Selection of Research Data

GUIDELINES FOR APPRAISING AND SELECTING RESEARCH DATA

About this publication

Selection of Research Data; Guidelines for appraising and selecting research data
A report by DANS and 3TU.Datacentrum

SURFfoundation
PO Box 2290
NL-3500 GG Utrecht
T + 31 30 234 66 00
F + 31 30 233 29 60

info@surf.nl
www.surf.nl

Authors
Heiko Tjalsma - *DANS*
Jeroen Rombouts - *Technische Univeristeit Delft*

Contributors:
Jaap de Lange - *Technische Universiteit Delft*
Eric Rumondor - *Technische Universiteit Delft*
Madeleine de Smaele - *Technische Universiteit Delft*
Ellen Verbakel - *Technische Universiteit Delft*

Editor
Heiko Tjalsma – *DANS*
Annemiek van der Kuil - *SURFfoundation*

SURF is the collaborative organisation for higher education institutions and research institutes aimed at breakthrough innovations in ICT (www.surf.nl/en)
This publication is online available through www.surffoundation.nl/en/publications

July 2010

Contents

Foreword

This report was commissioned by SURFfoundation as part of the SURFshare program, which was set up by SURFfoundation to create a common infrastructure that will facilitate access to research information and make it possible for researchers to share scientific and scholarly information.

The report is the result of a short study conducted by two data centres, DANS and 3TU.Datacentrum. DANS – Data Archiving and Networked Services – is an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organisation for Scientific Research (NWO). Since its establishment in 2005, DANS has been storing and making research data in the arts, humanities and social sciences permanently accessible. DANS itself develops permanent archiving services, encourages others to do the same, and works closely with data managers to ensure that as much data as possible is made freely available for use in scientific research. 3TU.Datacentrum is partnership between the libraries of the three universities of technology that together form the 3TU.Federation. Professional storage, retrieval and provision of technical-scientific publications have always been the core tasks of these libraries, and they have acquired considerable experience in digital archiving.

The subject of this report is selection of research data. Its main deliverable is a set of practical guidelines for appraising and selecting research data, intended for all those who are in a position to do so. It should be emphasised that these guidelines are general in nature. As we will make clear, it is impossible in a report of this length to give *specific* selection guidelines for every discipline or stakeholder. The guidelines can be found in the Management Summary.

The report summarises the 'state-of-the-art' on this subject, based on recent literature, a limited number of interviews with some key players in this field, and the lessons learned at DANS and 3TU.Datacentrum. The first section outlines the main issues. The following sections discuss some of these issues in more detail. The practical guidelines described in the Management Summary can be used by everyone involved in selecting research data. Section 5 looks at the different roles of the people who have to answer the questions put forward in the guidelines. The report ends with a set of conclusions.

We would very much like to thank all those who have collaborated with us in this project, in particular the persons we interviewed: Ir. Niels Batjes (International Soil Reference Information Centre, ISRIC - World Soil Information), Dr. Gerben de Boer (Open Earth, Deltares), Vincenzo Beruti (ESA-ESRIN), Drs. Milco Wansleeben (EDNA) and Prof. Dr. Charles Jeurgens (National Archives of the Netherlands), as well as Drs. Marcus van Leeuwen (NWO) for additional information and Dr. Dirk Roorda (DANS) for his critical remarks.

Management Summary: general guidelines

Although a great deal of research data should be preserved, either for use/reuse or to validate research results, that is not true of *all* such data. To determine which research data are valuable sources or resources for research, we have developed a set of practical guidelines in the form of a checklist.

It should be emphasised that this checklist is *general* in nature. It provides a framework for creating selection guidelines for specific disciplines or stakeholders. It summarises the main reasons for selecting research data for long-term preservation. It can be used by individual researchers or research groups, researchers working together in collaboratories, research institutes, university departments, national and international organisations focusing on a specific academic discipline, or funding bodies. It can also be used by managers of data archives, data repositories or heritage institutes.

Selection should ***preferably*** take place at the time the data are created, if possible in accordance with a data management policy or infrastructure. If that is not possible, selection decisions can also be taken while the data are being entered into a data repository, or at any later time (for example when re-appraising an existing data collection).

Reasons for preserving research data:

1. Is there an ***obligation*** to preserve the research data so that it can be ***used/reused***? For how long? An obligation of this kind might be imposed by research funding bodies (for example NWO or universities), academic publishers or others.

   If there is no obligation: are there valid reasons to preserve the research data so that it can be ***used/reused*** **for research purposes**? Are those reasons valid from the perspective of the *research discipline* where the data were created or *academic disciplines other than the original* research discipline?

   These reasons could be:
   - Value of the data: potential value in terms of reuse, national/international standing and quality, originality, size, scale, costs of data production or innovative nature of the research.
   - Uniqueness of the data: the data contain non-repeatable observations.
   - Importance of the data for history, in particular the history of science.

2. Is there any ***obligation*** to preserve the research data for ***verification purposes***? For how long? This could be an obligation based on an existing code of conduct for research, like the *Netherlands Code of Conduct for Scientific Practice*, which prescribes storage of raw data for at least five years.

3. Are there reasons to preserve the research data for ***general (basically non-academic) purposes***? For example, are the data important for cultural heritage reasons, for museums or for other presentations?

Pre-conditions for selecting data for preservation:

One of the reasons mentioned above should always be decisive in making the final selection decision. The points below should also be considered before or during the selection process. They should be seen as necessary conditions for preserving research data but not sufficient in themselves. There must be clarity on all these points before a final selection decision can be made. This means that they must be checked to ensure that the requirements are fulfilled, depending on the nature of the data and the research.

- Technical: which data formats, software (standard or tailor-made: research-specific tools), hardware?

- Metadata: available and sufficient? Technical information, codebooks, information on data structure, contextual information, information on intellectual property rights, links with publications or related data (in a collaboratory e.g.).
- Data: which data from which point of the digital life cycle: raw data, intermediate data, published data?
- Clarity on intellectual property rights, for example copyright, patent and/or database rights, privacy protection?
- Infrastructure available for preserving the data? Either a data archive or an institutional or thematic repository.
- Costs: how are the costs to be covered for selecting, converting, preserving and making the data available?

Managementsamenvatting: algemene richtlijnen

(Translation of management summary in Dutch)

Hoewel onderzoeksgegevens veelal bewaard dienen te worden voor gebruik/hergebruik of ter validatie van onderzoeksresultaten, geldt dit niet voor *alle* gegevens. Om te bepalen welke onderzoeksgegevens waardevol zijn als (bron)materiaal voor onderzoek hebben wij een aantal praktische richtlijnen opgesteld in de vorm van een checklist.

Met nadruk wordt erop gewezen dat deze checklist *algemeen* van aard is. Het is een leidraad voor het opstellen van selectierichtlijnen voor specifieke wetenschapsdisciplines of gegevensbeheerders. In de checklist worden de belangrijkste redenen opgesomd voor het opslaan van onderzoeksgegevens voor de lange termijn. De checklist kan gebruikt worden door individuele onderzoekers of onderzoeksgroepen, door onderzoekers die binnen een collaboratory samenwerken, onderzoeksinstituten, universitaire faculteiten, landelijke of internationale organisaties die zich op een specifieke wetenschappelijke discipline richten en door onderzoeksfinanciers. De checklist is ook geschikt voor beheerders van archieven, onderzoeksrepositories en erfgoedinstellingen.

De selectie moet **bij voorkeur** plaatsvinden op het moment dat de gegevens gecreëerd worden, zo mogelijk op grond van een beleidsplan of een infrastructuur voor gegevensbeheer. Wanneer dat niet mogelijk is, kan de selectiebeslissing ook genomen worden op het moment dat de gegevens in een onderzoeksrepository worden ingevoerd of op een later tijdstip, bijvoorbeeld wanneer een bestaande gegevensverzameling opnieuw wordt geëvalueerd.

