixed distance above the document. Many come with fixed lighting equipment for adjustable and maintainable illumination of documents to improve consistency of digitization. Copy stands' prices range between US\$ 50 and US\$ 1 000.

The following JISC document provides detailed information on the selection and use of copy stands: <http://www.jiscdigitalmedia.ac.uk/toolkit/digitisation-equipment/the-copy-stand>.

There are websites describing how to make similar equipment from readily available raw materials, and providing free scanning software (<http://www.diybookscanner.org>).



Example of a simple copy stand
(Courtesy of speedgraphic.co.uk)

3.5 **Digitizing microfilm and microfiche**

Microfilm and microfiche (collectively microform) are specialist long-term storage media used for storing high-quality photo reproductions of paper documents.

The wide accessibility of digital media, which are much more useful for searching and accessing information and require little storage space, has led to a decline in the popularity of microfilm and microfiche, and their use is no longer considered to be best practice. Within hydrology, microform has often been used to store images of water level charts and precipitation records.

Whilst microform is designed for long-term storage, there are valid reasons for converting it into a digital format:

- Risk of deterioration of the media;
- The storage environment must be carefully managed, as for paper;
- Specialist hardware is required to view the information;
- Access can be time consuming (the film containing the required record has to be identified, mounted on the hardware and scanned before the document can be viewed);
- The information is not digitally available for analysis without further work.



Example of microfilm

Specialist equipment is required to convert microform into digital images and it is often expensive (thousands of United States dollars). Even with this equipment, the process of digitization can be slow, as each image must be centred individually prior to scanning. More automated equipment is available but is even more expensive.



Microfilm scanning equipment

There are many specialist services that convert microform to digital images, but they can be very expensive. Digitization could include optical character recognition (OCR) which may make documents more usable. Digitization of microform images of water level chart records should be undertaken very carefully, ensuring that all the vital metadata written on the chart remain legible. As annotations of charts can vary significantly, it is recommended that the number of samples used for quality control be high and that samples be carefully inspected, in order to detect where information is being lost.

Digitization of water level traces from digital images into a time series of water levels should be accurately investigated, ideally comparing digitizing from the image to digitizing from the original paper record. Inaccuracies can occur due to bends in the paper or the angle of the paper surface at the point of digitizing. However, if the microfilm represents the only surviving copy of a record, this could be an important process yielding valuable data.

The International Environmental Data Rescue Organization (IEDRO) is currently working on software to automatically read traces from scanned images of rainfall charts, which may prove to be a time-saving and cost-effective means of producing water level traces (<http://www.iedro.org/en/datarescue/stripchartdigitizer.html>).

It may be necessary to print the images prior to digitizing the traces, in which case this should be done at a very early stage with a good sample of the whole dataset in order to assess the quality of the images and prepare the microform scanning process.

3.6 **Digitizing charts**

Historic water level or stage data may often be stored on charts from recorders. The chart trace can be digitized and converted into a series of dates, times and water level values representing the variation of water level through time. The resulting dataset can be simply stored in text files, spreadsheets, databases or hydrological data management systems; it can thus be easily accessed and rapidly analysed.

3.6.1 **Before digitizing charts**

Before digitizing water level charts, it is worth considering the value of the data. It may be that the water level data are of value on their own (for example, for understanding historic flood levels, or for lakes and reservoirs), but often they will only be of use if they can be converted to meaningful flow data.

To do this for river sections, one must have a valid rating, defining the relationship between water level (stage) and flow (discharge) or, alternatively, sufficient information concerning instantaneous discharge measurements to develop a rating, or a theoretical rating for a gauging structure (see section 3.7 for guidance on data rescue for gauging and ratings).

Assuming that the water level data are likely to be useful, there are some further checks that may be worthwhile carrying out prior to digitization of the charts.

3.6.2 **Cataloguing station information**

There may be information about the station site that is critical to understanding and use of the stage record. Has the station been located in the same place over time? What is the datum of the station and has this been constant over time? Are all stage measurements relative to the known datum? Are there gauge board check readings available? Has the channel structure been constant over time? Are there cross-sections available? Is there any information about bank erosion or other changes to the river in this area? Is the flow contained at this site? At what stage does the flow overtop the bank? Has the equipment for measuring the stage changed over time?

3.6.3 **Cataloguing available charts**

Try to create an inventory for each station, noting the start and end date of the period for which charts are available, the number of charts for each station and any large gaps (several months) in the record.

3.6.4 **Checking the quality of charts**

For each station, prior to digitization, visually inspect a good sample of the charts to check the quality of the paper and trace, and to note any consistent anomalies that suggest problems with the record or the way the equipment has been managed. Issues may include floats being stuck, causing periods of constant level (though this may be real); charts not being replaced, causing multiple traces which can be difficult to resolve; and charts of an incorrect scale being used. In addition, any consistent annotations to the charts should be noted.

As a minimum, each chart should state the date when it was added and removed, and the stage measured manually at the time the chart was changed, so that the record can be digitized. The level reading, including units, for each chart should be clear. Other information, for example, equipment failures, factors influencing the levels and manual corrections to the level, may also have been noted. During digitization, useful annotations should be extracted to a file of comments (with the date/time).

3.6.5 **Tips for digitizing charts**

The *Guide to Hydrological Practices* (WMO-No. 168), Vol. I, Chapter 9, describes data processing and quality control of hydrological data, much of which is as appropriate to data rescue activities as to contemporary processing. A sensible conservative approach to data processing is described, stressing the need for consistent use of station numbers, annotation of records, development of data quality codes and so on (WMO, 2008a). Section 9.4.2 deals specifically with capturing chart data and section 9.6.2.1 deals with primary processing procedures for water level series data. Below are some general suggestions for digitizing charts.

Number the charts for ease of reference. Keep a list of the charts digitized and note the start and end date of digitization alongside the chart numbers and any other useful information (date/time, equipment used and name of the person digitizing the charts).

Check that the start date and time align with the end date of the previous record. Check that the annotated start stage is equal to the end stage of the previous record.

Understand the implications of accuracy for the resulting data. If a rating exists for the data, the impact on the derived flow of an error in the stage can be calculated, and is a worthwhile exercise in understanding how accurate the digitizing process should be.

Instructions for use of digitizing equipment will vary depending on the make, but some general principles for accurate digitizing include ensuring charts are completely flat, well secured and cannot move during digitization. Digitize the centre line of the trace.

To check the quality of digitization, print the digitized data (ideally at the same scale) and compare it to the original plot. Plot each chart after digitization, looking for problems such as those listed in Box 1. Annotate the list of digitized charts with any issues that might have arisen.

