EXECUTIVE SUMMARY

The *Guidelines* that are embedded in this document offer constructive advice, and web-links where they exist, on methods, and relevant information where available, on how to research, locate, rescue, digitize, store and share, analogue or early-digital materials in electronic formats that are fully commensurate with modern ones.

1. Where and how to begin?

It seems unavoidable that the *Guidelines* are patchy and incomplete. The information they contain is reached by series of binary decisions such as:

Were the records in question known to exist, or were they discovered serendipitously? Are the original records in good, indifferent or parlous physical condition?
Is there living knowledge concerning the activities which produced them?
Where relevant, can all the necessary calibration observations be found?
Are the materials in danger of being discarded as "unwanted"?
Is special digitizing equipment needed?
Can the envisaged work be carried out by untrained personnel (with or without supervision)?
Is expert handling (e.g. by an archivist) needed at any point?
Are there data management plans (institutional or otherwise)?
Is there an identified repository for long-term preservation of originals?

Responses to the above dictate how a project to rescue and recover non-digital materials must be tackled. However, any one answer will be conditioned by circumstances, forces, opinions and luck, to an extent that makes it scarcely feasible to create a model to represent the general case. The real-life example in the writer's own experience ("Bundles of papers in a cupboard", page NN), could be typical of any "re-discovery" of raw materials, though there must be many unrecorded ones with less happy outcomes.

2. Success and Serendipity

Two extremes of situation illustrate the fortune or fate awaiting the investigator: the benign and the unlucky. In the former case, original materials have been stored carefully in a well-protected location, and are made available to the researcher. Their provenance has been described and recorded, and a few retired staff on hand have long enough memories to recall important specifics of the experiments in question and even fill in gaps in the written records. The materials (in papers, and journal reprints) are in good condition and can be photocopied cleanly. Appropriate OCR software is able to deal with the tabulations of numbers with very few errors, and all the necessary metadata can also be recovered from notebooks stored in the same location. Incorporating the recovered information into a modern analysis extended the time-base through a period which proved critical to the science, and revealed trends in the data which no models had foreseen.

In the second case, the researcher visited the location where the materials in question were said to be stored, but found access troublesome, and only part of the materials actually there. Notebooks that should have contained metadata were incomplete, damaged and rotting, and no-one was able to recall the circumstances of the experiments concerned. Photocopying the papers proved difficult because of the flimsy and damaged state of many sheets, and the OCR returned multiple errors because the ink was faded and the pencil nearly illegible, requiring the researcher to key-in by hand

the numerical columns. When finally the new sets of information were included in the researcher's programme, it was found that the most critical sets were the ones missing.

3. The Unsatisfactory Present

(a) Recognition of the need to rescue past observations, records and materials for modern research is growing, and growing healthily. It has driven two very active Groups (CODATA: *Data At Risk Task Group*, 2010–2016, and the Research Data Alliance *Data Rescue Interest Group*, 2015–present), and launched an International Workshop in Boulder (USA) in September 2016. Many who have long harboured a wish to do something useful with heritage materials are becoming heartened by the knowledge that the whole matter is being brought to the fore and given specific airing.

(b) However, too much is being left to serendipity. It is fortunate if the relevant players and pieces can be reassembled successfully, but too often it is already too late for action to resolve those situations. The need for action is therefore becoming critical, and growing so daily.

(c) There is no formal body or organization that carries, or can link to, the kinds of expertise and information sources which the seeker needs. In particular, there is no recognized pool of expertise or collected experience to which a seeker can refer.

(d) Establishing the needs outlined in (c) requires awareness raising. Case Studies benefit by communicating, sharing expertise where feasible, and sharing findings or experiences. Regular awareness-raising is nearly non-existent, too patchy, and rather ill-received.

4. Creating a Robust Future

At present, efforts to rescue data are not at all well supported by the community. Often they are perceived as competitors for dwindling resources of funding or manpower. Using volunteer labour can benefit from immediacy, but may be unable to enforce rigour beyond a basic level, or cope with uneven output quality. Furthermore, the status of a project is assessed by the qualifications of those who do it, so a project that can be done by volunteers becomes tainted by the preconception that it is low-grade work that even untrained people can carry out, and not worthy to be considered for funding from a category for more "serious" science. These attitudes can and must be reversed.

