IJDC / *Research Paper*

Application and Evaluation of the Hoffman et al. (2020) Data Rescue Framework Using an Historic Scottish Cloud and Rain Chemistry Dataset Exemplar

Shona J Ferguson
UKCEH

J.N. Cape
UKCEH Edinburgh

Alan Crossley
UKCEH Edinburgh

F. Harvey
UKCEH Edinburgh

D. Fowler
UKCEH Edinburgh

D. Leaver
UKCEH Edinburgh

C.F. Braban
UKCEH Edinburgh

Abstract

Environmental data are vitally important and valuable research outputs, and there are vast quantities in laboratory storage and on servers where they are not easily or openly accessible. It is imperative for preservation and for potential reuse purposes that historical data of long-term value are efficiently curated and made publicly available. We evaluated the Hoffman et al. (2020) data rescue framework (DRF) for the initial assessment stage of a data rescue by applying it to an historic Scottish cloud and rain chemistry dataset. The DRF facilitated workload prioritisation, anticipating potential obstacles, and approximating resources required. We used a novel points-based adaptation of the DRF to identify suitability of datasets for rescue and compare the dataset status before and after rescue, particularly taking FAIR principles into account. The reusability of the dataset was greatly improved by the Hoffman et al. (2020) framework, and it is now published in an appropriate open access data centre with detailed metadata. It is recommended that the traceable DRF and scoring system be adopted in future to begin moving twentieth and early twenty-first century environmental data into the public domain.

Submitted 26 July 2024 – *Accepted* 3 March 2025

Correspondence should be addressed to **Shona J Ferguson**. Email: shofer@ceh.ac.uk

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v19i1.1027
DOI: 10.2218/ ijdc.v19i1.1027

Introduction

Data are some of the most vital outputs from research, and, alongside answering research questions, they can open doors to new research topics and hypotheses. There is a wealth of untapped historical data resources held by research institutions globally. Many researchers are unaware that their data can be useful to other researchers, and often have concerns regarding the availability of their data (Tedersoo et al., 2021). In 2011, the magazine *Science* polled its researchers and found that 88.5% of the 1700 responders stored their data either in their laboratories or on university servers, and only 7.6% stored their data with a community repository (Science Staff, 2011). These numbers have improved over the years (Tedersoo et al., 2021), but much of the historical data stored in laboratories or on university servers ten or more years ago is likely still in the same state or lost. When one considers that there were approximately 8.8 million scientists around the world in 2018 (Naujokaityte, 2021), it is difficult to even comprehend the extent of data that are not currently publicly available for reuse.

Researchers reuse existing datasets in a variety of ways to conduct their own research. The primary use for environmental data is to use a data point in space and time that has passed. The research may be unrelated to the original purpose of the data or following on from it. It can be to evaluate methodologies, to verify one's own findings or models, or to increase the temporal extent of research. Data reuse can save time and resources that would otherwise have been spent conducting repeat field research, purchasing or constructing instrumentation, and collating the data. If research institutes and other bodies begin investing in rescuing data that are valuable but not currently reusable, the ongoing benefits of newly reusable data will outweigh the time and effort taken to retrieve them originally.

A high standard of data rescue is essential to ensure that the resulting publicly available data meet the FAIR principles of data (Wilkinson et al., 2016). The FAIR principles are a set of four characteristics (Findable, Accessible, Interoperable, Reusable) that published data should exhibit to be successfully discovered and reused by third parties (Wilkinson et al., 2016).

Using a defined framework or process to rescue data will ensure that all important factors surrounding the data are being considered and that the FAIR principles are met. There have been frameworks in the past based around digital collections.[1] However, a framework for practically undertaking data rescue assessment is needed. One such framework is the Hoffman et al. (2020) framework, which is a list of 18 assessment factors intended to assist data curation professionals in the initial stage of a data rescue, from prioritising next steps to assessing the labour and resources required. Each of the factors is accompanied by detailed guidelines to consider. A short summary of each of the Hoffman et al. (2020) assessment criteria can be seen in Table 1 below. A more comprehensive description can be found in the original paper.

Table 1. A summary of each of the Hoffman et al. (2020) framework assessment factors.

Number	Assessment factor	Description
1	Extent	How large is the collection (characterised in terms of linear feet, number of boxes, number of digital files, digital file size, etc.)? Within the collection, how much data is present? To what extent is the collection or the data within the collection already processed?
2	Data objects	What kinds of data exist in the collection, and in what forms? What file formats or physical materials are the data in?