Box 1. Possible problems in water level records

- Blocked intake pipes or stilling wells, which will tend to show up as rounded-off peaks and recessions that are unusually flat;
- Spikes or small numbers of values which are obviously out of context and incorrect due to an unnaturally large change in value between adjacent data points. Such situations can occur, for instance, with errors in the digital values between a sensor and a data logger input port;
- Gaps in the data, for which there should be known explanations, and a note as to the remedial action taken;
- Errors induced by the field staff, such as pumping water into a well to flush it;
- Restrictions to the movement of the float/counterweight system or the encoder (perhaps caused by a cable of incorrect length);
- Vandalism or interference by people or animals;
- Debris caught in the control structure, or other damming or backwater condition.

Source: Guide to Hydrological Practices (WMO-No. 168), Vol. 1, 9.6.2.1

Name the resulting data files in a consistent manner that explicitly links them to the chart from which the data were digitized. Store a copy of the raw digitized values before processing the data.

Resolve any issues with the digitized data at an early stage while original charts are still accessible. Annotate the data files appropriately to indicate changes to the data. Use quality codes, where possible, to identify changes made. If a spreadsheet or similar tabular software is used, adding a column for data quality codes and a column for comments is a useful way to store this information. Some hydrological databases will have useful software for annotating and managing quality codes as well as tools for editing data. If such software is used, the raw data could be imported and edited with the same software.

Some charts may have trace inversions or reversals, where the trace reaches the upper or lower edge of the chart and subsequent changes in the trace are inverted in order to provide space for the trace to continue, effectively creating a mirror image. These should be handled carefully to ensure the reversal is correctly understood and dealt with. A printed copy of the chart should help to clearly show where reversals have occurred. Some hydrological software will have tools for correcting these reversals. They can be fairly easily corrected within a spreadsheet once the stage at which the reversal has occurred has been identified, though care should be taken where the inversion occurs several times.

3.6.6 **Software for chart digitization**

Specialist software for digitizing water level records is available, often as part of a larger database or data handling system. Some examples are given in Table 2.

Table 2. Examples of software for chart digitization

<i>Software</i>	<i>Company</i>	<i>Cost estimate (2012)</i>	<i>Description</i>
HYDRAS	OTT hydrometry	Full version ~ US\$ 6 000	Software for downloading logger files, editing and plotting, which includes a chart-digitizing module for processing inversions and for drift editing.
TIDEDA	NIWA	~ US\$ 2 500	Fully functional hydrological database including digitizing package. It supports a variety of chart types for both water level and rainfall (strip charts, radial strip charts, circular charts, reversing and siphoning). This software also supports GTCO Digipad series, Summagraphics Sketchpad and Microgrid, and Graphtec, Numonics, Kurta and Logitech tools. http://www.niwa.co.nz/software/tideda-time-dependent-data
KiDIGI	KISTERS	~ € 2 500 plus € 470 per year	This software is a stand-alone digitizing package, which is compatible with Kisters' WISKI database software but also works without it. The software supports csv and zrxp file formats including comments and target points, supporting level (stage, temperature, etc.) and aggregated (rain) charts. Reversal points can be performed with the digitiser software. KiDIGI supports Summasketch, Summagrid, and Calcomp (usb and serial), Wacom, TDS, Digicon and Graphtec tools. http://www.kisters.eu/water/software/kidigi-digitalisation.html

More generic digitizing software, enabling connection to digitizing tablets and conversion of traces to a series of numbers, but without specialist hydrological functions, such as handling reversals, is available from around US\$ 200 (2012 estimate) (<http://www.logicgroup.com>, <http://www.calcomp.com>).

Digitizing tablets are available new or reconditioned. Often the main source of technical problems can be the compatibility of tablet, device drivers and digitizing software with computer operating systems (for example, specific versions of Windows) and with the available connections/ports to the computer. Before the equipment is purchased, the compatibility of all hardware and software should be checked.

There are now also numerous free software tools to extract a time series of values from a scanned image of a graph or chart, and they have the potential to perform the function of digitizing tablets and software when these are not available, or when images only (for example, from scanned microfilm) are available. In some cases, a large part of the digitizing can be automated as the software can recognize the trace in the chart.

Before any digitization project starts, the accuracy and quality of the results should be compared to the original information, considering the planned use of the resulting data.

3.7 **Digitizing rating and gauging information**

Often river flows will have been measured at open river sections using the velocity-area method where gaugings have been made on a regular basis, measuring the actual flow and recording this against the water level. These gaugings may be used to create a rating for the station. It will often be the case that ratings have already been made in this way, and that the records of these ratings and the gauging from which they are derived will also have been stored. In this case it is

important that the full gauging history, together with any comments from the field technicians who made the measurements, and the rating curves and equations are also digitized, and the information kept alongside the digitized chart data.

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

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

3.8 **Keying data from paper documents, images or charts**

Manually keying data from documents is an option for digitizing data held on paper records, or in scanned images of records, to an electronically readable format. For textual data, this allows records to be rapidly searched, grouped and analysed. For time series data, it allows graphing and numerical analysis. Once data have been digitized in this way, their preservation should be far simpler, assuming appropriate file formats are used and the sustainability of the storage medium is assessed regularly. The *Guide to Hydrological Practices*, Volume I, section 9.4.1, discusses data processing in relation to key entry of data.

Suggestions for data rescue through manual entry of data include maintaining a list of files/records being processed, with the name of the person entering the data, the station number and start and end date of the record (if appropriate).

Data entry and initial validation should be performed by the same person. Whilst data entry can be performed by non-specialists, it is recommended that an expert explain the requirements, provide adequate training and oversee the work. Selection and assessment of the original records should also be undertaken by an expert, as should the setting up of processing logs and verification of resulting data. If classification or category information is being entered, a controlled list can be set up in most spreadsheet software packages to ensure no erroneous categories are entered. If a spreadsheet is used to capture information, it can be useful to rank data in each column and scan the ranked data in order to find anomalies. If the data are numerical, some simple range checks can be made. Printing a graph of numerical results is often a useful way to identify mistakes.

Quality flags and comments should be used when data values are edited. Periods of missing data should be explicitly flagged as such to ensure erroneous zero values are not included.

3.9 **Recovering data from obsolete digital media and formats**

Historically, a wide variety of digital media have been used to store data and information, gaining popularity from the 1950s.

These include magnetic tapes and disks, widely used for digital backup and storage until the cheap availability of CDs in the late 1990s. These media require more careful storage than paper and so deterioration is often a problem. As new media have become popular, older media have



Boxes of tapes of digital computer data

effectively become obsolete and the hardware and software used to access the information can become difficult to maintain and replace. The formats in which the data was stored may have been proprietary and may now be hard to interpret. All of these factors can result in the data becoming very difficult to recover.