(a) A formal international body for "Data Rescue" needs to be established. Its members (those who carry out Data Rescue in the sense described here) will amass expertise, information and experience, and will maintain a Web-page with a suitable Help line. Its dedicated co-ordinator, communicator and fund-raiser will maintain a bibliography of published Case Studies.

(b) Links to Web-pages describing successful recovery projects can boost morale, build confidence, cut costs, save time and enhance knowledge derived from research. They will be the responsibility of the Co-ordinating Body.

(c) Scientists seem rarely to throw anything away, but losses have occurred and it is now within our power, ability and resources to ensure that further losses are minimal. This comes back to awareness raising, as in 3(d) above. Once an international body makes the case in public that heritage records can contain unique and unrepeatable information, of benefit to science and ultimately to society, efforts to protect and rescue historic legacies will be recognized before original media deteriorate or (worse) human ignorance orders their disposal.

BACKGROUND

A compilation of advice, help and suggestions

FOREWORD

Any document that tries to explain "data rescue" must define what "data" are being addressed, and describe what it includes under "rescue". As will quickly become apparent, there is no possibility of a "one size fits all" recipe in the domain of data rescue; the advise which these *Guidelines* offer needs to be both extremely general and at the same time extremely focussed. What is contained in the pages that follow is our effort at handling that ideal requirement as seen from the perspective of the reader (with whom there is considerable empathy since all of the points that are treated here are based on real-life experiences).

We first identify the context and the rationale which has nudged these *Guidelines* into being, and identify the audience for which they are written (herein referred to as the *seeker*), and the wider population of *users* who are ultimate likely to benefit through the ability to access heritage materials that have been converted to modern electronic formats. The *Guidelines* are lodged within a frame that describes the **context** in which "data" may need to be "rescued", and adds a **dictionary** of key terms and phrases which may seem obvious to some but may have unintentional duplicates that muddy the picture for others. The *Guidelines* adopt the language of the definitions which comprise the Dictionary to ensure that its contents are clear and unambiguous. They then present the large array of options which are part of the seeker's experience in practice, and conclude by suggesting actions, information sources and other snippets of intentionally helpful advice. An **Executive Summary** places the activity of "data rescue" against a background of research and science.

1. INTRODUCTION

Scientific research is largely empirical. It relies heavily on observations, records and measurements, and attempts their interpretation through models based substantially on theories or hypotheses. Recognizing the need to make and record accurate observations marked the commencement of scientific research as we know it today, whereby aspects of an object or system are measured and recorded to a quality that withstands external scrutiny, and bear all the necessary *metadata* descriptors to place the observation or measurement correctly in time, place, and context.

The behaviour of the objects or systems which come under examination in research is autonomous; objects with a proneness to change will manifest changes regardless of who or what is recording their parameters. If that statement appears tautological, it must be remembered that a considerable amount of "knowledge" about such systems is seriously limited by the scope of information available, either in range of date (as in the case of modern electronic archives) or in kind (wavelength/frequency, cadence, detector properties, observing conditions, etc.). In particular, advances in certain modes of observing can render one set of observations more productive than another, and in doing so may bias a whole series of facts derived from that particular collection.

2. CONTEXT

All digital files are to some degree "at risk", whether from willful annihilation, virus damage, interference by hacking, or mere human error. The advent of born-digital records (generally dated as post-1980) spawned a whole industry of digital management, storing, archiving, "data curation"

and "data science", with the result that many born-digital records can today be processed, shared, borrowed and incorporated into research smoothly and easily. Other major efforts now planned will enhance further those services and technologies.