Redenen om onderzoeksgegevens te bewaren:

1. Dienen de onderzoeksgegevens **verplicht** te worden bewaard voor **(her)gebruik**? Voor hoelang? Deze eis kan bijvoorbeeld gesteld worden door de onderzoeksfinancier (zoals NWO of een universiteit), een wetenschappelijke uitgever of anderen.

   Wanneer er geen sprake is van een verplichting, is er dan toch een goede reden om de onderzoeksgegevens voor **(her)gebruik ten behoeve van onderzoek** te bewaren? Ofwel ten behoeve van de onderzoeksdiscipline waar de gegevens gecreëerd zijn ofwel ten behoeve van andere wetenschapsdisciplines?

   Deze redenen kunnen zijn:
   - Belang van de gegevens: potentiële waarde voor hergebruik, nationale of internationale positionering en kwaliteit, oorspronkelijkheid, omvang, schaal, de productiekosten van de gegevens of het innovatieve karakter van het onderzoek.
   - Uniciteit van de gegevens: de gegevens omvatten niet-herhaalbare waarnemingen.
   - Belang van de gegevens voor historisch onderzoek, in het bijzonder wetenschapshistorisch onderzoek.

2. Dienen de onderzoeksgegevens **verplicht** te worden bewaard ten behoeve van **controle**? Voor hoelang? Deze eis kan voortvloeien uit een bestaande gedragscode voor onderzoekers, zoals de *Nederlandse Gedragscode Wetenschapsbeoefening, die voorschrijft dat ruwe gegevens minimaal vijf jaar moeten worden bewaard.*

3. Zijn er redenen om de onderzoeksgegevens voor **algemene (in principe niet-wetenschappelijke)** doeleinden te bewaren? Zijn de gegevens bijvoorbeeld van belang voor cultureel erfgoed, musea of presentaties?

Randvoorwaarden voor het selecteren van gegevens voor bewaring:

In alle gevallen geldt dat een van de hiervoor vermelde redenen doorslaggevend dient te zijn bij het nemen van de uiteindelijke selectiebeslissing. Voor of tijdens het selectieproces moet echter ook rekening worden gehouden met de navolgende punten. Deze punten dienen te worden gezien als noodzakelijke, maar op zichzelf niet voldoende randvoorwaarden voor het bewaren van onderzoeksgegevens. Over al deze punten dient duidelijkheid te bestaan voordat een definitief

selectiebesluit wordt genomen. Dit betekent dat alle punten gecontroleerd moeten worden om er zeker van te zijn dat aan alle eisen is voldaan, een en ander afhankelijk van de aard van de gegevens en het onderzoek.

- Technisch: welke data formats, software (standaard of op maat gemaakt: onderzoeksspecifieke tools), hardware?
- Metagegevens: beschikbaar en voldoende? Technische informatie, codeboeken, informatie over de structuur van de gegevens, contextuele informatie, informatie omtrent intellectuele eigendomsrechten,links naar publicaties of gerelateerde gegevens (bijvoorbeeld in een collaboratory).
- Gegevens: welke gegevens uit welk moment van de digitale cyclus: ruwe gegevens, halfbewerkte gegevens, gepubliceerde gegevens?
- Duidelijkheid met betrekking tot intellectuele eigendomsrechten, zoals auteurs- en octrooirechten en/of databankrechten, bescherming van persoonsgegevens?
- Infrastructuur: is deze beschikbaar voor het bewaren van de gegevens? Ofwel een gegevensarchief, ofwel een institutionele of thematische repository.
- Kosten: is in de kosten voorzien van het selecteren, converteren, langdurig bewaren en beschikbaar stellen van de gegevens?

1 Introduction

This report describes a set of practical guidelines for selecting research data, useful for anyone who is in a position to do so. This section introduces the subject and the issues involved.

A special report by *The Economist* says 'Information has gone from scarce to superabundant'. This is true for almost every corner of society, and certainly for the research world. The amount of digital research data is growing exponentially. Should all data be preserved forever? And if not, how should we select the data that we *do* preserve? That is the theme of this study.[1]

Since it is the main subject of this report, we must clarify what we mean by 'selecting' research data. Archival science often makes a distinction between appraisal and selection. Appraisal is the process of evaluating documents (leading to a 'valuable' or 'not valuable' qualification), while selection involves actually removing a document or archiving it, based on the earlier appraisal decision. This distinction has a legal background and is important in the world of administration. As decisions concerning the selection of research data are hardly ever made in such a bureaucratic context, we will not make this distinction in this report and restrict ourselves to the term 'selection', which here will cover both meanings.[2]

1.1 For whom are these guidelines intended?

It should be emphasised that those in a position to make selection decisions are a widely varying group. They can be researchers themselves, acting either individually or collaboratively in multidisciplinary or international groups, academic institutes and university departments. But they can also be research funding bodies or curators of data archives, data libraries, university repositories, heritage institutes or comparable institutions. They can have very different positions in the process of creating, using, collecting and preserving data and therefore different interests. Additionally, even a group of similar stakeholders in similar positions may have a wide variety of different interests and tools available, depending, for example, on the disciplines involved, the community culture or the resources at hand.

Although the stakeholders are a heterogeneous group, we have developed a general set of practical guidelines. Because many of the decision points are basically the same, the guidelines can be used by everyone. This is particularly true for the pre-conditions that have to be met when selecting data, for example the presence of metadata, data formats, etc.

1.2 Need for guidelines: the general framework

The study concerned the situation in the Netherlands, although the researchers also made use of international contacts and recent international literature on the subject. The outcome is therefore specifically relevant for this country, but it could also be useful in an international context.

We decided to develop a set of practical guidelines because such guidelines did not yet exist. This initiative fits in with the SURFshare programme, which pays particular attention to data curation and long-term preservation of digital data. Not only are we increasingly aware of the importance of keeping research data permanently available, but rapid advances in ICT have made it easier for us to share information in the form of data. Data-sharing allows us to reuse or combine data in far more efficient ways. This study has been deliberately limited to research data because there are already coherent structures in place to preserve publications, whereas that is certainly not yet the case for research data.

[1] Data, data everywhere 2010
[2] Jeurgens 2008, pp. 18-20

As used in this report, the term 'research data' is restricted to *digital* data. The study did not explicitly take non-digital research data into consideration. Documents created within an academic research environment, for example e-publications, e-prints, e-mails or Internet pages, have not been considered unless they contained research data. There is no precise definition for 'research data', but it certainly includes all research output other than such documents resulting from research activities.

1.3 Reasons and pre-conditions for preserving research data

There are three groups of reasons for the long-term preservation of data. Whether or not research data are selected for long-term preservation basically depends on at least one of these three groups, and sometimes on more than one, as they are non-exclusive. These groups are:

- *Reuse* within or outside the research discipline in which the data were created. Also often described as *secondary use*.
- *Verification* of the data on which publications are based. Existing codes of conduct for research, for example the *Netherlands Code of Conduct for Scientific Practice*, often prescribe keeping the data available for verification for a mostly limited period.[3]
- *Heritage*: for historical research, in particular for the history of science, or more generally cultural heritage.

There is another set of issues that should be considered when making selection decisions. These are technical, documentary, legal and financial issues. They are to a large extent new issues that have emerged from the relatively recent phenomenon of archiving *digital* material, and are therefore still developing. We see these issues as necessary *pre-conditions*. They should always be checked, but they do not in themselves constitute reasons to either preserve or not preserve data. The practical guidelines, in the form of a checklist, can be helpful when addressing the issues involved.

1.4 Answers to elementary questions in subsequent sections

The following sections will answer a few elementary questions.
Section 2: Is it even possible to formulate selection criteria that can be used in every discipline?
Section 3: Why exactly should we want to preserve research data?
Section 4: Which stakeholders play a role in selection decisions?
Section 5: At which point in the digital life cycle should decisions be taken?