This section provides some general advice on recovering data from obsolete and damaged digital media.

The limitations of digitization can be understood when comparing the task of digitizing paper records (which can at least be read by the human eye) with that of recovering data from damaged or obsolete media. Whilst a paper record digitization project may be time consuming, it can at least be performed gradually and cheaply by keying records into a computer or by carefully photographing paper documents. It will often be impossible to salvage data from damaged or obsolete media without

potentially expensive specialist assistance. However, there is much that can be done to speed up and possibly make the process cheaper.

First of all, any available knowledge about the media, the hardware used to access them, and the computers and software that were used should be obtained from documents or people.

As described in section 2.2 above, an inventory of the media should be obtained along with any information about the contents from labels, boxes or associated documents.

If you have the skills to resurrect old technology, an initial investigation could be made into whether it is still possible to retrieve the data contained in the media. However, expert advice should be obtained before attempting to recover the data, as there is a danger of (further) damaging the old media in the rescue process.

It may well be necessary to call on external experts, whose cost can be reduced if they do not have to spend time investigating the hardware, software or file formats themselves.

Whether the data recovery is performed internally or externally, a great deal of time can be saved if it is known which media may contain useful information. Some media may, in fact, simply contain backups of other media.

If data supports are damaged, it is very likely that specialists will be able to limit the damage and maximize the data recovered.

The task of recovering data from damaged media is often termed "data recovery". The retrieval of data from obsolete media is known as "data archaeology".

Companies providing these services exist worldwide. The cost of the work to be undertaken should be balanced with the cost (and risk) of shipping the media containing the data to be retrieved.

Again there may be institutions at the national level (archives or libraries) that could provide advice or assistance. There may even be other institutions within a country that still use the “obsolete” technologies and could provide facilities for recovering data far more cheaply than specialist companies.

Recovering data from a sample of the total media can give an idea as to the usefulness of processing the whole dataset, and the cost of doing so.

3.10 **Verification of digitized or recovered data**

An essential step in any digitization or data recovery project is quality control. Data rescue projects should have a specific plan for quality control, which should set out quality requirements and allow each digitized record to be verified, quantitatively where possible, against those requirements. The outcome of these checks should be recorded, summarized and communicated to others within the project. Deviations from the accepted quality levels should be addressed, by adjusting methods where possible and redigitizing any substandard records. A mechanism for making decisions about changes to digitization processes should be set out in the project plan.

Analysing data at an early stage can help with the quality assessment of the results. Simple checks on the maximum and minimum, monthly and annual means and totals (where appropriate) and comparisons between different records can help identify major problems. A printed graph of the results is a good means for assessing general data quality, though the plausibility of the absolute values should also be considered.

4. **GOOD PRACTICE FOR HYDROLOGICAL DATA MANAGEMENT**

4.1 **Hydrological databases and data management systems**

There are numerous types of digital storage systems, from flat files to dedicated hydrological databases. These are described in *Guide to Hydrological Practices*, Volume I, section 10.2, as are the principal criteria to consider when investigating these systems. In a data rescue project, the principal consideration will be the security of the resulting digital data and ensuring that the data do not have to be rescued again. However, in order to realize the full benefits of this kind of project, the rescued data should be readily incorporated in existing digital data, visualized, analysed and exported. An institution may already have a system for managing digital data and the most effective way to benefit from the project would be by incorporating the rescued data in the existing system. However, a large data rescue project may present an opportunity to improve digital data management, in which case the Guide would provide useful background information on how this could be achieved (http://www.whycos.org/hwrrp/guide/chapters/english/original/WMO168_Ed2008_Vol_I_Ch10_Up2008_en.pdf)

4.2 **Cataloguing metadata from rescued data**

Whilst hydrological databases and other data management systems will provide the means for storing, handling, analysing and visualizing the time series and numerical data, it is unlikely that they will provide the tools for storing all the metadata available. It may be necessary to capture and store these data outside any system for numerical data.

4.2.1 **Metadata standards for hydrological data**

Metadata describe or give information about other data, and are useful both for understanding the data, their utility and limitations, and for carrying out searches using specific criteria. It has

long been considered important to standardize the metadata captured about environmental data in a consistent way, to allow datasets to be joined and searched together. The development of international standards for metadata in hydrology has been discussed for some time (see Maurer, 2004) with some progress but few results to date.

The most significant progress to date has been the development and uptake of a metadata standard for geographical data, *ISO 19115-1:2014 – Geographic information – Metadata – Part 1: Fundamentals*. This defines a standard way to describe geographic datasets, including accuracy of locations, how the data were derived, how to access the data, access restrictions, contact information and so forth. Such standards are welcome and will help organizations make their geographical data searchable and comparable with other international datasets. All hydrological data have a geographical element, so ISO 19115 could be used to catalogue a hydrometric station network as well as other digital spatial datasets such as those for watersheds and digital river line networks.

However, the intended wide use of ISO 19115 means that it is very general and does not define ways to describe much of the metadata that would ideally be needed for hydrological measurements, such as how readings are taken, regularity and accuracy of measurement and equipment used.

The World Meteorological Organization has developed a core metadata standard (a profile of ISO 19115) for describing meteorological datasets. This contains some useful elements, such as a controlled list for the frequency of data measurements, but it is very general (<http://www.wmo.int/pages/prog/www/metadata/WMO-core-metadata.html>).

More detailed standards are currently being developed by WMO (through the Global Runoff Data Centre) and other organizations such as the Open Geospatial Consortium (OGC) (through its Hydrology Domain Working Group), including WaterML2, based on the OGC Observations and Measurements Standard, which defines monitoring, hydrological and time series concepts in relation to datasets. However, whilst standards for describing hydrological metadata are on the horizon, tools for capturing and searching the metadata (essential for data rescue projects) are still some way off.

4.2.2 **Capturing metadata**

Until such standards and tools become available, a pragmatic approach is for data rescue projects to attempt to capture metadata relating to hydrological data in the most appropriate way possible. If sufficient metadata are captured during the rescue process, it should be possible to adapt the data to fit a standard developed in the future. If metadata are not captured at this point, capturing them later will be much more difficult.

Hydrological database packages generally contain a means to enter and store metadata to some degree. If data are stored using these packages, then the metadata storage options should be used as much as possible. In general these will provide, as a minimum, station information (station numbers, river names, locations, catchment areas, etc.) and details of time series (specific parameters being measured, units of measurement, frequency of measurement, etc.). In general, these should implicitly define the process for deriving time series (for example, by linking a flow series to a water level series and a rating curve), which is important. However, it may not be possible to store other detailed information such as instrumentation, accuracy and location of original records.