That situation is sadly not true at all for pre-digital records – those on analogue media such as paper, photographic plates or films, books, pro-formas, etc., or (where traditional knowledge is concerned) even the spoken word. The situation which *they* collectively face amounts to a severe deficit of attention. Extra, usually discipline-specific, steps are required to render pre-digital materials accessible electronically. Most of the transformation steps do not not demand technology that is challenging or costly, or requires skilled operators; nevertheless, it is at this point that the upgrading of the world's legacy of analogue materials is now stalled. The reason is only partly one of resources; attitudes have a great deal to answer for: the shiny and new has undiscovered potential and and can appear more exciting than the "old stuff" with its presumed inferior quality and content. Nevertheless, that "old stuff" has the unquestionable advantage of a date-stamp that can pre-date anything born-digital by long margins, and which renders it indispensable for studies of long-term variability.

Analogue materials are unique, since (unlike electronic files) they cannot be copied easily or well, a property which puts all the information which they contain "at risk" from physical deterioration of the media through to damage or destruction from natural disasters. It is possible to prioritize the urgency with which "data recovery" needs to be applied, e.g., according to physical state or storage conditions (not to mention human indifference), so in effect *all* analogue materials should be seriously considered for urgent rescue as soon as suitable resources can be made available. The *Guidelines* suggest modes of searching for those unique analogue materials and relevant metadata, and discuss schemes to perform digital transformations into acceptable electronic versions of the information which they bear. Early magnetic tapes should be included within the scope of "data at risk" if (as is so often the case) they lack formatting information and metadata, and the correct software to read them is no longer useable (if by any slender chance it is still available).

3. RATIONALE

Too often one hears the question, "Why bother with old materials when there are so many *modern* [superior] ones nowadays? Superior they may be, but none can reveal the prevailing parameters or conditions which pertained in the decades before the digital era. Many sciences are deeply involved in studies of variability of some type, whether periodic, explosive, repeating or irregular. Those properties reveal key information about the objects in question, even of their actual evolution. The "old stuff" can therefore be an indispensable adjunct to the modern records. There are numerous cases (some of which are cited on pp. NN) in which recovering heritage materials settled a dispute, relayed unexpected new information, refined understanding, revealed errors in present thinking, or displayed results that seemed at first sight to be too controversial to publish. It was only by putting the heritage observations alongside their modern counterparts that the new science could be uncovered. While the older materials were not of themselves able to provide all the necessary information, it was by extending the usefulness of the modern ones that the true richness was added. That property has been, and remains, the central driver for recovering science's heritage observations. Cultural studies also have their counterpart, as the recovery of past observations can stimulate curiosity into history for its own sake.

Another important reason for recovering heritage observations is the ability to employ them in a trans-disciplinary mode, as science itself has evolved during the sometimes long intervals of time

that have elapsed since the heritage observations were made. This is of course an option for modern analyses too, but the lengths of time involved in the case of analogue materials have sometimes been sufficient to see substantial changes in thinking or application, and the old materials can then be used with great effect in a situation which is rather far removed from the purposes of the original observers. Observations of stars in the 1920s–1930s, for instance, have been re-worked to determine the concentration of ozone in the Earth's stratosphere at that epoch, even though the presence of ozone up there was barely recognized when the stellar records were actually made.

4. DICTIONARY

(a) Data: Raw observations, records or samples that include descriptors and metadata, and measurements as carried out. (*Origin:* Latin, plural, meaning "things given"). **Data** are therefore *objective* inasmuch as they cannot be altered, though they can be transformed through calibrations into units that have meaningful scientific application. In everyday speech, especially English, the word is used loosely to refer to *things, stuff, columns of numbers* or *facts*, but the *Guidelines* only employ it in the sense of original observations. Ideally, **data** should be made freely available to any *bona fide* researcher, after suitable anonymization if necessary, except if protected by a short-term proprietary restriction.

((b) Metadata: Essential information about the **data**, e.g., date, location, context of experiment, present storage venue). Without **Metadata**, the **data** lose meaning and relevance. **metadata** belong with, or should always be traceable from, the corresponding **data**.