[1] A Framework of Guidance for Building Good Digital Collections: https://web.archive.org/web/20171201033840/http://framework.niso.org/

3	User communities	What groups of potential users should be able to understand and use the data in this collection?
4	Stakeholders	Who has invested in the data or in the research it supports? Who would be affected by use or reuse of the data?
5	Reuse value	What are the intended, demonstrated, anticipated, or plausible reuse opportunities for the collection?
6	Reusable objects	Are there specific components of the collection that carry reuse opportunities?
7	Historical value	What is the potential historical value of the collection? What important or noteworthy scientific approaches, results, or advances are documented or evidenced by the data?
8	Historical objects	Are there specific components of the collection that carry historical value?
9	Completeness	How complete or incomplete is the collection? In other words, are there gaps in the collection that would limit either reuse or historical value?
10	Sensitivity	Are there aspects of the collection that may be considered sensitive to unintended or undesirable access, use, or interpretations, whether from the standpoint of privacy, ethics, security, or scientific accuracy?
11	Access and use constraints	What constraints will be placed on access to and use of the data?
12	Rarity or uniqueness	Is any part of the collection or data within the collection duplicated elsewhere, or actively stewarded, curated, or maintained by another group or institution? This factor may also be used to address other, distinctive strands of rarity: whether the data are fundamentally irreplaceable, or whether aspects of them could be recreated.
13	Reproducibility	In what ways, if any, are the data within the collection reproducible?
14	Relevant collections	Are there other collections of research materials that are relevant to this collection, and which demonstrate a wider network of interest or investment in the research documented by the collection?
15	Associated publications	Are there identifiable publications associated with the collection, such as scientific journal articles that report, rely on, or cite the data or methods represented by the collection?
16	Fit for purpose	To what extent are the data ready or suitable for actual or potential uses identified in reuse value, historical value, and reproducibility (above)? How much additional documentation, interpretation, and processing are required to prepare data, either for reuse or adequately to serve as historical evidence?
17	Obstacles to recovery	What are the anticipated or observed obstacles to recovering data from the collection?
18	Priorities	What are the most immediate priorities for data recovery, as opposed to the optimal or long-term objectives of recovery?

The datasets considered for rescue in this project are the result of research measurements conducted by scientists at the UK Centre for Ecology and Hydrology (UKCEH) to continuously monitor and model cloud and rainwater chemical composition at high elevation. These measurements began in 1993, after unexpected results from campaigns at the Great Dun Fell field site (GDF campaigns). The GDF campaigns showed that rainfall amount and rainfall composition increase with altitude due to a mechanism known as the 'seeder-feeder' effect (Choularton et al., 1988, Fowler et al., 1988). This is the mechanism by which precipitation droplets from high-level cloud (seeder) above a hill fall through low-level cloud (feeder) collecting cloud water, thus causing greater precipitation on the hill under cap cloud than on nearby flat land (Figure 1; Cape et al., 2010). Measurement sites were set up at Holme Moss and Dunslair Heights for long-term monitoring of these effects. However, in 2003, the Dunslair site was retired, and the measurements relocated to Bowbeat due to decreased capture of hill cloud caused by sheltering from an adjacent forest (Cape et al., 2010). Due to a renewed research activity studying pollution deposition in complex terrain, these data were requested to be rescued.

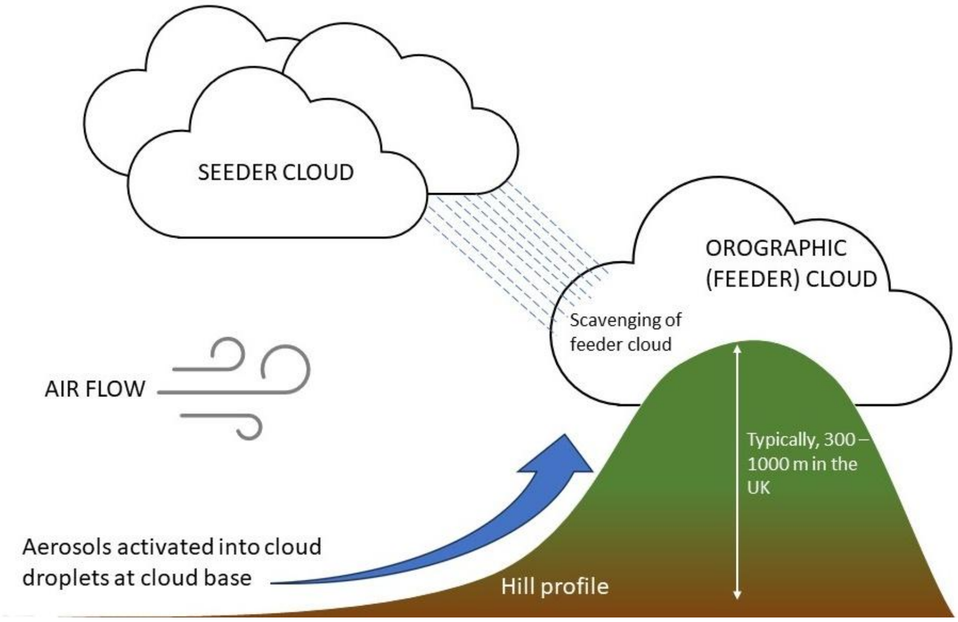
SEEDER CLOUD
OROGRAPHIC (FEEDER) CLOUD
Scavenging of feeder cloud
AIR FLOW
Typically, 300 – 1000 m in the UK
Aerosols activated into cloud droplets at cloud base
Hill profile

Figure 1: Schematic of the seeder-feeder process (adapted from Cape et al., 2010).