[3] www.vsnu.nl/Media-item-1/Code-of-conduct-for-scientific-practice-2004.htm

2 Can selection criteria be established that cover every discipline?

Is it possible to formulate *general* guidelines for appraising research data? This important question was addressed by this study. Research methods vary widely between the various academic disciplines, implying that both the data collection methods and the nature of the data involved may be very diverse. There are major differences in the way data, either original or taken from other sources, are processed, used and made available. In addition, technical and documentary (metadata) standards for describing and storing data vary between disciplines from non-existent to well established. There are major differences even within disciplines. A discipline is therefore too broad to serve as a category, so this study looks at the next level down, i.e. domains or specialised research groups within disciplines.[4] It considers whether it is even possible to formulate guidelines that will suit the many research disciplines in all their variety.

The conclusion of this section is that reuse, verification and heritage are the three main reasons for preserving data and that it is possible to formulate general guidelines based on these three reasons. These general guidelines will, however, have to be adjusted at a more detailed level for each discipline or domain, as will be made clear in this section. The guidelines can be used by researchers and by repository managers, regardless of the context. They could be part of the preservation policy of an institute or a data repository – either a thematic or institutional one – or even the policy of a research group or an individual researcher (Treloar 2007). They can include explicit instructions concerning significant properties.

It would be advisable to collect more examples of data archiving policies and their underlying considerations. It would also be advisable to document best practices of data handling from a very wide variety of communities.

2.1 Selection criteria in practice

2.1.1 Primary and secondary research data

After data are collected, they can undergo various transformations. There is an essential difference between primary and secondary data.

Primary data are data in their most basic and elementary form: unembellished, pure observations. These are often the *raw data*, i.e. data not yet influenced or shaped by researchers. Once researchers do something to these primary data, they become secondary data. Secondary data, or *processed data*, can be a combination or recombination of data or data that are recoded, categorised or visualised. Secondary data are often the data that are communicated to the outside world in one way or another, depending on the discipline or domain, either in publications (enhanced or otherwise) or in data collaboratories.

Primary data are not published in articles or books very often, at least not yet. For verification purposes, primary data tend to be preferred over secondary data so as to enable reconstruction of analyses performed during research. That is why codes of conduct for academic research prescribe preservation of primary data. Data repositories often preserve both raw data and processed data. The archaeological repository of the Netherlands, EDNA,[5] retains both raw and processed excavation data, for example. The primary (raw) data come directly from the excavations. The secondary data are the raw data after being processed by archaeological researchers to answer research questions. The Open Earth initiative (climate and ocean data) preserves all raw data and processing scripts, so that secondary data can be generated/regenerated.[6]

[4] Lyon 2010, p. 4
[5] www.edna.nl
[6] *The Netherlands Code of Conduct* 2005; EDNA and OpenEarth interviews

2.1.2 Levels of decision making

A useful starting point is to decide at which levels selection decisions will be made. Generally speaking, the existing literature identifies the following levels:

1. The institute: this is the general data *policy* of a research institute. That policy may consists of the institute's mission, goals, available resources (financial), and its legal obligations. Institutes can differ significantly on policy, especially with respect to the scope of their data repositories (thematic or institutional; restricted to their own data production or wider), the level and scale of adaptation by a community, and last but not least, the available resources. Long-term preservation is often part of this policy.
2. Data repository: this level involves the *collection criteria* formulated by the repository manager for determining which research data should be preserved, for whom and why. These criteria may be in line with an institute's policy, but data repositories can also cover much wider data collection areas than research institutes.
3. Designated community: this level concerns the *significant properties* of digital objects, or those properties that are essential to making permanent access possible.[7] What versions and specific properties (such as presentation, format, raw or processed data) should be saved and how much context and documentation needs to be preserved? At this level, issues such as the degree of standardisation, 'openness' and other legal or cultural aspects may be important factors for preservation.

2.1.3 Selection strategies

Section 4 looks more closely at who, or which parties, take selection decisions. Decision-making takes place at different levels, but the various levels also employ differing collection strategies. An examination of current practices applied by the various organisations involved in curating and archiving data reveals that there are basically two strategies:

- No pre-selection, all data presented are kept. Acceptance, however, always depends on re-usability, often related to metadata quality. There should also always be enough funding to process the data. This strategy can be combined with re-appraisal at defined intervals.
- Pre-selection: first of all, programmes, projects or experiments are selected for preservation. Secondly, data from the selected programmes, projects or experiments are filtered top-down by programme or project significance. Decisions may be taken by engaging the relevant academic communities, but that is not guaranteed.

Both of the above strategies can be further refined. To limit the amount of data, for example, a decision might be taken to preserve only the primary data or only the secondary, published data.

2.1.4 Application of criteria by repositories

Viewed from the perspective of data repositories, it is not possible to draw general conclusions from the limited number of interviews held within the context of this study. It is, however, striking that most data repositories do not appear to have well-defined selection criteria, or indeed to have any such criteria at all. The American data archive ICPSR[8] might be an exception, as it has a clearly formulated appraisal policy. Of course, data repositories remain within the boundaries of their data acquisition profiles, and most data repositories select on the basis of quality and completeness of the data and metadata. But they do not have explicit quality criteria. How is quality determined, then? And do we mean scientific quality or technical quality, for example the presence of metadata or data formats suitable for archiving? The interviews raised the point that selection criteria are needed when investing in retrospective archiving projects.[9] The data producers themselves probably applied criteria to the data offered for archiving, but it is unknown what these criteria are and whether they were applied consciously or not. What is clear, however,

[7] Van Horik 2009, p. 10
[8] www.icpsr.umich.edu/icpsrweb/ICPSR/
[9] EDNA interview

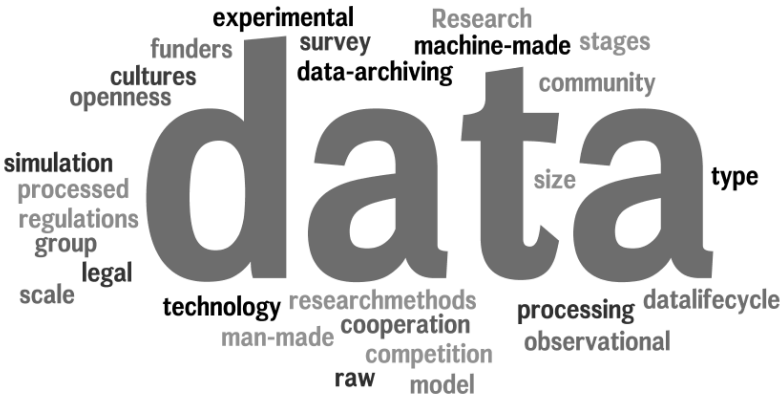
is that the reuse argument usually plays a decisive role for data repositories. Verification and heritage criteria, on the other hand, are mostly non-existent or secondary.

Does this mean that selection is not currently regarded as a major challenge? There is certainly awareness that in the future, selection might become more and more inevitable, as the costs associated with archiving data are expected to increase. These costs are related more to data handling (ingesting, checking, documenting) than to the actual storage facilities.

2.2 Selection criteria: is there common ground?

2.2.1 Differences between and within disciplines

Recent literature, such as the report *Data Dimensions,*[10] and a knowledge of current practice in data repositories indicate that general approaches may not be sufficient to take into account all the different types of research data from the various research communities. A number of related issues have arisen in that respect. There are differences in the type and scale of data and in the research methods or stages in the data processing or data life cycle, for example observational, experimental, model or simulation data, man-made versus machine-made (sensor) data, raw data or processed data. Research community cultures vary, depending on the group size, its degree of openness, its level of cooperation or competition, and its degree of technology-mindedness. Relevant legal obligations or other regulations, for example codes of conduct or contractual obligations by funding bodies, may play a role. Some research disciplines or communities have long experience in data archiving – the social sciences have, for example, been working with survey data for more than fifty years – while others have only been engaged in it for a relatively short period of time, such as archaeology. Yet other disciplines, such as psychology, have yet to begin.