(c) Information: Facts such as physical or chemical parameters, classification, population, derived from and about individual **data**. **Information** is published in catalogues and tables, and as supporting materials to a journal paper, report, or other form of dissemination. **Information** should be placed in the public domain unless protected by a proprietary restriction (e.g., in a thesis awaiting defence). **Information** is *subjective* since its derivation depends on choices (often guided by personal preferences) of algorithms, theories and laws.

(d) Data At Risk: (i) *In the context of analogue data*: Observations, records, files or measurements, mostly (but not exclusively) on analogue media and which face potential loss through physical degradation, damage from any cause, or destruction from natural disasters, or because of a supposed lack of scientific value. Data sets can become orphaned when a project ends and the P.I. moves on without having created a sustainable management system, and the project's website is no longer maintained; even the location of the data set can be in doubt. Data are also "at risk" when specific technology or software that is required to read or process them is no longer available, or if metadata or calibration files have not been linked to the set and there is no guidance on how to analyze the data. Any of these factors can trigger a skepticism as to the value of such materials occupying valuable storage space, and the materials risk being thrown away. (ii) *In a general context*: Any data which seem prone to being deleted, more through malicious effort than through accident or natural disaster.

(e) Dark Data: The information assets which organizations collect, process and store during regular [business] activities, but generally fail to use for other purposes (*from Gartner IT Glossary*). In everyday parlance, equated with stuff (junk . . .) that is hoarded for no expressed reason.

(f) Data Rescue: The pursuance of activities such as salvaging, retrieving and recovering heritage **data at risk**, and placing them in a protected environment where they can be curated and shared.

The rescue of a data set at risk usually becomes more difficult as time passes.

(g) Archive: An active repository for **data**, either electronic or analogue. An active **archive** curates the data, performs quality checks, and migrates the contents if upgrading of computing facilities so requires. An archive of physical data may contain a wide selection of materials, from collections of seeds, historical photographs of glaciers, historical newspapers or biological slides to museum collections. Often used synonymously with "database".

(h) Data Curation: Ensuring that **data** are reliably retrievable for future research purposes or re-use (*from WhatIs.com*). Normally carried out within, or associated with, an **archive**.

(i) Data Repository: A secure, trusted, long-term storage location for physical objects such as books, papers, photographic film or plates, or samples (e.g., seeds, rocks). The emphasis is on storing and cataloguing the contents, as in a museum rather than an archive.

(j) Data Base: A collection of digital files derived from raw **data**, possibly including calibrated versions as well.

Others? To be added.

5. CASE STUDIES (Incomplete)

1. Bundles of paper in a cupboard (*Pure serendipity*)

2. Sniffing out the evidence (*Trans-disciplinary research*)

An epidemiologist demonstrated recently how he could follow the progress of cholera through Spain by sniffing paper records dating from the 17th century. Vinegar was used by the post offices of the time to disinfect correspondence coming from infected sites, and its smell never goes away.

3. The significance of cholesterol in the blood (*Peer-reviews misjudged the study*)

4. Re-reading astronauts' Moon tapes (*Improved technology made all the difference*)

5. Notes from a study of national relevance (*A desperate search for files once known to exist*)

6. Vital research from a despised and outdated technology in astronomy (*Eating humble pie*)

THE GUIDELINES

A. Basic Questions

The researcher who is seeking a set of legacy materials will already appreciate that nothing is standard about possibilities and procedures. Each situation needs to be addressed individually, and assessed from whatever facts are presented. We commence by listing situations which are most likely to greet the seeker, and then discuss in more detail the kinds of binary decisions which must be required. Examples that were met in the Case Studies (pp. 6–7) may prove useful.