It is possible to store metadata in spreadsheets or other files, though they are often not ideal for hierarchical data storage. A relational database could be more appropriate, but the skills for developing and maintaining such a system are not always available. The format of any metadata files should be carefully considered to ensure that metadata can be readily found and, to some extent, searched. The location of the metadata files should be carefully considered.

When deciding which metadata to capture, consider who would want to access the data currently and in the future, and why they would want to access them (for example, technicians wanting to find sites where certain instruments are deployed and hydrologists wanting to find stations with good-quality flood records).

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

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

Reduce text descriptions to a series of simple common elements where possible. For example, instead of storing text descriptions of the way in which data from each station are measured and processed, you could list the common elements for all the stations such as instruments, frequency of measurement, parameters being measured, digitization method and post-processing method. These could be described in series of fields, which would allow the information to be searched and used more readily. The benefits of this approach can be understood when considering, for example, the storage of address information as a single text field or as separate fields for building number, building name, road, town/city, country and postal code. However, it is always better to maintain a text description, if useful information would be lost without it.

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

Consider what someone with no knowledge of the data might want to know, or what an expert with no prior knowledge of the dataset would want to find out (for example, exactly how data were captured and the accuracy of instruments).

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

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

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

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

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

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

Box 2 below contains a list of useful basic metadata.

Box 2. Essential basic metadata

- Latitude and longitude of the measurement location, and geographic datum for those values;
- Name of the organization that collected the data;
- Start and end time of the hydrological record;
- Parameters and units being measured;
- Method used to measure the parameter (for example, flow calculated from water level through rating) and its accuracy where possible (for example, water level measured from staff gauge to nearest 0.5 cm);
- Time interval of the data and the time zone applicable;
- Information on what the data value for each time interval represents (for example, mean over the preceding interval, total over the subsequent interval or instantaneous value);
- How the record was digitized;
- Where the record is stored, when it is not possible to include such metadata with the record.

4.3 **Securing rescued data for the longer term**

An essential consideration of a data rescue project should be to ensure that data do not need rescuing in future. It is unlikely that any single step will produce this outcome, due to the changing nature of technology: as storage media and the means for reading them evolve, it is almost inevitable that older media and readers will become obsolete and hard to access or use. Media will also deteriorate with time. A solution for long-term data security will involve a sensible selection of durable media, software and file formats, and procedures for periodic checks of the current status and accessibility of media, upgrading these where necessary.

Tools based on standards, combined with standard and non-proprietary file formats are likely to last the longest without need for updating. ASCII files should remain readable for a long time.

Bespoke databases and databases accessed through specific packages may become redundant if the software is not kept up to date or if the software company is no longer operating. For this reason, a process for exporting all data to non-proprietary file formats should be set up and used on a regular basis, for example annually, with exported data files stored as a back up to the main database.

Processes for backing up electronic files on a regular basis should be in place. The frequency should be determined by the effort required to insert changes made since the previous backup. It can be useful to have several stages of backup in case of different levels of disaster. A networked copy of data files or databases could be made daily, which would be useful in the case of file deletion or corruption. A weekly or monthly backup could be made to removable media, which could be stored in a fire-safe room or on a different site, in case of serious damage such as flood or fire. It is highly recommended that data be regularly deposited with an international database as this provides an almost fail-safe means for retrieving data in the case of disaster.

The procedure for checking the current security of data should be documented and should be carried out at appropriate intervals (perhaps yearly or every 2–3 years). The procedure should be simple, clear and unambiguous as to what constitutes a risk. Checklists allowing each risk to be assessed are a useful way to document that the procedure has been completed. Any new or developing risks should be added to the checklist.

Responsibility for these checks should rest at an appropriately high level, so that they are not overlooked. A lack of ultimate responsibility for data at an appropriate level is often the reason why data rescue requirements develop.

Data security checks should ensure that any issues are raised and dealt with at the appropriate level.

Verification of media and file formats should also cover deterioration of media and obsolescence of file formats, ensuring that all different file types and formats can be opened and accessed using the most up-to-date hardware available, and that software tools run on the most up-to-date systems. It is not sufficient that data can still be accessed using one or more older computers, as data could be lost should this older equipment fail.

Processes for ensuring the security of physical documents should assess the state of a representative sample of different documents from all locations in the archive and attempt to identify deterioration through time, as a minimum. If possible, atmospheric conditions (temperature and humidity) should be monitored and the reasons for any deviations from acceptable range identified and resolved.

4.4 **Sharing and securing data through storage in an international database**

As described in section 1.6 above, data rescue activities should aspire to meet the aims set out in Resolution 25 (WMO, 1999). In addition to the benefits of improved international cooperation and access to scientific information, there is much to be gained in terms of data security. For example, allowing data to be held by an external organization can remove the risk of catastrophic loss in the event of a disaster. There are international organizations that will store hydrological datasets and maintain them in an accessible format at no cost, and with little effort on the part of the data provider.

The Global Runoff Data Centre (GRDC) operates under the auspices of WMO and stores daily and monthly flow series from currently 9 000 gauging stations in 160 countries. In addition to hosting river flow data, GRDC aims to help earth scientists analyse global climate trends and assess environmental impacts and risks. To support these activities, the Centre makes data available for non-commercial research purposes. It is recommended that data rescue activities consider the provision of data to GRDC for archiving and access by researchers.

5. **CONCLUSION**

Hydrological data rescue can contribute hugely to our awareness of the hydrological cycle at a global and local level, including in the areas of flooding, water resources and climate change assessment. It is an increasingly important issue that warrants an effort commensurate with the benefits it can bring.

Risks to data through deterioration of records, obsolescence of formats and natural disasters are great and increase with time but, once rescued, hydrological data can be widely used and their security relatively easily provided.

Special care should be taken during data rescue to avoid further damaging fragile records. In addition, badly digitized documents may be a poor replacement for a paper archive. Careful consideration of each step is essential to a successful data rescue project throughout the planning and implementation stages, including quality control. It may be necessary to obtain outside guidance from experts and to provide training for those doing the work.

The work involved in data rescue can be painstaking and complex, but there is a wealth of guidance material freely available, some of which is referenced here, which can help to inform the approach taken. There are also many lessons to be learnt from previous projects, and it is recommended that the references be carefully consulted before a data rescue project is begun.

The results of a data rescue project should be shared through an international database to foster a wider understanding of the hydrological cycle. The experiences of a data rescue project should be documented and shared. In doing so, the huge benefits of the activities undertaken and the potential impacts of the newly available data can be used to demonstrate the worth of the project and advance our understanding of good practice in data rescue.