Disciplines that have extensive experience in data archiving have well-established standards, in particular on metadata, for example the DDI[11] metadata for the social sciences. Many other disciplines, including history or psychology, do not have such standards, and existing data archiving infrastructures may vary greatly as a result. The main reasons for data archiving are verification, reuse or the degree of reproducibility.[12] There are considerable differences of opinion between and within disciplines and domains with respect to Open Access and Open Access standards. Opinions also differ on preferred formats, the average size, state (raw, processed) and structure of the data, and the level of investment required for metadata annotation.

Despite all these differences, the lists of *selection criteria* presented in the literature all have elements in common.
The reasons for preserving and maintaining access to digital objects each have their own specific demands. For research data in general it can be said that the three most important reasons are first of all reuse, followed by providing evidence (verification) and heritage.
Last but not least additional elements play a role, such as the presence of good metadata, re-usability of the data format, and financial considerations. These elements are regarded as *pre-*

[10] Lyon 2010
[11] Data Documentation Initiative: www.ddialliance.org/
[12] Van Horik 2009

conditions. They should be seen as necessary, but they are not sufficient in themselves to decide to preserve data.

a) Reuse

Reuse of research data is the most important and common motive for preservation. When referring to possible reuse of data, this is usually conceived as reuse within a given discipline. Reuse can mean re-analysing data from a new research perspective, based on general advances made in science. It can also mean combining/re-combining or simply comparing older data with new data or model outputs in order to obtain a fuller picture or a longitudinal series of data.

Increasingly, however, research data are used by researchers in other disciplines for interdisciplinary research, for example the ESA (European Space Agency) earth observation data.[13]

Regarding reuse, the value of research data depends largely on various factors such as quality, uniqueness, repeatability, production costs, scholarly use (now and in the future), risk of loss and indications for reuse (in publications or by request). The question is 'How do we measure this?', since all the factors mentioned above serve to illustrate that there are indeed major differences between and within research disciplines.[14]. Significantly, the DCC Scarp report recommends that 'researchers should work with colleagues and their discipline community to develop selection/appraisal criteria to identify priority data'. To which one can add: not only at the level of a discipline community, but also at the level of the domain or research group.

In our view, developing selection and appraisal criteria at the level of the domain or research group is the only way to overcome all the differences in data collection and curation that we have noted. Only researchers working in the academic discipline/subdiscipline itself can judge the academic value of 'their' data on the basis of content, whether it be data from experiments, simulations, observations or ancient but now digitised texts. They are the only ones who can decide what the value of research data is: are the data unique or easily repeatable? When data are used in interdisciplinary research, as indicated above, the academic disciplines/subdisciplines *using* these data may also be involved in selection decisions. It should be stressed that the actual decision as to whether or not to preserve data will not be taken on the basis of these criteria alone. Pre-conditions may play a role as well. These are criteria of a technical and data-archival nature. Section 3 describes the reasons for reuse, in other words the value, of research data and the pre-conditions in greater detail.

b) Verification

Verification is a very different category. Verification is almost always based on obligations, for example codes of conduct for researchers. This means that in fact there is no question of selection, nor do differences in scientific disciplines or domains play a role. The Netherlands Code of Conduct for Scientific Practice prescribes storage of raw data for at least five years. The preservation terms are therefore not that long and cannot be considered 'long-term preservation'. For verification purposes, it will usually be sufficient for the researchers themselves to preserve the data. As mentioned before, the overwhelming majority of data in data repositories, data archives and so forth is preserved for reuse reasons, but those data are also always available for verification purposes. We could even say that there is a 'grey area' between reuse and verification, with a newer generation of scientists in a certain field wanting to re-analyse old data because of doubts concerning the earlier analysis. In addition, new analysis methods may emerge in a discipline years after the original analysis was performed.[15]

Some scientific publishers require publication of underlying data with articles but they do not always facilitate storage of and access to large datasets and are less interested in re-usability and long-term preservation.

c) Heritage

[13] http://earth.esa.int/
[14] DCC Scarp report (Lyon 2010)
[15] The Netherlands Code of Conduct for Scientific Practice, VSNU 2005

Data can also be preserved for other reasons. These reasons are often summarised in the literature as 'heritage'. This means preserving research data for general historical research, in particular for the study of the history of science. The data may be used for all kind of purposes, comparable with the way documents are used in public archives.

Data are also increasingly used by non-academic 'researchers', for example journalists, interested amateurs such local historians, genealogists or others. This is particularly the case for data available on the Internet.[16]

2.2.2 Accountability: Open Access to data

A new set of selection guidelines for preserving administrative resources was recently developed in the Netherlands. The new guidelines are currently being implemented by the National Archives of the Netherlands and are based on the advice given by the Jeurgens Committee. The guidelines can be useful when looking at selection methods for research data, but they certainly cannot be copied without significant modifications. The main reasons they give for preservation are government accountability and historical importance. These two main points resemble the final two reasons mentioned earlier: verification and, to a lesser degree, reuse. Reuse as such is not an important motive in selecting government documents. This is a major point of difference with research data, as mentioned above.[17] Government accountability is one of the main reasons for making documents and data from public administration available to the general public, mostly in public archives.

Accountability could, however, become a new and important factor for research data as well. It is very much in line with the Open Access initiative (the Berlin Declaration),[18] primarily aimed at Open Access to scientific journals, but now extended to 'Open Data'. The outcome of publicly funded research should be available to the public, including the data. The recent row in the United Kingdom about access to historical climate data clearly demonstrates that public interest in scientific data is growing. Certainly not all data will become available: datasets are inaccessible to the general public if they contain personal or other protected data.

At the moment, however, research data is hardly ever preserved for reasons of general accountability as conceived here. Public archives do not play an active role in this, or any role at all, even when they have the legal means to do so.[19] Selection criteria are therefore non-existent in this respect. Having mentioned this dimension here, this report will not explore it further.

It is interesting for historians (history of science) and e-scientists to reconstruct the research process. They need the research data do to so, but even more important is contextual information on the origin and background of the project. This information is often contained in administrative documents: correspondence with university boards, funding bodies and colleagues. These more bureaucratic documents may well be preserved, but preservation of e-mails between researchers is far more problematical, even though they often contain the most interesting insights.

2.3 Conclusion

To conclude: reuse, verification and heritage are the three main reasons for preserving data, and it is possible to formulate general guidelines based on these three reasons. These general guidelines will, however, need to be adjusted to each discipline or domain. They can be used both by researchers and repositories, regardless of context. They could be part of the preservation policy of an institute, a data repository – either a thematic or institutional one – or even the policy

[16] Adams 1997
[17] Jeurgens 2008; Jeurgens interview
[18] http://oa.mpg.de/openaccess-berlin/berlindeclaration.html
[19] Jeurgens interview

of a research group or an individual researcher.[20] They can include explicit instructions concerning significant properties.
It would be advisable to collect more examples of data archiving policies and the considerations underlying these policies, and to document data-handling best practices from the widest possible variety of communities.

[20] Treloar 2007

3 Preserving research data: why and for how long?

This section looks more closely at the reasons for selecting research data for long-term preservation, as well as the reasons not to do so. We will also take a closer look at the necessary pre-conditions. An additional question is: if we select data for preservation, how long should it be preserved?

Our conclusion is that when trying to establish the value of research data for long-term preservation, uniqueness and repeatability are often each other's opposites. Peers in each discipline or domain must decide whether the data's uniqueness is significant enough to justify the investment in preservation.

There are *pre-conditions* that, although they are not the deciding factor in deciding whether or not to preserve research data, must nevertheless be taken into consideration. They concern issues such as providing sufficient metadata, legal or contractual rights, the availability of an infrastructure (including the necessary staff expertise), and the costs of preservation. Experienced staff from data repositories can advise on such pre-conditions.

3.1 Reasons for preserving data

As pointed out in section 2, the main question to be asked when selecting data for preservation is: which data will be reused in the future? It is difficult to give a general answer to this question, as this is very much dependent on the research discipline. However, some general guidelines can be sketched.