Q1: Motivation and starting position
Q2: Physical condition of materials
Q3: Metadata, and other germane information
Q4: Calibration and other subsidiary files
Q5: Digitizing and scanning equipment
Q6: The role of experts
Q7: Data management
Q8: What to preserve?
Q9: Repositories for physical materials
Q10: Education to support the need for preservation

B. Detailed Responses

Q1: Motivation and starting position

In most data-rescue projects, the recovery and retrieval of information from heritage materials is a means to an end. The researcher will be looking for records, observations, etc., that are known to exist, even if their precise location and storage history are unknown, because it is believed that they will enhance a specific research study. Occasionally other, or additional, materials may be found that were unanticipated, or presumed long vanished. Local information, reports, relevant papers and colleagues working in the same field usually constitute the best sources of help. A personal visit to the site of the data store is strongly recommended unless that is not feasible; learning how the materials have been stored, and assessing their present condition, can be a valuable guide to their present potential. Word of mouth is acceptable if it comes from an interested collaborator who does have that first-hand information. Borrowing a few samples (if permitted) can also reveal the state of the materials and their re-useability. If the store of materials is well maintained, there may be conditions of access to the location; an active archive may not permit public access without prior arrangement.

Sometimes a cache of papers, books, or other forms of record, often long orphaned, is brought to light by chance, and its value needs to be determined. If the location of the cache is (say) within a Department that has occupied the building for some time, it should not be difficult to identify the materials, or to find someone who can. More rarely, the cache of materials will be spotted by a visitor who has no familiarity at all with the Department or its research but can fortuitously act as a go-between (see Case Study N, page X: *Bundles of papers*).

Just occasionally a researcher, librarian or archivist may be looking through abandoned storage, eager to examine whatever turns up. Finding something of potential use is then an end in itself. The searches can prove to be exciting, disappointing, rewarding or frustrating, but invariably (as

with all research) pose more questions than they attempt to answer. Information – of any kind – is the overarching requirement, and is best sought by beginning with the department or individual most likely to be able to respond positively. Even sorting out the best direction to get to that point can take time, meet numerous brick-walls or end in cul-de-sacs. Describing the find on a public website may be one way of getting out of that particular rut.

Q2: Physical condition of materials

Are the records, papers, etc., in a sufficiently robust condition to withstand removal to a different site? Is paper brittle or decaying, the writing too faded to read, showing signs of attack by vermin? These matters are obviously of major concern. If paper is too fragile to handle normally, photocopying it could endanger the integrity of the information it contains, but the quantity might be too large for comprehensive repairs.

Materials in poor condition encourage opinions that they are unwanted and not of scientific value, and risk being discarded. It is therefore important to have a firm idea of at least the qualitative, if not the fully quantitative, potential of the materials for present-day research.

Q3: Metadata and other germane information

Is there living or documented information concerning the experiments which produced the original data? Are there conditions attached to accessing that information? A seeker who is attempting to work in a trans-disciplinary mode on data belonging to a field that is not his[1] own may lack either right or privilege to request help. It can also be difficult for someone not versed in the intricacies of domain-specific data to use them optimally, even correctly. A possible solution is to seek a collaborator from within the organization which owns the data in question.

When metadata are being added to data for storing in a repository, it is important to ascertain at an early stage whether the repository wants incoming files to adhere to a specific set of metadata standards.

Q4: Calibration and other subsidiary files

Can all the necessary calibration observations be found? Are the raw data still valuable without them? The answer will depend on the*kind* of information needed. If the seeker is looking for evidence of the *present–absent* type, then even if the raw data have to be used in a qualitative (e.g., uncalibrated), rather than quantitative, mode their usefulness can still be assured. However, if the raw data cannot be used correctly because (for instance) the response of a photographic emulsion to light cannot be specified, then the lack of that ability to calibrate the raw data undermines their value to a serious degree. It can sometimes be possible to proceed by "bootstrapping, i.e., comparing with a correctly-calibrated record to deduce what calibration recipe would have worked well enough for the sample that lacks its specific one.

Q5: Digitizing and scanning equipment

"Digitizing" means different things to different aspects of data recovery. A universal adage is that a digital version is only an observation of an observation; digitizing cannot add to the information content of a sample, and may reduce it in a manner that is deleterious for the intended purpose(s).

[1]For simplicity, the masculine pronoun is used to represent both the masculine and feminine forms.

(a) For handwriting or printed wording on paper (loose sheets or bound), where the words themselves contain all the information sought rather than creases, water-marks or other shadows, it should be sufficient to use a common flat-bed scanner. The mode of output can usually be selected at will.