APPENDIX A. EXAMPLES OF PREVIOUS DATA RESCUE PROJECTS

1. **Securing meteorological data through microfilming in Africa**

The World Meteorological Organization, through its World Climate Programme, first started data rescue in 1979 through its Regional Association I (RA I) data bank project, with funding from the Belgian government. Between 1979 and 1988, over 1 million meteorological documents from nine member countries of the Permanent Inter-State Committee on Drought Control in the Sahel (Comité permanent Inter-Etats de Lutte contre la Sécheresse dans le Sahel (CILSS)) were microfilmed and microfiche backups stored in the participating countries, at the AGRHYMET Regional Centre in Niamey and at the International Data Rescue Coordination Centre in Brussels. Resources for both personnel and equipment were provided by the project.

The subsequent Data Rescue for RA I (DARE I) project was set up to address data rescue in the non-CILSS countries in Africa, and ran from 1989 to 1996. This provided equipment for microfilming data, and producing and reading microfiche, to each country involved. Staff from the DARE I project visited each country to help to prioritize stations for microfilming, based on length of record and spatial coverage. A minimum gauge density was specified. Training was also provided during the visits. Country staff were then left to complete the acquisition of data, usually involving the microfilming of monthly meteorological tables. The project resulted in about 3.5 million paper records being rescued onto microfilm, with microfiche copies stored in several locations, and CDs of all images sent to each country involved. Annex 4 of the Report of the International Data Rescue Meeting (Geneva, 11–13 September 2001) (WCDMP-49) (WMO, 2002) describes the data rescue process involved in these two projects, listing problems encountered and lessons learned.

The widely acknowledged success of the DARE I project makes the following lessons invaluable to further data rescue projects:

- The importance of well prepared project documents, clear priorities and guidelines;
- The need for a flexible approach to data rescue work as well as a formal approach to defining the project and requiring regular activity reports;
- The use of a local institute with a regional and academic focus (in this case the African Centre of Meteorological Application for Development) for the archiving of documents eased fears over data security;
- Rescuing a limited subset of national data encouraged participation – a comprehensive digitization could have been deemed acquisitive and rejected by some countries;
- The importance of cataloguing all stations (particularly those not digitized);
- Capture on microfilm, whilst inferior to complete digitization, provided the speed required to rescue documents before they deteriorated beyond recovery;
- Training is essential to the successful rescue of documents;
- A mixture of local and international staff improved the efficacy of training.

A further, recent review of the project (WMO, 2008b) highlights the potential loss of the data rescued due to the microfiche medium becoming obsolete, and recommends that this be upgraded.

2. **Climate data digitization in the Netherlands**

From 1981 to 1987, the Royal Netherlands Meteorological Institute digitized a huge amount of pre-1850 meteorological data held in its archives through a project partly financed by the European Union (WMO, 2008c). In total, 0.4 million sub-daily records were digitized. A subsequent project has digitized over 5 million further records to date. Lessons from this work include warnings concerning the limitations of optical character recognition and speech recognition software to increase the speed of digitization as well as a valuable list of “dos” and “don’ts”:

- Start with an inventory of historical data;
- Check if data are already available somewhere in digital form;
- A professional should check the data source before having it digitized;
- Gather all relevant metadata and supply them together with the digitized data;
- Use university students for typing in your data;
- Rigorously quality control your data;
- For many data sources, it is worthwhile supplying the image of the original (keep images of the original data alongside datasets in databases, for reference)
- Use operational infrastructure for central archiving (including backup) or digitized data and images;
- Assess the sustainability of the storage medium;
- Put the data freely on the internet and provide world databases with a copy (this adds to the value of data, publicises project and encourages validation).

Information from Brandsma T., 2008: Data rescue and digitization: tips and tricks resulting from the Dutch experience; Brandsma T., 2010: Personal communication; and World Meteorological Organization, 2008: *Proceedings of the International Workshop on Rescue and Digitization of Climate Records in the Mediterranean Basin* (WMO/TD-No. 1432).

3. **Hydrological data rescue and training in Africa**

In 1998, a WMO hydrological data rescue pilot project (WMO, 2006) was initiated to address concerns over continuing storage of hydrological data on paper records in several countries in RA I (Africa). Following identification of areas at risk, eleven countries were provided with a hydrological database plus 10–15 days training by a consultant from the region. Between 2000 and 2005, three French-speaking countries, predominantly in West Africa (Chad, Niger and Togo) were provided with the French HYDROM software from the Research Institute for Development (Institut de recherche pour le développement (IRD)) and eight (Egypt, Eritrea, Gambia, Ghana, Kenya, Nigeria, Rwanda and United Republic of Tanzania) were provided with the HYDATA software from the Centre for Ecology and Hydrology (United Kingdom). The training consultant for the HYDATA courses was from the Institute for Meteorological Training and Research in Nairobi, himself trained under a WMO project. In addition, each participating country was provided with a PC, printer and scanner to enable data rescue.

The aim of the project was mainly the preservation of hydrological data (locally) and also the enhancement of capacity in the region. No incentive to share data regionally or internationally was provided and data remained within the participating countries. However, the potential data rescue benefits of the projects were huge, given the low total cost of US\$ 100 000, although the

initial impact can tend to decrease over time depending on continued availability of resources, a point that was recognized in the recommendations of the project review (WMO, 2006). Other problems encountered included the lack of time for training and the lack of computing facilities. Unfortunately, the reports on the country visits make no mention of the state of archives or of further data rescue requirements beyond those related to the software training provided. The lack of documentation on the state of national archives is an ongoing limitation to the effective targeting of hydrological data rescue activities.

APPENDIX B. HYDROLOGICAL DATA RESCUE: THE CURRENT STATE OF AFFAIRS

Matthew Fry*

Introduction

Historical hydrological data records are important at a national and international level for process understanding; water resources management and modelling; climate change detection and climate modelling; flood modelling and prediction; and other hydrological and engineering activities. The loss of data can have a significant impact on the ability to undertake these activities and reduces the quality of results obtained from them. In addition, the loss of descriptive station information can mean that rating curves – sensitive relationships between water level and flow – and the quality of flow records cannot be reassessed, and that the impact of changes on the catchment over the period of the record cannot be understood.

Hydrological data are costly, in terms of both effort and resources, to record and collect. Despite this, large volumes of data are lost due to inadequate archiving of collected data and poor maintenance of the archives. A decline in the monitoring of the Earth's hydrology in the late twentieth century has been widely documented (World Bank, 1993; WMO, 1996; Giles, 2005). This has been seen in the neglect and abandonment of stations, reductions in budgets for field maintenance and inspection, and insufficient discharge measurements being made to adequately define rating curves. The result is a significantly reduced coverage of the river monitoring network, and a reduction in the network of stations with long records (Vorösmarty, 2002).