As a pilot study, the historical cloud and rain chemistry dataset from one of the field sites mentioned above was selected. Rather than just re-use for one application, the data was assessed for FAIR rescue using the Hoffman et al. (2020) framework, as it is easily repeatable and well-suited for assessing small datasets in a digital tabular format. As part of this, we also evaluate the Hoffman et al. (2020) data rescue framework. Although the framework was originally intended for initial assessment only, we have found that it can be utilised alongside a slight adaptation to successfully curate a dataset from start to finish and that the method used here may be helpful to other future data rescuers.

Methods

Using the Hoffman et al. (2020) Framework to Identify a Suitable Dataset for Rescue

Four pollutant deposition datasets with accessible measurement data on UKCEH servers were identified as being suitable to test the Hoffman et al. (2020) framework (Table 2). These datasets were all created around the 1990s and were the result of research projects involving observations of atmospheric pollutants in complex terrain (see introduction for more detail). They were chosen as they are expected be of long-term value, and they were primarily funded by the UK Natural Environment Research Council (NERC) and/or UK Department for the Environment, Farming and Rural Affairs (Defra) so would be suitable for archive with a NERC data centre, such as the Environmental Information Data Centre (EIDC).

Table 2. Initial datasets identified for potential use to test Hoffman et al. (2020) data rescue framework.

Dataset name	Type	Current location	Data accessible (Y/N)	Metadata accessible (Y/N)
Bowbeat	Cloud/rain	Institute servers	Y	tbd
Dunslair	Pollution deposition experiment	Institute servers	Y	tbd
Holme Moss	Pollution deposition experiment	Institute servers	Y	tbd
Great Dun Fell	Pollution deposition	Institute servers	Y	tbd

The archive files for the datasets were given a preliminary assessment to identify relevant files for the rescue such as datasets, images, location/site information, and metadata. As a result, Dunslair and Bowbeat were chosen for further assessment as their format and volume of data were the most suitable for the six-month timeframe of this project. The version control used in files for both datasets was unclear, so each file needed to be opened to determine which were the most complete and up to date. File formats and sizes were noted for reference, and the potential issues with the files were also recorded.

To decide which dataset to move forward with, the Hoffman et al. (2020) framework was adapted to a points-based system where each assessment factor could be scored out of five based on an initial assessment of how suitable the files were for rescue (please see Appendix 1 for a full list of the grade descriptors created for this project based on the assessment factors). To allow some discrimination between datasets and to assist with prioritisation, a threshold of 60 out of a total of 90 possible points was used to determine if the dataset was “worthwhile” to rescue. When this approach was applied to Dunslair and Bowbeat, Dunslair scored 73/90, and Bowbeat scored 78/90 (Table 3). Despite covering a longer time series, Dunslair scored lower as the spreadsheets containing the measurements included a lot of superfluous data and graphs that would have required a significant amount of processing time to clean. The additional resources and time this processing would have taken were not within the scope of this project, and therefore, Bowbeat was chosen as the most appropriate for this data rescue. More comprehensive notes on the assessment factors where Bowbeat scored higher than Dunslair are found in Table 4.

Table 3. A summary of the outcome of the assessment of Bowbeat and Dunslair datasets using a points-based adaptation of Hoffman et al.'s (2020) data rescue framework.

Number	Assessment factor	Bowbeat	Dunslair
1	Extent	4	3
2	Data objects	4	4
3	User communities	5	5
4	Stakeholders	5	5
5	Reuse value	4	4
6	Reusable objects	5	4
7	Historical value	5	5
8	Historical objects	5	5
9	Completeness	4	3
10	Sensitivity	5	5
11	Access and use constraints	3	3
12	Rarity or uniqueness	5	5
13	Reproducibility	3	3
14	Relevant collections	3	3
15	Associated publications	5	5
16	Fit for purpose	3	3
17	Obstacles to recovery	5	3
18	Priorities	Immediate (5)	Immediate (5)
	Total	78/90	73/90

Table 4. Detailed notes on the Hoffman et al. (2020) framework assessment factors where Dunslair and Bowbeat scored differently.

Number	Assessment factor	Bowbeat	Dunslair
1	Extent	The main dataset suitable for reuse is the chemistry dataset which is 837 kB. The data are not well structured or annotated so will require some processing.	The main dataset suitable for reuse is the chemistry dataset which is 6.12 MB. The data are disorganised with a lot of unexplained data flagging. The spreadsheets are very large and contain data for several sites within the same sheets.
6	Reusable objects	In addition to the chemistry dataset, there are some images of rain and cloud	In addition to the chemistry dataset, there are some meteorological datasets that may be

		collecting equipment, maps of the site, and some description of methodology.	suitable for reuse. There are not many files containing useful metadata to assist with reuse.
9	Completeness	The chemistry dataset has some data gaps where there was no collection. This shouldn't limit the value of the data.	It is difficult to estimate the completeness of this dataset due to its lack of structure and organisation. There are some years with very large periods of missing data with limited explanation.
17	Obstacles to recovery	Obstacles to recovery include somewhat unclear versioning and lack of metadata. These can be resolved through communication with the original researchers and research into associated publications.	Obstacles to recovery include very unclear versioning, unexplained data flagging/data gaps and lack of metadata. These can be resolved through communication with the original researchers and research into associated publications. This dataset would require significant data processing time.