Uniqueness or repeatability is often mentioned as a criterion, but so are value, quality, production costs, scholarly use (now and in the future), risk of loss, and explicit indications for reuse in publications or by user requests. There is clear consensus that the reuse and consequently the preservation of older data generated by easy-to-repeat processes, for example computer models or simulations or small-scale experiments, is not considered as important. When, on the other hand, data are unique, non-repeatable, this is often seen as a good reason to preserve them. Non-repeatable data are produced by archaeological excavations, earth observations or opinion polls conducted forty years ago, to mention just a few examples. Their potential value in the future is hard to predict.

3.1.1 Value of research data

Of these criteria, the value of the research data is probably the most important one. It might be useful to look at the guidelines of the American National Archives and Records Administration (NARA) from 2003.[21] These concentrate on the value as well as on physical considerations such as usability and costs:

- How significant are the records for research?
- How significant is the source and context of the records?
- Is the information unique?
- How useable are the records?
- Do the records document decisions that set precedents?
- Are the records related to other permanent records?
- What is the time frame covered by the information?
- What are the cost considerations for permanent maintenance of the records?

These guidelines are general in nature and not specifically aimed at research data, but they can serve as a starting point for the selection of research data. Gutmann et al. have evaluated these guidelines from the perspective of the social science community. They concluded that, when used

[21] Gutmann 2004

together, the guidelines are a good example of the types of issue that arise when selecting data. They added a primary consideration to the selection process: the extent to which the data will advance knowledge. To put it briefly, data must have 'substantive value, enduring archival value and uniqueness'. Another factor that can be added is the academic status of the research group that has created or generated the data: its size, scale and above all its importance. One complication is that it is difficult to predict value in the future. Research interests or even research paradigms fluctuate in most sciences, and so does the demand for certain data as a result.[22] Documentary and technical properties (formats) play a supplementary role in making the final decision.

3.1.2 Uniqueness of data

Uniqueness is another important, even essential, factor in selecting data. There are many disciplines in which uniqueness is the primary reason for data preservation; see also section 2.2.1. 'Repeatability of data generation'. In fact, these are all domains in the arts and sciences where research cannot be recreated. This covers a very broad field, ranging from the earth sciences, earth observation data, archaeology, astronomy, and the social sciences to many disciplines in the humanities. In most cases, irreplaceable data consist of observations created either by humans or by machines. A number of disciplines in the humanities, however, use resources that are not originally digital; they consist of the digitised data of analogous originals. Historical research, for example, may use digital copies of documents that are centuries old. The most important sources in history, but also in literature or philosophy, have been digitised, but as long as the originals still exist, these digital copies are not unique. Only the most recent archival documents are digital originals. That means that, strictly speaking, the data or documents in a number of source-based domains are not unique.

Even this can be debated, however; historical databases or collaboratories containing data from history or literature[23] are new entities that combine digital data. These new entities can be considered unique in themselves because they select or combine data as well as offer hitherto unknown research facilities, for example search functions or GIS applications. The website 'Slave Voyages',[24] which contains historical data on the voyages of slave ships, is a good example of such a database. The underlying database combines data originating from various sources in many countries that have never been brought together before.

3.1.3 Other reasons

There are other reasons to preserve data, for example legal or contractual reasons. These are not necessarily related only to verification. Legal reasons can be based on archival law, in which the important criteria are accountability and cultural heritage.[25] In actual practice, however, these laws are of little relevance when it comes to research data. Contractual reasons can, on the other hand, be based on reuse considerations as well as on 'Open Access' [26]. This is certainly the case for obligations imposed by research funding bodies such as NWO. One explicit reason for NWO is to ensure that it is not wasting its huge investment in data infrastructure when subsidising research projects.

Editors or peer reviewers impose data preservation obligations on researchers who publish in their journals for more or less the same reason mentioned above. Journals can have a 'Data Availability Policy'; examples include *Nature* and the *American Economic Review*. Researchers publishing in these journals have to provide at least the final 'dataset for analysis' related to a particular publication.[27] However, in some cases researchers and project funding bodies can have conflicting

[22] Gutmann 2004, pp. 212-213
[23] IISH 2009
[24] www.slavevoyages.org
[25] Jeurgens 2008
[26] www.nwo.nl/openaccess
[27] IISH 2009, p. 9

interests with regard to the project results. In those cases, the research results are published without the datasets concerned, and long-term preservation of data is not included.[28] This does not necessarily mean that scientific publishers see a task for themselves regarding long-term digital preservation. A recent survey among publishers showed that most believe that 'other parties are better equipped to handle the preservation of research data'.[29] The size of datasets is also sometimes too limited for publishers, and the datasets often do not meet the requirements of re-usability, such as open formats and high-quality metadata.

Other important factors are the state of the data, meaning which stage of the digital life cycle they come from: are they raw, or have they been processed (re-organised, recoded, combined/re-combined, etc.); have they been published, for example in the form of tables, or have they at least formed the basis for publication?

Also of relevance are existing differences in the cultures of research groups or disciplines when it comes to their willingness to provide access to research data. Some are less inclined to do so than others; some are more open, some more restricted; some are innovative and some traditional. This is an academic culture problem that can also be observed in the attitude towards Open Access. Here, factors such as the degree of competitiveness and the use of personal data within a research discipline can play a role.

Even within a discipline, some data may be unique and irreplaceable while other data produced by data models or simulations are not. A clear example is climate science, where observational data are considered unique and valuable enough to preserve, but where climate data models are no longer seen as useful after only five years. These data are, in any event, rarely used by other researchers.[30]

3.2 Reasons for not preserving research data

Just as there are explicit reasons to preserve data, there are also explicit reasons not to do so. These reasons may, of course, be the reverse of the data preservation reasons given above. Apart from these reasons, pre-conditions may also play a decisive role. Ingest of data can simply be too expensive or too difficult owing to a lack of sufficient metadata or codebooks, for example.

3.2.1 Repeatability of data generation

Repeatability is the reverse of uniqueness. This is the case in all sciences where empirical data are collected in circumstances that can be recreated, enabling repeatability of the research. Whether the data are empirical is not decisive as such; what is decisive is whether these observations are repeatable, assuming that the observed phenomena have not changed. This might be the case with medical, biological or psychological research, in particular in laboratory environments. It is essential, of course, that the whole experiment can be copied exactly from the description of the experiment, simulation or other research setting, and this is compulsory in most sciences. In many experimental sciences, the idea is that, when experiments are repeatable, it is often preferable to redo them after a time to take advantage of ever-improving measuring techniques. On the other hand, the availability of large data collections offers new opportunities in scientific discovery and can improve the quality of research as well as ensure the efficient use of research resources.

There are other ways of considering whether experiments are actually repeatable, for example: when is it safe to say that research conditions in the future will not be different from present or past ones? One discipline in which this is the case is experimental psychology: do the subjects of behavioural tests behave in the same way now as thirty years ago?[31]

[28] Research Information Network 2008
[29] PARSE, Insight interim report 2009, p. 18
[30] Lyon 2010, p. 4
[31] Voorbrood 2010, p. 19 and p. 24

3.3 Pre-conditions for preserving data

It is important to realise that there is no real difference between digital and paper documents. This means that, fundamentally, we do not distinguish between selecting older data on paper and digital data. In practice, however, the new and unexpected features of digital data – both technical and intellectual – have raised a whole series of issues. One major point of concern is the as-yet underestimated challenge of providing permanent access to digital data owing to the rapid obsolescence of both hardware and software (media, platforms, operating systems). The following issues with respect to reuse should also be mentioned: providing sufficient metadata, the various legal or contractual rights involved, the availability of an infrastructure, including the necessary staff expertise, and, last but not least, the costs of preservation.

Many of the issues related to archiving digital material are new and still very much in development. They should be taken into consideration as necessary pre-conditions when making selection decisions. They should always be checked. Not all conditions will always be relevant; that depends on the specific research environment.