In these circumstances, there would appear to be a need to maximize the usage and impact of the hydrological data that has been captured in the past. There is also an increasing need for good quality hydrological data and contextual metadata within the international community, for the detection of potential climate change signals in rivers and for the assessment, or calibration, of hydrological models to link to ever-improving Global Climate Models. There is, however, anecdotal evidence of a reduction in the budgets for, or efforts towards, the management of hydrological data archives and databases in some countries. It is vital to understand the level of hydrological data at risk and to direct efforts towards safeguarding the most important data at the greatest risk.

Data rescue initiatives

To this end there have been several data rescue initiatives (often termed DARE). Data rescue is the process of securing data that might be lost through natural hazard, degradation or redundancy of storage medium, and of providing access to data through digitization and computerization. Data rescue has been a major focus in meteorology for several decades, recently driven by the need for higher-quality and further-reaching "reconstructions" of past climates for climate change detection and climate modelling. The World Meteorological Organization started data rescue in 1979 with a project, supported by the Belgian Government, that successfully digitized over one million meteorological documents in northern and western Africa (WMO, 2002). The Atmospheric Circulation Reconstructions over the Earth (ACRE) project (www.met-acre.org) has focussed largely on the identification and digitization of historic ship's logs in order to introduce valuable data from periods and locations whose data coverage is poor. These are two examples of large-scale data rescue projects. Many data rescue initiatives are funded within individual National Meteorological Services, but provide the resulting digitized data to the international community.

* Centre for Ecology and Hydrology, Maclean Building, Crowmarsh Gifford, Wallingford, Oxon, OX10 8BB, United Kingdom, mfry@ceh.ac.uk

Hydrological data rescue

The progress achieved in meteorological data rescue has not been matched in hydrology. The Flow Regimes from International Experimental and Network Data (FRIEND) programme has not undertaken any specific hydrological data rescue activities, though it has collated regional databases such as the European Water Archive, Southern Africa FRIEND river flow and spatial databases, for research purposes (Servat and Demuth, 2006). These databases constitute substantial archives of secured data, available for international research. The Global Runoff Data Centre (GRDC) has been hugely successful in accumulating river flow data and obtaining recognition from many countries of the need to share data. Many national and international projects have succeeded in providing capacity to manage hydrological data electronically. But defining hydrological data quality is much more complex than defining meteorological data, and the rescue and provision of river flow time series without data quality information is not sufficient to preserve the data. Precipitation or temperature measurements, with adequate validation against data from nearby locations, can generally be taken as accurate. But river flow data have numerous sources of potential error, for example instrumentation, measurements of cross-sectional area and rating equations, hence they can vary widely. An accuracy of 5–10% can typically be achieved under good conditions (Hirsch and Costa, 2004), though it is often not possible to obtain. Therefore, an understanding of the factors that influence data quality is absolutely essential when utilising the data. In addition, there are anthropogenic factors that can affect catchment runoff, such as reservoirs, increased abstraction or discharge. An understanding of these factors, and how they have changed over the period of record, is also essential when using river flow data for almost any purpose.

However, neither data quality information, nor data on human impacts, are generally available internationally, and are rarely stored alongside the river flow time series data at national level. In addition, there is a perception that the digitization of national archives has only been undertaken piecemeal, with records from many, generally less operational, stations remaining in some earlier medium.

National-level need for data rescue

However, as stated, current knowledge of both the level of data loss and the need for more intensive hydrological data rescue is based on anecdotal evidence. Between 2006 and 2008, the WMO Commission for Hydrology (CHy) surveyed its member NHSs with the aim of producing a clearer picture of the extent of the hydrological data rescue problem. A series of questions was asked concerning the need for data rescue, the nature of data rescue required and previous data rescue activities. The findings of the survey are summarized below.

Out of 183 Members, 58 responses were received from NHSs in 56 countries (30%). This was considered a good response rate for such a survey, and broadly indicative of the need globally. The nature of this type of survey means that countries requiring data rescue assistance could have been more active in responding than those with no need for data rescue. This paper aims to describe the need for hydrological data rescue on the basis of the responses given, but does not attempt to scale the numbers up to reflect total global demand. Table 1 shows the breakdown of responses by WMO Regional Association (RA), and the number of countries specifying a need for data rescue.

Table 1. Data rescue survey responses by WMO Regional Association

	<i>RA I Africa</i>	<i>RA II Asia</i>	<i>RA III South America</i>	<i>RA IV North America, Central America and the Caribbean</i>	<i>RA V South-West Pacific</i>	<i>RA VI Europe</i>	<i>Total</i>
Responses (countries)	16	4	5	6	10	15	56
Countries in need of data rescue	14	2	5	4	4	8	37
Countries requiring urgent data rescue	8	0	2	4	2	3	19

The need for data rescue did not correlate with a country's income. Many developed countries, in Europe and North America, where hydrological data management systems are known to be well developed and resourced, identified data rescue needs.

Experience of previous data rescue activities did not seem to determine whether data were secure. The survey revealed that 11 of the 19 countries that had not declared a need for data rescue had undergone previous successful data rescue projects, all of which were funded and undertaken internally. A similar proportion (21 out of 37) of countries in need of data rescue had experienced previous data rescue activities, 16 of which had been entirely internally funded. However, those countries with no current need generally described these past projects as involving comprehensive digitization of paper records, whereas those declaring a need described a partial digitization of records, involving principal stations only. Descriptions of the unsuccessful projects showed a number of reasons for failure: limited funds and personnel; equipment failure; inappropriate database capabilities; problems with conversion; loss of raw data; lack of space for storage of paper records; and data stored on redundant media.

Usage of a range of hydrological database management systems (DBMSs) was reported. All of the countries that did not need data rescue used a hydrological DBMS for managing data, and all new data was entered into the system; 84% of those countries stored gauging station metadata within the system. Of those countries in need of data rescue, 84% used a hydrological DBMS and 97% of these entered all new data into the system, while only 61% of them stored gauging station metadata within the system. A wide range of tools were mentioned, many being proprietary ones developed within the NHS, though the tool used was not related to the need for data rescue.

Figure 1 shows the volumes of data described as needing rescue, by region. These numbers can only be seen as broadly indicative, as the information provided was inconsistent and in some cases incomplete. They suggest the volumes of data requiring rescue within Europe are far higher, but this is presumably because of the higher density of river monitoring networks in this region historically.

The survey requested information concerning the type of data rescue required. Table 2 summarizes the results from the 39 NHSs requiring data rescue.