Applying the Hoffman et al. (2020) Framework to the Data Rescue of the Bowbeat Dataset

During the initial assessment of the Bowbeat files, the issues detailed in Table 5 were noted.

Table 5. Issues noted during the initial assessment of the Bowbeat files.

Category	Issue
Spreadsheet formatting	The measurement data were in Excel 97-2003 format (most data centres require CSVs). Layout of spreadsheet was complex and unclear.
Spreadsheet formulae	Some formulae in the spreadsheet were resulting in errors. Formulae were not explained.
Measurement data	Some missing measurement data with no explanation. Most measurement variables and their units were unspecified/undefined.

As per Assessment Factor 18 of the Hoffman et al. (2020) framework, a list of priorities for rescuing the Bowbeat cloud and rain pollutant dataset was written to address the issues found during the initial analysis:

1. Quality-check the dataset (formulae, errors, completeness, notes flagged on cells of spreadsheet).
2. Define the variable names and units used.
3. Check what file format the data should be in (data centre requirements and atmospheric chemistry modellers' requirements).

4. Clean and convert the file to meet the requirements of the modellers and of the data centre.

Missing data were identified during the quality check, however there was no explanation found in the archive files beyond the notes in the spreadsheet stating 'no data collection'. Some formulae errors were corrected easily as they had clearly resulted from accidental changes while filling formulae down columns in Excel. At first glance, the data appeared to have a weekly temporal resolution. However, upon further inspection, not all data collection occurred exactly seven days apart. This was noted for inclusion in the supporting documentation that would be created at the point of publishing the dataset.

The EIDC (Environmental Information Data Centre) was chosen as the most appropriate data centre for the data to be made public. This data centre requires Excel data to be converted to an open, non-proprietary format such as CSV to allow the data to be used in a wide variety of common software tools and to ensure they are robust and future-proof. They also have several guidelines on structure and content of tabular data. The Bowbeat dataset was, therefore, cleaned and quality-checked using an R script, ensuring the validity of the output and meeting the EIDC requirements of high reuse quality. From CSV, data users can transform the data easily into a suitable format for their application.

An atmospheric chemistry researcher at UKCEH was consulted to help identify the measurement variables/units and to verify the formulae. While this was very helpful, there were still some outstanding issues with the dataset and missing metadata, so the decision was made to contact the original researchers involved in the data collection/processing to make some queries and to gather any additional information they could provide.

We were able to contact a researcher who led the original research, and they helped to annotate a map of the field site and were able to provide additional information on the funding of the project. They also provided some additional insight into possible reasons for missing data (e.g., lack of rain/cloud, obvious contamination of sample) and gave thorough reasoning behind the calculation used to derive non-marine sulphate concentration from sulphate and sodium concentrations. They recommended conducting an ion balance check for quality assurance as they noticed some clear discrepancies in the data that they surmised were caused by contamination.

An ion balance was constructed for the rain and cloud/rain chemistry datasets but not for the cloud chemistry as this is a derived dataset. As a result of this, some data had to be removed for high contamination or analytical errors. Data flags were added to the data where there were unusual ion imbalances, and the percentage completeness of the datasets decreased due to the removed data. These changes were made by adapting the R script and regenerating the data files.

To publish the completed dataset, supporting documentation was written containing a detailed overview of the original research and the additional quality control steps taken as part of this rescue. Associated publications were extremely helpful for this document as the historical files contained very little information on the experimental design, collection methods or fieldwork instrumentation.

A Review of the Bowbeat Dataset after Rescue

When the dataset was in its finalised form, the points-based adaptation of the Hoffman et al. (2020) framework was applied to the dataset again to see the improvements made at a glance. Table 6 shows the assessment factor scores before the data rescue versus after the rescue.

Table 6. A comparison of the Bowbeat dataset before and after it was rescued using a points-based adaptation of Hoffman et al.'s (2020) data rescue framework.

Number	Assessment factor	Before rescue	After rescue
1	Extent	4	5
2	Data objects	4	5

3	User communities	5	5
4	Stakeholders	5	5
5	Reuse value	4	5
6	Reusable objects	5	5
7	Historical value	5	5
8	Historical objects	5	5
9	Completeness	4	4
10	Sensitivity	5	5
11	Access and use constraints	3	5
12	Rarity or uniqueness	5	5
13	Reproducibility	3	4
14	Relevant collections	3	3
15	Associated publications	5	4
16	Fit for purpose	3	5
17	Obstacles to recovery	5	5
18	Priorities	Immediate (5)	Immediate (5)
	Total	78/90	85/90

Table 7 contains more detailed notes on each assessment factor and the changes resulting from the data rescue.