Individual researchers do not consider preservation capacity a major constraint. Some scientific communities tend to believe that selection of research data is certainly required, however, because if all data sets are preserved, major investments in preservation infrastructure will be beyond the available resources.[32]

Apart from capacity, the following constraints can be identified:

- **Preservation format**

Preservation format, in particular, is considered to be a constraint when it comes to multidisciplinary data. Data within a certain specific research area, such as soil science, can be preserved and made available in a well-defined format; they can be accessed using commercial software, which may not be affordable for individual users, or through the OSG.[33] Researchers in climate and ocean sciences tend to avoid this problem by promoting the use of freely available (open source) software.

- **Metadata management**

Most of the sources consulted in this study agree that the availability and management of high-quality metadata are essential for reuse, verification and common heritage. As a matter of fact, the availability of metadata is a primary quality assurance factor. Preservation without adequate documentation makes no sense at all.

- **Legal and ethical limitations**

Legal and ethical constraints are particularly significant in the medical and social sciences. Medical data, for example, must not be used in scientific research unless they are disconnected from the persons (patients) concerned. Both the obstacles to and the opportunities for using/re-using medical data, in particular epidemiological data, in the Netherlands are described in the report on public health data *Van gegevens verzekerd*.[34]

In the social sciences, researchers would like to link personal data from the same person in different datasets in order to gain insights at the micro level.[35] There are, however, strict legal constraints in this area relating to the protection of personal data.

- **Infrastructure and expertise**

Preservation capacity and financial resources are interconnected. Capacity involves more than funding, however; other requirements are physical space, energy management, expertise and

[32] Lyon 2010
[33] ISRIC interview
[34] *Van gegevens verzekerd* 2008, pp. 45-54
[35] Jeurgens interview

infrastructure, such as the Internet capacity. A data repository should have sufficiently trained personnel to handle discipline-specific data.

- **Financial resources**

The importance of financial resources is underestimated. Only recently have both electronic libraries and data repositories attempted to get a grip on the costs of digital archiving. As this is totally uncharted water, there are no models available. Business models have yet to be developed. Digital preservation investments must be made, but with so much uncertainty regarding preservation strategies, it is virtually impossible to calculate prices for the long term.

3.4 Which factors determine the preservation period?

Most data experts claim that it is difficult to determine the preservation period for research data. Without specifying the research area concerned, it is almost impossible to predict, in general, how long research data should be preserved. Contrary to public administration documents and business financial records, there is, according to Harvey, often no legal obligation to preserve research data.[36]
In the Netherlands a legal obligation to preserve research data in most cases exists, but is hardly ever effectuated in practice.

Some sources within specific research areas are quite clear about the preservation period; in the case of earth observation, for example, et ernal preservation is common, since observations cannot be repeated.[37] The same applies to the climate and ocean data used in hydraulic engineering, in particular coastal engineering and management,[38] or to archaeology, where excavations 'destroy themselves' and can never be redone.[39] Similarly, in soil profiles, descriptions and sampling are inherently destructive, hence the need to safeguard established reference collections.[40] In his report, Harvey stresses the abundance of generated research data. According to him, preserving all these data will require significant, although not unaffordable, investments. Researchers consider preservation capacity sufficient, but there may be constraints on preserving new research data or collections.[41]

The IISH's report on the preservation of research data suggests waiting for a period of ten years after a collaboratory has been updated with new data before making a final preservation decision.[42]

In many cases, datasets are generated within projects and made available through a website developed specifically for the project. When the project initiator or funding body does not or is unable to maintain the website significantly after the project has concluded, the website, as well as the underlying datasets, are typically not preserved.[43] Preserving websites is a specific and particularly complicated challenge, especially when the sites contain underlying data. Many websites have been lost after a shorter or longer period of time. For more information on the selection of websites, see Masanès 2006.

3.5 Conclusion

When trying to establish the value of research data for long-term preservation, uniqueness and repeatability are often each others opposites. Peers in each discipline or domain must decide whether the data's uniqueness is significant enough to justify investment in preservation. For those cases where unique data can be (easily) recreated the investment might not be necessary.

[36] Harvey 2007
[37] PARSE, Insight interim report 2009, pp. 10-11; ESA interview
[38] Open Earth interview
[39] EDNA interview
[40] ISRIC interview
[41] ISRIC and OpenEarth interviews; Harvey 2007
[42] IISH 2009, p. 10
[43] Research Information Network, 2008

There are *pre-conditions* that, although they are not the deciding factor in whether or not to preserve research data, must be taken into consideration. They involve issues such as providing sufficient metadata, legal or contractual rights, the availability of an infrastructure, including the necessary staff expertise, and the costs of preservation. Experienced staff from data repositories can advise on the pre-conditions.

4 Selection: the various stakeholders

One important question to be asked when producing selection criteria is: for whom are these criteria being developed? In other words: who will be applying the criteria in practice? It should be noted that there are various different stakeholders to be taken into account. In this section, we will try to identify these stakeholders in relation to the possible roles they may play in selection. A UNESCO report from 2003 states that '[d]ecisions should be based primarily on the value of material in supporting the mission of the organisation taking preservation responsibility'.[44] The term 'organisation' can mean many different things, as we will illustrate here.

The responsibility to preserve research data is now scattered among many players, both within and between disciplines. One of the challenges is how to improve this situation: should disciplines, the researchers themselves, shoulder the responsibility or should funding organisations or universities do this? There are very few organisations that are able to apply appraisal and selection criteria on a disciplinary level. At the funding level, NWO in the Netherlands has started including the responsibility for proper preservation of data in its grant contracts. Universities do not have a policy on digital preservation of their research data, but some research departments or institutes do. A university's digital repository, which is mostly used for publications, has the potential to collect and preserve research data.

4.1 Stakeholders

Throughout the life cycle of digital data, in particular research data, we see that various different stakeholders are active at different times. It is therefore not that easy to make stakeholders responsible for preservation. The data creators are often not the ones who attend to long-term preservation: they are not (or do not feel) responsible for the data after the research project for which they were created has finished. On the other hand, the organisations that are set up to take care of long-term preservation (in particular data archives) have almost no influence on the creation of the data. In an ideal world, these two groups would collaborate closely when datasets are being created and processed. It is common wisdom in digital archiving that the curation process leading to digital preservation should preferably start when the data are created. Data archives now often have to perform rescue and salvage work, although this is gradually changing. Only data created and preserved by the same research institute remain in the hands of the same stakeholder.

As early as 1998, Beagrie and Greenstein[45] concluded that decisions concerning the prospects and costs of preservation are divided over different stakeholders. Funding agencies play a very important role, as they provide the investments necessary for creating the data and thus are in a position to influence the long-term life of the data. In the Netherlands, this role is indeed gradually being taken on by funding organisations, in particular by NWO. Furthermore, cooperation between the various institutes that preserve digital data should be encouraged. There is a clear task here for the Netherlands Coalition for Digital Preservation (NCDD).[46]

4.2 Stakeholders on a disciplinary level

There are not many stakeholders operating on a national disciplinary level in the Netherlands. One of the few organisations is EDNA, E-Depot Netherlands Archaeology.[47] This organisation works for all the archaeologists in the Netherlands, whether they are employed by universities or commercial firms. When an excavation has taken place, all reports and all data objects created must be sent to EDNA. No selection takes place.

[44] Harvey 2007, p. 27; UNESCO 2003
[45] Beagrie and Greenstein 2001
[46] www.ncdd.nl
[47] www.edna.nl

There are a few other data repositories working on a national level in the Netherlands. The national organisation DANS (Data Archiving and Networked Services)[48] collects both historical data and data in the social sciences. The 3TU.Datacentrum[49] at Delft University of Technology collects data from the natural sciences and engineering. It cannot, however, be said that DANS contains all the research data in either the field of history or the social sciences. The same holds true for the 3TU.Datacentrum.