(b) For photographic materials, the choice will depend to some extent on the nature and purpose of the originals. If the photographs display equipment, say, in order to enhance understanding of how an experiment was conducted, a flat-bed scanner will be adequate (it is partly what they were designed for). But if the photographic plate or film is a record of an image whose properties need to be quantified (e.g., determining the relative brightnesses or positions of objects, or measuring absorption lines in a spectrum), then a commercial flat-bed scanner – especially the cheaper ones – will not suffice since they introduce scattered light to an extent that can seriously compromise the scientific output. Scanners of that kind were not designed for quantitative scientific work. For those needs, a custom-built digitizing microphotometer is essential. Such instruments are not cheap, and require trained operators. For low-resolution originals, it can sometimes be found that one loses next to nothing by trading output quality with cost and expedience, but that should never be an overriding choice; it must be better to over-sample and (in effect) to over-scan than to risk limiting the effort, as it may never be possible to repeat the effort and re-scan an large amount of materials.

Q6: The role of experts

Archivists have specialist knowledge about handling heritage data, and will advise in the case of fragile or damaged specimens. It is therefore a good idea to identify one or more such people who can be contacted if difficulties arise. Are there resources to pay an expert if it proves necessary, or are there skilled personnel available in-house?

To what extent can the envisaged work be carried out by untrained personnel under supervision? Are there precedents for such attempts in the field in question? Involving untrained labour reduces costs but can involve other types of expenditure. "citizen scientists" have proved valuable assets for repetitive tasks such as classifying images against a set of templates, or manual jobs like photocopying or scanning papers. However, there are imitations and they need to be observed. Digitizing needs to be tackled at the highest available level, e.g., for medium-to-high quality photographic plates or films, special digitizing equipment will be needed and fairly extensive training is required before the activity can be left safely in the hands of a volunteer who may have rather little background knowledge.

To safeguard against errors or stakes like typos, it is usually best for the work to be duplicated by at least one other person, so that erroneous values or "personal equations" can be identified and annulled. It can be difficult to ensure completeness and to maintain standards of work when the labour is unpaid and may be a bit unreliable.

There is also a sociological downside to using voluntary help. Tasks of any kind tend to take on the status of the people who do them. Thus, being a CEO in a major Bank may be thought of as very important as it attracts a large salary, while the stay-at-home parent of young children is presumed to have given up on anything "more worthy". Whether raising a family of well-behaved and interested citizens is any less worthy than managing economies that someone else may undo the next day is of course debatable, but the comparison is not altogether absurd. Equally, if the saving and digitizing of precious heritage data for research purposes can be carried out by volunteers, it becomes presumed that it is of low grade if it can be done by anyone regardless of qualifications, and may be unfairly dismissed by one's peers as of small importance.

Q7: Data management

Establishing a routine for handling digitized versions of analogue materials is highly important; we do not want the output to become topics for data rescue in the near future. If a scientific department, in-house library, etc.) has its own systems for managing digital data, then that should be the optimal place to hold those derived from analogue materials. Collaboration with the database personnel should be established when digitization commences, in order that preferred data formats are absorbed at the start, all required metadata fields are completed, and a sustainable data-flow procedure is installed. will the newly-created digital files also be placed in the public domain, from where they will also be methodically migrated when necessary, and will access be granted freely?

Planning for managing the data in the long term can be critical. If the present sole manager (e.g., the P.I.) moves on, is there a contingency plan so that the digitized files will be secure into the future? It is also important to have a back-up plan in case the computer which stores the newly-created electronic files fails, or upgrades to the necessary software become imperative. Is there provision for retaining copies of any essential software, and carrying out backward-compatible upgrades?

Whether or not the Cloud proves sufficiently stable and constant in the sort of long-term management which will often need to be considered in cases of data extracted from heritage materials, only time will tell. For a relatively short term, particularly as regards isolated efforts by individuals, the Cloud may be a solution to be considered, but for the longer term it seems more likely that domain repositories would serve better because they can probably guarantee higher levels of security, and because they are likely to have a vested, and an historical, interest in the contents of those materials. Whether a domain repository is part of (or linked to) an academic institution or to a domain "data centre" depends at present upon the discipline in question; there is no standardization regarding the levels of user support either achieved or even intended, but that may improve in the future.