Table 7. Detailed notes on the Hoffman et al. (2020) framework assessment factors before versus after the data rescue.

Number	Assessment factor	Before rescue	After rescue
1	Extent	The main dataset suitable for reuse is the chemistry dataset which is 837 kB. The data is not well structured or annotated so will require some processing.	The dataset has been separated into three CSVs, each under 30 kB. They have been fully cleaned, processed, and quality controlled.
2	Data objects	The dataset suitable for reuse is the rain and cloud chemistry dataset but it is not yet in the correct format.	The chemistry dataset has been cleaned and processed and is ready for publication/reuse. They are now in a machine-readable open format.
3	User communities	Atmospheric chemistry modellers and scientists.	No change.
4	Stakeholders	The project was DEFRA-funded. This organisation would not be affected by reuse.	No change.

5	Reuse value	The data could be used to test atmospheric chemistry models.	The supporting documentation generated during the rescue could be used in research into retired field sites.
6	Reusable objects	In addition to the chemistry dataset, there are some images of rain and cloud collecting equipment, maps of the site, and some description of methodology.	No change—the reusable objects were used to create supporting documentation.
7	Historical value	These datasets have important historical value as they were obtained by prominent UKCEH researchers and cannot be recreated for the time periods in which they were taken.	No change.
8	Historical objects	The datasets, images and maps all have good historical value.	No change.
9	Completeness	The chemistry dataset has some data gaps where there was no collection. This shouldn't limit the value of the data.	The ion balance quality control step resulted in the removal of some data. The completeness reduced from approximately 81–89% to 70–83%. This still won't limit the value of the data.
10	Sensitivity	No sensitivity issues.	No change.
11	Access and use constraints	The data cannot be accessed as it is held on UKCEH servers.	The data has been published with the EIDC, where it is freely available with no constraints (Crossley et al., 2023).
12	Rarity or uniqueness	The data is completely unique as there was no other data capture at the Bowbeat site during the same time period as this research. It cannot be recaptured as it is unique to the time period in which it was taken.	No change.
13	Reproducibility	The methods used in obtaining this data could be reproduced, but the specific data cannot be replicated.	The completed supporting documentation accompanying the data can be used to help reproduce the methods used.
14	Relevant collections	There are collections of similar data produced in associated projects at other field sites. These data have not yet been rescued and are stored on UKCEH servers.	No change.

15	Associated publications	The final report to DEFRA directly referenced the Bowbeat chemistry data.	There was very little information on the methods found in this report, so other publications referencing the Dunslair dataset were used as both projects used the same experimental methods.
16	Fit for purpose	The data is not yet fit for purpose as it requires quality checking and processing to be in a suitable state for reuse.	The data has been cleaned and converted to CSV format. It is now fit for purpose.
17	Obstacles to recovery	There are no obstacles to recovery.	No change.
18	Priorities	Immediate (list of priorities found above).	No change.

The whole process from the initial assessment of the datasets to the submission of the data to the EIDC took approximately eight months of intermittent work performed by an early career researcher, including liaison with the original researchers who no longer work at UKCEH and the creation of the repeatable methodology. A full transferrable method for rescuing digital tabular-based data written as part of this research has been included in Appendix 2.

Discussion

Issues Encountered and How They Were Resolved

As described above, there were several issues with the original spreadsheets before any processing had begun. They were complexly structured with several tables within the same sheet, contained missing data with limited explanation, and did not fully describe the variables measured. Layout and formatting issues were easily resolved by creating an R script, and GitHub was used for clear version control and to allow for reuse for the rescue of similar data.

The original researchers responsible for the collection and processing of the datasets provided invaluable contributions to this project, including supporting information about the site and project and a method with which to do a further quality check. Without their assistance, it would have been very difficult to provide complete supporting documentation, and the errors resulting from sample contamination may not have been found.

Associated publications also provided a great deal of supporting information that was not available in the historical files. This is a good example of the importance of keeping a comprehensive record of research methods and data processing steps while conducting research.

Benefits of Using the Hoffman et al. (2020) Assessment Framework

The Hoffman et al. (2020) framework was very helpful in a number of ways during the rescue of the Bowbeat chemistry dataset. It was used to identify the most suitable dataset for rescue, to identify priorities and what steps needed taken, and to clearly see the improvement in the reusability of the dataset after rescue.

For instances such as this where there are several datasets requiring rescue, the framework helps to quickly determine the time and resources required to complete the data rescues and, therefore, not take on more workload than is within the scope of the project.

Although some of the assessment factors do overlap, all of them are essential to encourage data rescuers to consider all aspects of the data. In particular, the framework helped to identify and prioritise the steps required to process the dataset to a state suitable for reuse. Addressing each assessment factor also assisted in flagging issues needing resolved with formatting and completeness.