Some disciplines already have or are building an international data infrastructure. In astronomy, there is the 'astronomical information network' of the Euro-VO Data Centre Alliance.[50] This is currently mainly a coordinating body, however. OpenEarth[51] is a coordinating initiative for marine and coastal science. It should be mentioned that European data infrastructures are being constructed in several (broadly defined) disciplines, such as CESSDA[52] for the social sciences, CLARIN[53] for linguistics and DARIAH[54] for the humanities. It is too early to tell whether these infrastructures will or even could play an important role in selecting data, when and if they are actually set up. At the very least, they intend to play a role in determining the pre-conditions.

Selection is not always carried out by peers, especially in data archives. See, for example, the difference in the archaeology data archives of the United Kingdom (selection by peers) and the Netherlands (selection by others). In the UK Data Archive[55] an 'Acquisitions Review Committee' takes selection decisions based on criteria such as reuse value, sample size, copyright, legal or ethical issues.[56]

In short, there are currently very few organisations working at a single disciplinary level that are able to apply appraisal and selection criteria. EDNA could do this, as it covers one discipline at the national level, but it does not do so yet. EDNA's project leader thinks it likely that this will happen at some point as data volumes continue to increase. Selection should be based on a national research agenda for archaeology.

4.3 Funding organisations as stakeholders

Funding organisations can play an important role in the pre-selection and acquisition process for digital data. It is very important that they take on this role. Until recently, this was very rare in the Netherlands. Although the Netherlands Code of Conduct for Scientific Practice and the main funding agents in the Netherlands (such as the European Union and NWO) state that due care should be taken with research data produced by projects, hardly any steps have been taken to ensure compliance with these conditions.

Recently, however, NWO – the main research funding body in the Netherlands – has included requirements on data preservation and access in its contracts, but this requirement has not yet been implemented in all of its various programmes. This is an implicit example of the pre-selection strategy described in section 1.5.3. Peers decide which research projects are to receive funding in the NWO programmes. By doing so they are also, implicitly, selecting the data generated in these projects. Scientific significance and relevance play a very important role in the appraisal process, but there are also other considerations, such as possible relevance for society in general or the importance for Dutch national science or the arts as a whole. One could say that the data produced by some of the most important research projects are being selected for long-term preservation in this way. It is not the case, however, that all the data generated by the most important research

[48] www.dans.knaw..nl
[49] http://data.3tu.nl/repository/
[50] www.euro-vo.org/pub/dca/overview.html
[51] www.open-earth.org
[52] www.cessda.org/
[53] www.clarin.eu
[54] www.dariah.eu/
[55] www.data-archive.ac.uk/
[56] Gutmann 2004, p. 213

projects are being preserved, in particular not data produced in projects funded by the universities themselves.

When it comes to data generated through research funded by international bodies, the situation is somewhat mixed, with some data being preserved because there is an obligation to do so. Another interesting example can be found in the earth observation data collected by ESA.[57] These data are preserved not because of an obligation but because of the increasing demand for that data by both science and the public.

4.4 Research institutes and universities as stakeholders

Dutch universities do not, as yet, have a comprehensive, corporate policy on the digital preservation of research data. Some research departments or institutes both within and outside the universities do have a policy of keeping, or acquiring, data in their field. These data repositories can be either purely institutional or, more broadly, subject-related. Data are selected on the basis of an acquisitions policy, just like museums or archives. Many universities now have a digital repository, but this repository is mostly intended for digital publications. These university repositories have the potential to collect and make data available as well, however.

The acquisition policy of research institutes is aimed mainly at small, specialised data niches in the field in which the institute is active. Only a few institutes have more sweeping acquisitions policies, for example the Nijmegen-based MPI institute on linguistics or Amsterdam's IISH on social history.

Collaboratories, in which scientists or scholars work together by uploading and sharing data and publications, are a special case. Ideally these researchers should decide where, in which repository, the collaboratory should be kept. This could be the institute which is primarily organising and hosting the collaboratory; it could also be another repository or archive in the field.

The case of the collaboratories leads to a wider issue: what should a local research group do when it submits its data to a large international data repository, for example Genbank?[58] Should it preserve the data in its own repository as well? That depends largely on the degree of trust it has in the international repository. The arrangements should be made clear in the relevant data licence agreements.

4.5 Conclusion

The main question is: whose is responsible for preserving research data? That responsibility is now scattered among many different players within and between disciplines. One of the challenges is how to improve this situation: should disciplines, the researchers themselves, shoulder the responsibility or should funding organisations or universities do so?

There are very few organisations that are able to apply appraisal and selection criteria on a disciplinary level. At the funding level, NWO in the Netherlands has started including the responsibility for proper preservation of data in its grant contracts.

Dutch universities do not have a policy on the digital preservation of their research data, but some research departments or institutes do. The university's digital repository, which is used mostly for publications, has the potential to collect and preserve research data.

[57] www.esa.int/esaCP/index.html
[58] www.ncbi.nlm.nih.gov/genbank/

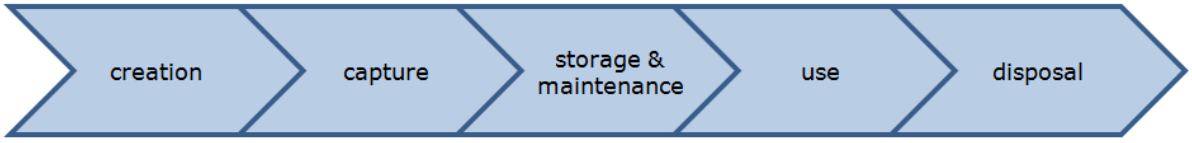
5 Selection points in the digital life cycle

At which points in the digital life cycle should selection decisions be taken? This is the key question of this section, and it requires a closer look at the digital life cycle of research data. What are the best points for taking decisions and by which stakeholders?

The conclusion is that selection at the source, meaning at the moment of creation, is clearly the most preferable option. Selection at later points might lead to all kinds of problems and irreversible and hidden costs. Most of these problems have to do with the pre-conditions. It is also very difficult to predict the potential value and future demand for reuse, especially when the impact of a research project is still unknown. Selection should, in principle, not be dependent on the end of a research project or on the change or closing down of a server. If selection at the source is not feasible, there are other appraisal/re-appraisal points: when transferring data to a data repository, or, in the case of long-range research projects or collaboratories, after a certain period of time has passed.

5.1 Digital life cycle and records continuum

There are many different models of the digital life cycle,all of them are broken down into successive stages, often these five: creation, capture (entered into an information system or database), storage and maintenance, use and disposal.



The number of stages may vary, but all these models are based on traditional paper archives and assume that data (and data records) are first kept for use in the current administration and then may be transferred into archival custody.[59] For paper records, this transfer really is a physical movement from one location to another, for example from a ministry to an archive.

As electronic records are not tied to a physical location in this way, this model has come in for criticism in the digital age. This has led to the new concept of the *records continuum*. To cite Sue McKemmish, who elaborated this concept: 'A continuum is something continuous of which no separate parts are discernible, a continuous series of elements passing into each other. A records continuum perspective can be contrasted with the life cycle model. The life cycle model argues that there are clearly definable stages in recordkeeping, and creates a sharp distinction between current and historical recordkeeping.' The consequence of the records continuum model is that recordkeeping (in the active administrative organisation) and archiving (passive) are processes which are integrated.[60]

5.2 Digital life cycles in practice

What are the practical consequences of these two concepts of digital life cycles for making selection decisions? Simplified, three different approaches are possible concerning the point at which data are archived:

1. *Transfer of the data to external data repositories* such as national and/or disciplinary data archives, or transfer of data to research institutes that host collaboratories or data collections, including data from external researchers (like the IISH or the MPI Nijmegen, for example). Data are transferred from a research group to a data repository at a certain point in time, either during or after completion of the research. This means a change in management and,

[59] Shepherd and Yeo 2003, pp. 5-8
[60] McKemmish 1997

usually, in the physical location of the data.