Comment: In the spirit of handling heritage data which have outgrown their original usefulness, the expectation is that electronic outputs resulting from digitizing such data will be placed in the public domain for free use (though protection for a proprietary period may be granted).

Q8: Tough decisions regarding preservation

Should all heritage materials be preserved? An accredited repository will set at least minimal recommendations, and can be costly in overheads, space and manpower. Attitudes will vary, depending on the institution and on the background and nature of the materials. Not all heritage materials will be of a quality that can justify the costs of long-term storage, but the general advice is to keep all materials until they have been assessed by an expert. Part of the decision will centre on the likely research potential, and part on their possible cultural or historical significance; in the latter case, it may suffice to retain examples only. The decision will also be influenced by the current level of management: good housekeeping maintains orderliness and efficient searchability, and will also add confidence that the log-books or notebooks are reliable and complete (as far as they may go). It has to be said that records which are no longer readable or useable, however excellent their book-keeping, will rank low when it comes to competing for resources to retain historic materials.

If pruning a collection becomes unavoidable, then another point to consider is the degree to which faithful digitization can be achieved. The uniqueness of a collection will be the first and foremost criterion, but if all else is equal then the condition of the physical samples (such as blemishes, scratches or cracks on photographic plates, crinkles or raggedness of papers, loss of bindings of

books, etc.) will necessarily have to become a factor too. In those cases it will help to carry out test digitization on a sub-sample. The loss of specific platform-dependent software once used for handling the kinds of information that could be extracted by digitization should not be an issue nowadays, unless the samples contain very idiomatic elements that have not been documented adequately in the metadata. At all events, simply photographing the samples, if that is feasible, will retain a useful record of what has had to be discarded. In any matter of pruning, knowledgeable people in the domain in question *must* be consulted for leading advice regarding the potential of all such materials. Final decisions may best be reached through consultation between domain specialists and archivists.

Sample digitization may also help the archive management to decide if a transformation from analogue to digital can be (or has been) carried out to almost 100% perfection and that it is unlikely that anything would be lost by losing sight of the originals. Faithful digital versions of papers will probably render the papers themselves redundant, but photographs which could bear (say) a faint haziness that is not reproduced adequately in a digital version should be retained so that they can be examined again by eye if necessary. Keeping more than one copy of digitized materials is fairly routine, and should be adopted as a universal policy. If any changes or corrections have been made to the original *information* too, it is essential that those changes be fully recorded by annotating version numbers. Making digital copies by different techniques could be considered in cases of highly sensitive materials, or where materials are in very poor physical condition, though doing so may mean less digitization of other parts of the collection and has to be weighed carefully against the expected returns of the investment of resources.

Materials that include traditional knowledge (e.g., maps, artefacts, taped recordings) are a special case for total preservation. The volume of materials is likely to be small. Recorded interviews should be transcribed onto paper to guard against loss through deteriorating tapes.

Q9: Repositories for physical materials

The recommended conditions for storage of various types of materials are the "archivists' bible", and have been listed elsewhere (URL??). They should be endorsed by the RDA, or by whichever body or organization(s) support this *Data Rescue* initiative in the longer term. Formal adherence to the policies outlined in this matter will help to ensure a high level of standardization for user and operator, and its significance will assist in engendering support from funders. A repository should be certified as "trusted"[2]. It must be able to manage the materials in the long term, so adequate provision for the attendant costs, including staff, maintenance, and possibly access fees needs to be assured, if not already absorbed by (say) institutional overheads. An active repository may have its own standards for metadata, cataloguing, proprietary access, etc., and it is as well to become wise to those in case extensive work later proves to be in unacceptable style. One should recall that a professional repository or archive is the locale of archivists, not research scientists, so the specific scientific value of a particular set of materials are likely to benefit from explanatory texts prepared for the intelligent layman.