The point-based adaptation of the original Hoffman et al. (2020) framework used here was developed as a semi-quantitative method of comparing datasets on a surface level and to show the value that the data rescues add to datasets. It is easily repeatable and could be used for any future data rescue projects. Although a threshold was used here for prioritisation between datasets, an alternative, simple rules-based method could have been used that is more nuanced and requires a level of judgement call:

A dataset is seen as 'worthwhile' to rescue if:

- Any assessment factors marked with a * are scored >1 (see Appendix 1);
- Extent of processing requirement identified in Assessment Factors 1 and 16 does not exceed any budget/time limits;
- Reuse and historical value are identified and strong enough to warrant recovery effort;
- Completeness, sensitivity, and access and use constraints do not significantly reduce value of rescuing the data;
- There are sufficient data objects, relevant collections, and associated publications to create fully descriptive metadata and supporting documentation.

Conclusion

Overall, the Bowbeat chemistry data rescue was very successful, and the Hoffman et al. (2020) framework is strongly recommended for future data rescues. It is very effective for prioritising workload, anticipating potential obstacles, and approximating resources required for a data rescue. It also has the benefit with the additional scoring system of having a traceable decision process and information available for future data rescuers. It is, however, important to note that unforeseen barriers can arise at any stage in a data rescue, and so following the framework may not identify every possible issue. For example, the extra quality control steps needed in the Bowbeat data rescue could not have been anticipated before a thorough analysis of the data and a consultation with the original researchers. As a result of this data rescue, the Bowbeat dataset now meets the requirements of the FAIR principles. The dataset and detailed metadata are publicly available on the EIDC website without need for authentication (Crossley et al., 2023), and it is in a well-organised CSV format so is easily readable by both humans and computers. The data is intended for use in testing atmospheric models, and it is hoped that other UKCEH historical files will be rescued in a similar way once required resources are obtained. Fog and rain chemistry remain little studied yet have significant effects on the ecology of uplands, where there are the main routes for nutrient inputs.

It is critical that the importance of timeliness for data rescue is considered alongside the economics of data rescue and immediate science need. Had this data rescue been carried out in ten or 15 years' time, it could have been considerably more difficult. As software develops, old data formats become harder to use and process. It can also become harder over time to get contact details for the original researchers, who may find it difficult to recall specific details from tens of years ago. If data of historical and/or reuse value are rescued earlier, the time and economic resources (i.e., staff time) required can be minimised.

For future datasets, those yet to be made, the need for data rescue should gradually be removed—the Bowbeat data rescue was required, in part, due to a lack of knowledge and awareness of data management practice at the time the data were collected. To make data rescues easier, or even to eradicate the need for data rescues at all, awareness of the Hoffmann Framework at the time of research could be considered by researchers as part of good data management practice during their projects. For example, data managers in research projects could complete a Hoffman Framework assessment for each experiment, including awareness of diverse data use. The support of a Data Stewardship expert can be valuable here. While this practice has improved drastically in recent years, there is still significant progress to be made.

Acknowledgements

This work was funded through the research partnership agreements between the Natural Capital and Ecosystem Assessment programme (NCEA Defra) and the UK Centre for Ecology & Hydrology (UKCEH) Numbers 32156 and 60978, which builds upon work supported by the Natural Environment Research Council Award Number NE/R016429/1 as part of the UK-SCAPE programme delivering National Capability.

Appendices

Appendix 1 – Point Grade Descriptors Used to Grade a Dataset from 1 to 5 for Each Hoffman et al. Assessment Factor

Number	Assessment factor	Point grade descriptors				
		1	2	3	4	5
1	Extent	Collection/ data are large (e.g., several boxes, >1TB, etc.) and completely unprocessed.	Collection/ data are medium (>10 GB and <1TB) and completely unprocessed OR Collection/data are large (>1TB) and partially processed.	Collection/ data are small and completely unprocessed OR Collection/data are medium and partially processed OR Collection/data are large and almost fully processed.	Collection/ data are small and partially processed OR Collection/data are medium and almost fully processed.	Collection/ data are fully processed.
2*	Data objects	Data are in proprietary, not commonly used, or physical formats that cannot be read or converted.	Data are in proprietary, not commonly used, or physical formats that will be very difficult to convert.	Data are in proprietary, not commonly used, or physical formats that will take some time to convert.	Data are in proprietary or not commonly used formats that will be very easy to convert.	Data are in a machine-readable, open format.