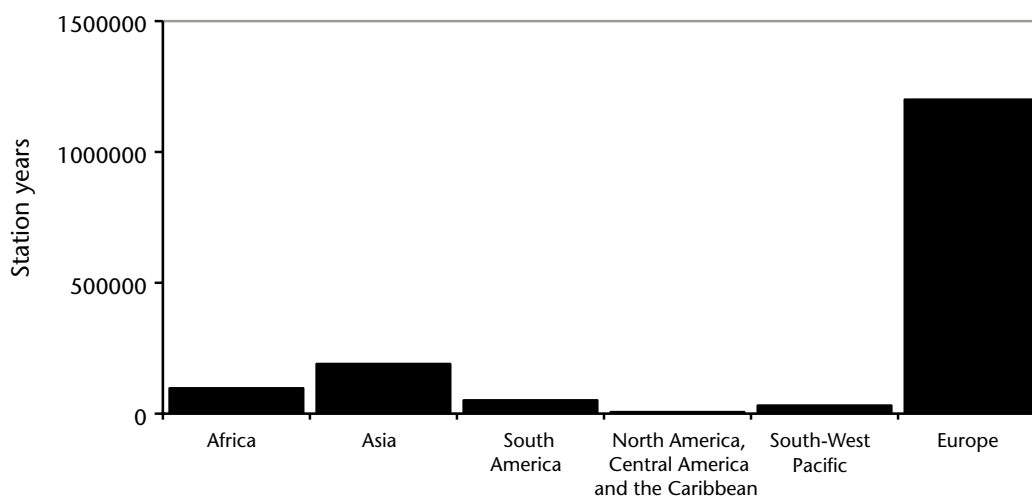


Figure 1. Volumes of data in need of rescue by region

Table 2. Summary of data rescue needs

Question	Details	Number of countries	Percentage of the 37 countries in need of data rescue
Is the volume of data in need of rescue increasing?		26	70%
What types of data are in need of rescue?	Rainfall	29	78%
	Water level	34	92%
	Discharge	31	84%
	Snow	11	30%
	Groundwater	22	59%
	Water quality	26	70%
	Metadata	32	86%
	Rating	26	70%
What are the reasons for data rescue need?	Deteriorating media	32	86%
	No hardware	16	43%
In what storage media are the data in need of rescue?	Paper chart	31	84%
	Paper manuscript	29	78%
	Paper tape	3	8%
	Floppy disk	14	38%
	Magnetic tape	9	24%
	Other	10	27%
Is the data in need of rescue in one location?		18	49%
Is there a catalogue/inventory of the data in need of rescue?		18	49%

These results reveal a palpable data rescue need internationally. Whilst the scale of the problem is large, with key datasets such as rainfall and river flows at risk in many countries, the availability of inventories is encouraging. The potential benefits to the international scientific community of these data are large, and there must be significant possibilities for improving understanding of river flow data through the capture of station metadata and rating information.

The variety of media requiring data rescue represents a significant challenge and could be used as an indicator of the urgency of projects. The window for retrieving data from redundant electronic formats, such as magnetic tape and floppy disk, is limited and could provide a focus for efforts.

Paper records do not represent a large risk in themselves, but deterioration can be rapid. Unfortunately, the survey responses regarding the deterioration of media did not differentiate

between records held on paper and those on magnetic media. However, additional information regarding the urgency of the data rescue need was requested and indicated a variety of specific issues: paper records at risk due to rodent and termite attack and damage from humidity; potential for loss due to fire; and lack of electronic backup facilities. Nineteen of the 37 countries described the need as urgent, due to risk of data loss. Others suggested that the need was being gradually met, or that additional information was required, for example for policy developments, climate change studies and flood modelling.

The data rescue survey was considered successful. The number of responses was high for this type of survey and responses were elicited from a broad range of countries, highlighting the severity of the issue in even the most developed nations. The responses provide good information both for increasing our understanding of the international data rescue need and for targeting assistance to meet this need. The urgency of data rescue requirements varies, but there is a very clear message concerning the urgent need to rescue documents at risk in 25% of countries responding to the survey. Follow-up work attempting to access data inventories and to catalogue the specific need in countries describing a data rescue requirement would be an appropriate step to obtain more detailed information about the scale of effort required to tackle the problem.

International need for hydrological data rescue

The survey findings outlined above clearly describe the requirements for data rescue at a national level. However, the importance to the international community of preserving hydrological data is often not understood. Some of the potential benefits could be improved spatial coverage, increased length of record and better access to gauging station metadata.

Figure 2 shows the number of station years of daily flow data within the GRDC database for countries describing a need for data rescue. Whilst many such countries are well represented in the GRDC database, for 9 of them there are 5 or less station years of daily flow data available, and 16 have 10 years or less. Data rescue in these countries alone, assuming agreement concerning data sharing, would lead to a major improvement in the international data available. The area of those 9 countries, currently represented in the GRDC by 19 station years of daily flow data, totals 13.1 million km².

In addition to the benefits of increasing areal coverage of flow measurement, there is potential for increasing the length of records available. However, the hydrological data rescue survey did not request information regarding specific stations or lengths of record so these cannot be quantified on the basis of the information currently available.

There is an international demand for gauging station metadata. Lack of information pertaining to data quality and anthropogenic impacts fundamentally undermines the use of river flow data for many applications. Svensson et al. (2005) described flood and low flow trend analysis in data from 21 carefully selected GRDC flow series, but pointed out that it was not possible to assess their

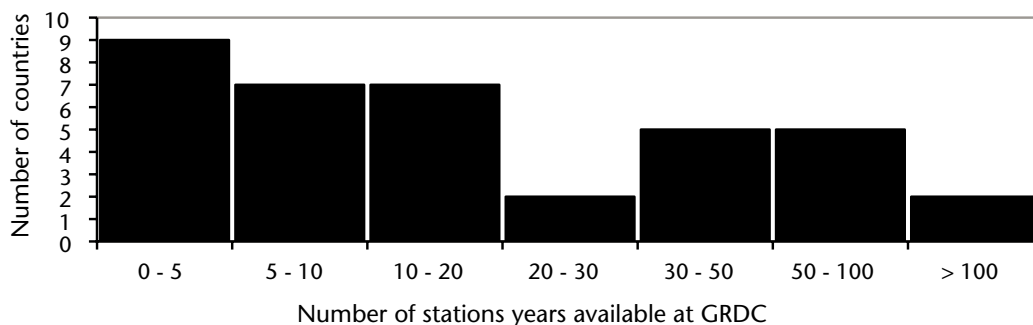


Figure 2. Data currently available internationally for countries declaring a data rescue need

suitability for identifying changes to the stations or within the catchments due to lack of information in the data centre. A means for capturing this metadata from countries providing flow data is essential. To this end, the GRDC has produced a metadata standard (Dornblut, 2009), based on common international standards for monitoring from the Open Geospatial Consortium (OGC). The GRDC metadata standard provides the means for describing metadata including catchment information, data processing steps and data quality information. Such standards are welcome but they are reliant upon NHSs capturing the information and, specifically, having the ability to ensure their data can meet these standards. Appropriate tools for NHSs, adequate support and encouragement for providing, in particular, rating quality information, could enhance the utility of a dataset such as that of the GRDC for a range of applications, including more accurate understanding of climate change impacts on river flows. The FRIEND programme could play a significant role in improving international hydrological databases by taking the lead in gathering these metadata and assessing their utility.