Q10: Education to support the need for preservation

Are the materials in danger of being discarded as "unwanted"? If opinions circulate that "old stuff" has little or no relevance to "modern science", heritage materials can – and have been – thrown

[2] e.g., `http://www.trusteddigitalrepository.eu/Trusted%20Digital%20Repository.html`

out as "no longer of any value or interest", regardless of their physical condition. The situation is potentially dangerous, and calls urgently for Education. It helps if one can refer to case studies of successful data rescue and re-use projects, especially within the same or closely allied field of research, and to provide bibliographical links to relevant publications. It therefore serves everyone attempting data-rescue projects to publicize as broadly as possible the need and the rationale of Data Rescue, particularly from the perspective of supporting research based on modern materials (see also item C2). Whenever the opportunity arises at a conference on some aspect of "data", it is important to get the topic of rescuing heritage data put on the agenda.

C. Useful resources, and future efforts

A. Useful Resources (Can be expanded)

(1) European Space Agency (2014): Long Term Data Preservation Framework, http://wiki.services.eoportal.org/tiki-index.php?page=Long+Term+Data+Preservation+Framework

(2) US Geological Survey (2014): *Guidelines for the Preservation of Digital Scientific Data*, http://ndsa.org/documents/USGS_Guidelines_for_the_Preservation_of_Digital_Scientific_Data_Final.pdf

(3) North-East Document Conservation Center (https://www.nedcc.org/) offers free preservation advice, leaflets (https://www.nedcc.org/free-resources/preservation-leaflets/overview), and practical strategies (https://www.nedcc.org/preservation-training/digital-directions/dd-17)

(4) Data Refuge: Building a "data refuge", in particular for federal climate and environmental data that are vulnerable under an administration that denies the fact of ongoing climate change (http://www.ppehlab.org/datarefuge)

(5) Digital Preservation Network (https://dpn.org/): more than merely backing up digital content.

B. Future Efforts

(1) *Contact personnel.* Each set of heritage materials that is being prepared for digitization has unique properties. No two are strictly alike in condition, needs or potential, so each experiment sets its own precedent. It is therefore useful to build a list of Internet links, guides, and people who could be the first ports of call as occasions arise. Chairs of relevant *Interest Groups* of the *Research Data Alliance*, and consulted for feedback.

(2) *Broadcasting and publicizing.* Telling others what is being accomplished has a snowball effect, and is always worthwhile, from the level of the Department to top-level events in research. Giving presentations at conferences and writing them up in *Proceedings* anchors one's stories in the printed memory of the world, from where they can always be discovered by someone. What is out of sight soon becomes out of mind, and useage of heritage materials will dwindle unless access is both possible and broadcast. Even requests for access to nearly-lost data triggers interest in them. Enough triggers spur action.

(3) *Join forces.* Establish a link to other data-rescue projects, in any field (not necessarily one's own), and build networks. Exchanging experiences can provide valuable help if difficulties are encountered.

(4) *Tabulate incentives.* A bibliography of rescue stories is inspirational. Even efforts that fail in their endeavours are instructive.

(5) *Sources of funding support.* The funding needed for rescuing information from heritage data will vary according to its nature (volume, venue, uniqueness, etc.) and will very likely have to support manpower. The major foreseen costs involve some or all of the following: (a) Operational costs of recovery. (b) Storing and preserving, perhaps involving a change of venue, and long-term operational preservation. (c) Digitization: instruments, resources and manpower, including quality checks. (d) Transfer and safeguarding of digital data, especially if the undertaking is a new one, or cannot be absorbed within an organization such as a domain data centre.

(6) *Central Co-ordinating Organization.* A need is foreseen for a central (international) body to (i) represent the motives of Data Rescue at both world and national levels, (ii) maintain active communication between data-rescue projects, (iii) co-ordinate information, (iv) maintain a bibliography of successful recovery programmes, (v) set up a Help facility (a.k.a. Consultative Service), (vi) draw up a Guide to Best Practice, (vii) assist with local Workshops on "data rescue" ...