3*	User communities	There are no user communities who would be able to reuse these data.	There are very few user communities that may be able to reuse these data, and we have not contacted them.	There are user communities that may be able to reuse these data, but we have not contacted them.	There are user communities that can reuse these data, but we have not contacted them.	There are user communities that can reuse these data who have been contacted and have expressed interest in reusing the data.
4*	Stakeholders	There are investors in the data or research it supports who would not allow reuse of the data.	There are investors in the data or research it supports who would require it to have strict licensing/access conditions.	There are investors in the data or research it supports who would require it to have some access conditions.	There are investors in the data or research it supports who may allow it to be completely open for reuse.	Any investors in the data or the research it supports would not be affected by its reuse.
5*	Reuse value	There are no reuse opportunities for this collection.	The reuse opportunities for this collection are limited.	There are potential reuse opportunities for this collection that haven't yet been explored.	There is one good reuse opportunity for this collection.	There are a few good reuse opportunities for the collection.
6*	Reusable objects	There are no specific components of the collection that carry reuse opportunities.	There is one component of the collection that carries minimal reuse opportunities.	There are a few components of the collection that carry minimal reuse opportunities.	There is one component of the collection that carries good reuse opportunities.	There are a few specific components of the collection that carry good reuse opportunities.
7*	Historical value	There is no historical value to this collection—the scientific approaches, results, or advances are not noteworthy	This collection has minimal historical value—the scientific approaches, results, or advances are	This collection has some historical value—the scientific approaches, results, or advances were	This collection has good historical value—noteworthy scientific approaches, results, or advances	This collection has strong historical value—noteworthy scientific approaches, results, or advances

		and are not documented or evidenced by the data.	documented but were not noteworthy.	noteworthy but are not documented or evidenced by the data.	are partially documented or evidenced by the data.	are documented or evidenced by the data.
8*	Historical objects	There are no components of the collection that carry any historical value.	There is one component of the collection that carries minimal historical value.	There are a few components of the collection that carry minimal historical value.	There is one component of the collection that carries good historical value.	There are a few specific components of the collection that carry good historical value.
9*	Completeness	There are gaps in the data collection that will erase reuse and historical value.	There are gaps in the data collection that significantly limit reuse and historical value.	There are gaps in the data collection that somewhat limit reuse and historical value.	There are gaps in the data collection that slightly limit reuse and historical value.	There are no gaps in the data collection OR there are gaps in the data collection that do not limit reuse or historical value.
10*	Sensitivity	There are aspects of the collection that are extremely sensitive to unintended or undesirable access, use, or interpretations.	There are aspects of the collection that are very sensitive to unintended or undesirable access, use, or interpretations.	There are aspects of the collection that are somewhat sensitive to unintended or undesirable access, use, or interpretations.	There are aspects of the collection that are mildly sensitive to unintended or undesirable access, use, or interpretations.	There are no aspects of the collection sensitive to unintended or undesirable access, use, or interpretations.
11*	Access and use constraints	The data will be closed (not shared with anyone).	The data will be controlled (to access the data, a bespoke licence will need to be negotiated).	The data will be restricted (access will be restricted due to the nature of the data)	The data will be shared (with a predefined list of people).	The data will be open (shared with anyone, as long as they obey conditions of licence).
12*	Rarity or uniqueness	The data are already	The data could be	Some parts of the data	Few parts of the data	The data are

		available elsewhere.	fully recreated OR most of the data are actively stewarded, curated, or maintained by another group or institution.	could be recreated OR some parts of the data are actively stewarded, curated, or maintained by another group or institution.	could be recreated OR few parts of the data are actively stewarded, curated, or maintained by another group or institution.	fundamentally irreplaceable and not available anywhere else.
13*	Reproducibility	The data is fully reproducible.	Many aspects of the data are reproducible.	Some aspects of the data are reproducible.	Few aspects of the data are reproducible.	No aspects of the data are reproducible.
14	Relevant collections	There are no other collections of research materials that are relevant to this collection.	There are minimal collections of research materials that are relevant to this collection.	There are some collections of research materials that are relevant to this collection.	There are many collections of research materials that are relevant to this collection.	There are many open collections of research materials that are relevant to this collection and demonstrate a wider network of interest or investment in the research.
15	Associated publications	There are no identifiable publications associated with the collection.	There are minimally helpful identifiable publications associated with the collection.	There are somewhat helpful identifiable publications associated with the collection.	There are helpful identifiable publications associated with the collection.	There are very helpful identifiable publications associated with the collection, such as scientific journal articles that report, rely on, or cite the data or methods represented by the collection.

16	Fit for purpose	The data are not at all ready for reuse. There is a significant amount of documentation, interpretation and processing required to prepare the data.	There is a lot of additional documentation, interpretation and processing required to prepare the data for reuse.	There is some additional documentation, interpretation and processing required to prepare the data for reuse.	There is minimal additional documentation, interpretation and processing required to prepare the data for reuse.	The data are fully ready for reuse.
17*	Obstacles to recovery	There are significant obstacles that would prevent data recovery.	There are significant obstacles that would make data recovery very difficult.	There are some obstacles that would make data recovery somewhat difficult.	There are a few obstacles that would make data recovery slightly difficult.	There are no anticipated or observed obstacles to data recovery.
18	Priorities	There is no date in mind for beginning to tackle priorities for data recovery OR priorities for data recovery can start to be tackled more than a year from now.	Priorities for data recovery can start to be tackled within the next year.	Priorities for data recovery can start to be tackled within the next six months.	Priorities for data recovery can start to be tackled within the next month.	Priorities for data recovery can start to be tackled immediately.