Conclusion

Investigation into the need for hydrological data rescue within NHSs for the purpose of maintaining national capability and supporting international science has shown that:

- There is a huge need for hydrological data rescue in most countries around the world;
- A quarter of countries have urgent data rescue requirements;
- The majority of gauging station data in need of rescue is within European NHSs, whilst the most urgent need for data rescue is in the regions of Africa, South America, and North America, Central America and the Caribbean;
- Countries needing data rescue are often poorly represented in international river flow databases;
- Most countries wish to rescue gauging station metadata, which is urgently needed to improve the utility of international river flow databases;
- The FRIEND programme could help by including metadata gathering activities in its work on regional hydrological databases.

Acknowledgements

The Hydrological Data Rescue survey was coordinated by the World Meteorological Organization, with the input of Ross James at the Australian Bureau of Meteorology.

REFERENCES

- Bradford, R.B and T.J. Marsh, 2003: Defining a network of benchmark catchments for the UK. In: *Proceedings of the ICE - Water and Maritime Engineering, Volume 156, Issue 2*, 109 –116.
- Brandsma T., 2008: Data rescue and digitization: tips and tricks resulting from the Dutch experience. In: *Proceedings of the International Workshop on Rescue and Digitization of Climate Records in the Mediterranean Basin* (M. Brunet and F. Kuglitsch, eds.) (WMO/TD-No. 1432). Geneva.
- Brandsma, T., 2010: Personal communication.
- Dornblut, I., 2009: *Hydrologic Information – Metadata: Semantic structure for the description of hydrologic data (GRDC Hydrologic Metadata)*. Global Runoff Data Centre Report 39r2. Koblenz, Federal Institute of Hydrology.
- Fry, M., 2010: Hydrological Data Rescue – The current state of affairs. In: *Global Change: Facing Risks and Threats to Water Resources* (E. Servat, S. Demuth, A. Dezetter and T. Daniell, eds.), IAHS 340.
- Giles, J., 2005: Solving Africa's climate data problem. *Nature*, 435, 863.
- Hirsch, R. M. and J. E. Costa, 2004: U.S. streamflow measurement and data dissemination improve. *EOS*, 85(21).
- Maurer, T., 2004: *Globally agreed standards for metadata and data on variables describing geophysical processes. A fundamental prerequisite to improve the management of the Earth System for our all future*. Global Runoff Data Centre Report 31. Koblenz, Federal Institute of Hydrology.
- Servat, E. and S. Demuth (eds), 2006: *FRIEND – A Global Perspective 2002–2006*. Koblenz.
- Svensson, C., W.Z Kundzewicz and T. Maurer, 2005: Trend detection in river flow series: 2. Flood and low-flow index series/Détection de tendance dans des séries de débit fluvial: 2. Séries d'indices de crue et d'étiage. *Hydrological Sciences Journal*, 50(5):811–824, DOI: 10.1623/hysj.2005.50.5.811.
- Vorösmarty, C. J., 2002: Global water assessment and potential contributions from Earth systems science. *Aquatic Sciences*, 63(4):328–351.
- Whitfield, P.H., D.H. Burn, J.A. Hannaford, H. Higgins, G.A. Hodgkins, T. Marsh and U. Looser, 2012: Reference hydrologic networks I. The status and potential future directions of national reference hydrologic networks for detecting trends. *Hydrological Sciences Journal*, 57(8), DOI: 10.1080/02626667.2012.728706.
- World Bank, 1993: *Sub-Saharan Africa hydrological assessment: project summary document*. Washington DC.
- World Meteorological Organization, 1996: *The adequacy of hydrological networks: a global assessment* (WMO/TD- No. 740). Geneva.
- , 1999: *Abridged Final Report with Resolutions of the Thirteenth World Meteorological Congress* (WMO-No. 902), http://www.wmo.int/pages/prog/hwrrp/documents/Resolution_25.pdf.
- , 2002: *Report of the CLICOM-DARE Workshop* (July 2000) and *Report of the International Data Rescue Meeting* (September 2001) (WMO/TD-No. 1128). Geneva.
- , 2006: *Report on Hydrological Data Rescue Pilot Project in Africa*. WMO internal report. Geneva.
- , 2008a: *Guide to Hydrological Practices* (WMO-No. 168), Volume I. Geneva, http://www.whycos.org/hwrrp/guide/chapters/english/original/WMO168_Ed2008_Vol_I_Ch9_Up2008_en.pdf.
- , 2008b: *WCDMP Data Rescue Update – GCOS/WCRP Atmospheric Observation Panel for Climate (AOPC), XIV Session*.
- , 2008c: *Proceedings of the International Workshop on Rescue and Digitization of Climate Records in the Mediterranean Basin* (M. Brunet and F. Kuglitsch, eds.) (WMO/TD-No. 1432). Geneva.

Links to data rescue advice

British Library Preservation Advisory Centre:

<http://www.bl.uk/blpac>

British Standards Institute:

<http://www.bsigroup.co.uk>

International Organization for Standardization:

<http://www.iso.org>

International Environmental Data Rescue Organization:

<http://www.iedro.org>

JISC digitization advice:

<http://www.jiscdigitalmedia.ac.uk/digitisation>

Queensland State Archives (list of publications):

<http://www.archives.qld.gov.au/Recordkeeping/Pages/Publications.aspx>

Northeast Document Conservation Center:

<http://www.nedcc.org/free-resources/preservation-leaflets/overview>

The UK National Archives:

<http://www.nationalarchives.gov.uk/information-management/>

UNESCO Guidelines for digitization projects (draft):

<http://archive.ifla.org/VII/s19/pubs/digit-guide.pdf>

WMO Guide to Hydrological Practices:

<http://www.whycos.org/hwrrp/guide>

For more information, please contact:

World Meteorological Organization

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

Communications and Public Affairs Office

Tel.: +41 (0) 22 730 83 14 – Fax: +41 (0) 22 730 80 27

E-mail: cpa@wmo.int

Hydrology and Water Resources Branch

Climate and Water Department

Tel.: +41 (0) 22 730 84 79 – Fax: +41 (0) 22 730 80 43

E-mail: DHWR@wmo.int

www.wmo.int