Appendix 2 – Full Transferrable Method for Rescuing Digital Tabular-based Data

1. Initial assessment of files
    a. Identify any potentially relevant/important files for the data rescue (datasets, images, maps/location info, metadata, project information, publications, instruments/equipment information, etc.)
    b. If version control is unclear, open the files to determine which file is the most complete/up-to-date (note that this may not necessarily be the most recently edited file)
    c. Note file formats and file sizes

    d. Briefly identify any potential issues with files (sensitive data, out-of-date format, lack of units, lack of key, formulae errors/issues, lack of date/time info, missing data, unclear/unspecified variables, etc.)

2. Assess identified files using Hoffman et al (2020) framework
    a. Construct report briefly assessing against each factor in framework
    b. Score each factor out of five based on how suitable the file(s) are for rescue
    c. If files(s) score at least 60/90 then dataset is "worthwhile" to rescue and can continue
    d. Make list of next steps required for rescuing dataset (address any issues identified in Step 1d)
3. Conduct a detailed quality check of the data
    a. Check that all variables are clear, identifiable, and have units
    b. Identify any missing data, and check for any explanation/notes in the file or in associated files
    c. Check all formulae in document are correct (consult any researchers of similar research if needed)
    d. Identify temporal resolution, e.g., if weekly, check if all dates reflect this
4. Address any issues with the dataset if possible
    a. Search the internet for any associated publications that could assist in understanding of the data
    b. Use any supporting documentation found in the archive files
    c. If there are still unresolved issues, contact any researchers involved in the original data collection or involved in associated projects
5. Clean the dataset(s) if required
    a. Choose an appropriate data centre with which to ingest your data
    b. Check the requirements for datasets
    c. Determine the format and spreadsheet layout most suitable for reuse of the dataset
    d. Create a code script to clean the data to a stage where it is understandable and could be ingested into the data centre (remove any sensitive data, resolve any formula errors, ensure suitable column headings are used, etc.)
6. Create supporting documentation for the dataset
    a. Check the requirements for supporting documentation for the data centre
    b. Use all information gathered so far to write a supporting document
    c. If any missing information is identified that cannot be found in associated publications, contact any researchers involved in the original data collection
    d. Conduct any extra quality control required if any further issues are identified through this process
7. Assess data against Hoffman et al (2020) framework again (optional)
    a. Change any notes for any changes and quality checks
    b. Score each factor out of five again
    c. Compare the scores to discover how the measures taken have altered the suitability for data rescue
8. Submit data to be deposited into data centre

References

Cape, J. N., Smith, R. I., & Fowler, D. (2010). *Long-term monitoring of cloud chemical composition in the UK and implications for estimating wet deposition.* Retrieved from NERC Open Research Archive website: https://nora.nerc.ac.uk/id/eprint/505413/

Choularton T. W., Gay M. J., Jones A., Fowler, D., Cape J. N., & Leith, I. D. (1988). The influence of altitude on wet deposition comparison between field-measurements at Great Dun Fell and the predictions of a seeder-feeder model. *Atmospheric Environment* 22, 1363–1371. https://doi.org/10.1016/0004-6981(88)90161-8

Crossley, A., Harvey, F. J., Ferguson, S., Leaver, D., Fowler, D., & Cape, J. N. (2023). *Cloud and rain pollutant concentration and deposition data at Bowbeat, Peebles, 2003–2006* [Data set]. Lancaster, UK: NERC EDS Environmental Information Data Centre. https://doi.org/10.5285/2dca8f2f-a21b-4f77-bc8c-326269ab58d1

Fowler, D., Cape, J. N., Leith, I. D., Choularton, T. W., Gay M. J., & Jones, A. (1988). The influence of altitude on rainfall composition at Great Dun Fell. *Atmospheric Environment* 22, 1355–1362. https://doi.org/10.1016/0004-6981(88)90160-6

Hoffman, K. M., Clarke, C. T., Shiue, H. S. Y., Nicholas, P., Shaw, M., & Fenlon, K. (2020). *Data Rescue: An assessment framework for legacy research collections* [White paper]. University of Maryland College of Information Studies. https://doi.org/10.13016/1zmx-ghhq

Naujokaityte, G. (2021, June 14). Number of scientists worldwide reaches 8.8M, as global research spending grows faster than the economy. *Science Business.* Retrieved from https://sciencebusiness.net/news/number-scientists-worldwide-reaches-88m-global-research-spending-grows-faster-economy

Science Staff. (2011). *Challenges and Opportunities 331,* 692–693. https://doi.org/10.1126/science.331.6018.692

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä, Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., & Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data* 8, 192. https://doi.org/10.1038/s41597-021-00981-0

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3(1). https://doi.org/10.1038/sdata.2016.18