2. *The research groups retain their own data, in the environment where they were created*. This comes closer to the records continuum model. There is a growing tendency and wish to do so, especially in large-scale research groups or centres where research projects have a long-term, sometimes even indefinite, time scale. The Netherlands Kinship Panel Study (NKPS)[61] is a good example of this: this database is continually updated by successive new 'waves' of data. This has led to the development of the concept of 'trusted digital repositories' and guidelines that attempt to prescribe when such a repository really can be trusted, for example the Data Seal of Approval.[62]

3. *Rescue and salvage: this comes down to doing nothing at all regarding preservation*. This is in fact the situation for many data created in the sixties, seventies, eighties and nineties of the previous century. These data have to be rescued and often reconstructed in special projects.[63] One could also say that this is a form of digital archaeology, which means that no form of data curation is carried out and data reconstruction is only in response to demand.

All three options have their advantages and disadvantages. The second option involves a situation of continuity, but selection decisions must still be made. The data concerned are often stored in larger databases or collaboratories, for example for longitudinal research (long time series). The dataset is often upgraded by newer series or supplemented by other variables. The data have prescribed, standardised data formats. Large research projects often have specialised IT personnel and they can advise on the selection decisions, in particular on the pre-conditions (see section 3.3). The data that are kept are usually not the raw data, but can be seen as published data. Access, however, is often restricted to colleagues in the same discipline or even only in the same specialisation.

5.3 Selection at creation time

Digital archivists agree unanimously that the best way to preserve digital data is to make the initial decisions at the earliest possible stage, preferably at the time that an information system or database is created. Shepherd and Yeo[64] argue that for a retention (preservation) procedure 'rationally made retention decisions should be made as part of a records management programme', in other words a data policy. The advantages of this are:

- it is easier to retrieve those records that are needed;
- it helps avoid inadvertent destruction;
- it eliminates the cost of storing and maintaining unwanted records.

The 2003 UNESCO report on data policy says: 'A policy allows informed, consistent and accountable decisions about appraisal and selection to be made in situations where judgments are subjective and speculative'. This policy should include a statement on re-appraisal principles and a re-appraisal schedule.[65] The same is argued in the recommendations of the Jeurgens Committee, based mainly on experiences in the administrative world. There is no reason to doubt that this would be different in the academic world.[66] The Committee's advice stresses that rescue and salvage operations to retrieve data from obsolete media, hardware or software environments are a tedious and expensive business.

Taking selection decisions at such an early point in time means that the value of the data has to be determined at that early stage as well. This is a top-down approach that applies not to individual datasets but to research projects or at least clearly defined parts of such projects. In principle, the

[61] www.nkps.nl
[62] www.datasealofapproval.org
[63] Balkestein and Tjalsma 2007
[64] Shepherd and Yeo 2003, pp. 146-172
[65] UNESCO 2003, 12.7
[66] Jeurgens 2007, pp. 64-65

selection can be done using the guidelines given in the management summary. There is no guarantee, however, that later insights will not change these selection decisions. The ESA data are an example: only later was there an appreciation (or re-appreciation) of their value, owing to the emerging interest in climate change.[67] Another implication is that in the early design stages of new information systems, an understanding in electronic archiving is required, which may be more difficult to implement in the academic world than in the administrative world. Even in the more bureaucratic administrative world the early involvement of archivists in system design already encounters serious difficulties in practice.

We need only mention the issue of the different stakeholders to realise that it might be quite difficult in practice to involve archivists at an early stage. This is particularly the case when data are to be transferred to data repositories or archives. It might be very difficult or even impossible to take early measures aimed at much later preservation.

Harvey mentions that it is important to 'engage your stakeholders'. However a data repository is organised, the 'perhaps most crucial step' is to involve a specific and relevant user community in the selection decision-making process.[68] This is easier in some disciplines than in others.

5.4 Collaboratories

Collaboratories should be considered a special case within the context of data repositories. A collaboratory can be defined as a team of distributed researchers who, generally for the purpose of a specific research project, create datasets collectively (or share individually collected data).[69] According to the IISH report on guidelines for preserving research data, there are three potential situations in which preservation decisions have to be taken concerning data in a collaboratory:

1. A data collaboratory is hosted by the archiving institution and is still collecting data. The archive will preserve all datasets on condition that within ten years, publications or new grant proposals are based on them.
2. A data collaboratory is hosted by the archiving institute, but has ceased data collection. The archive will preserve all datasets on condition that within ten years, publications or new grant proposals are based on them.
3. A data collaboratory is not hosted by the archiving institute, but wants to deposit its datasets after the lifetime of the collaborator. Preservation of the data will then be dependent on the data's relevance and the availability of proper metadata.[70]

A period of ten years is seen as the time frame for active data. If no new research activities take place within that period, for example new publications or additional research funding, a decision must be taken on permanent archiving. Ten years is an arbitrary period, but it indicates that after a certain length of time, a selection decision must be taken. This is contrary to the idea of selection at the source. Maybe it is not possible to predict how a data collaboratory is going to develop when it first begins; on the other hand, the hosting institute should be able to draw up a preservation policy based on its research profile.

This seems to be in line with an interesting suggestion made by Treloar et al., who, while applying the 'Data Curation Continuum', have constructed two kinds of repositories: the collaboration repository and the publication or preservation repository. The collaboration repository is where researchers actively work with and analyse data. The publication repository is for research that is 'finished', with some of the results being available for public viewing. In the latter repository, there is more emphasis on preservation, Open Access and organisational management. When migrating

[67] PARSE, Insight interim report 2009, p. 10
[68] Harvey 2007, p. 28
[69] IISH 2009, p. 4
[70] IISH 2009, p. 10

data from the collaboration to the publication repository, selection decisions must be made by a combination of human and computer actions.[71]

5.5 Conclusion

It is clear that selection at the source, meaning at the moment of creation, is the most preferable option. Selection at later points may lead to all kinds of problems and hidden costs. Most of these problems have to do with the pre-conditions. It is also very difficult to predict the potential value of and future demand for reuse, especially when the impact of a research project is still unknown. Selection should, in principle, not be dependent on the end of a research project or the change or closing down of a server. If selection at the source not feasible, there are other appraisal/re-appraisal points: when transferring data to a data repository, or in the case of long-range research projects or collaboratories, after a certain period of time has passed.

[71] Treloar 2007, p. 6-7

6 Conclusion: applying selection criteria

We have described which considerations should play a role in selecting research data for preservation and indicated that selection criteria are defined at different levels. This is summarised in the general guidelines, set out in the Management Summary.

We have also indicated that these general guidelines can be applied by a wide variety of different stakeholders with different responsibilities and positions in the academic world. Most of them apply pre-conditions. These are summarised in the guidelines. The pre-conditions should be seen as necessary conditions for preserving research data but not sufficient in themselves. The guidelines can be considered as a *general* set of guidelines for selecting data. Specific guidelines will have to be developed for each discipline/subdiscipline.

At the moment, selection criteria are being applied in only a very small number of disciplines and mostly implicitly. There is a mixed situation in a number of disciplines: some stakeholders within that discipline apply selection criteria, whereas others do not. An example of such a mixed situation can be found in linguistics in the Netherlands.

A fairly large number of disciplines appear to make no provision at all for preserving research data. A national funding organisation such as NWO applies de facto selection criteria, which are not concerned directly with the appraisal of data but have an indirect influence because they serve to select research projects. These criteria are both scientific/scholarly and societal in nature. They are not coordinated with the policies of universities regarding the expansion of specialisations within or between disciplines. There is no national coordination on preserving research data, either within or outside disciplines, with only a few exceptions. There is also no international coordination on this point, again with a few exceptions (such as astronomy). Future European data infrastructures (CESSDA, DARIAH and CLARIN) could play a role in this respect. Attempts to set and apply selection criteria at international level are just as fragmented as those at national level.

To conclude, permanent and open access to research data is certainly not widespread in the Dutch research world. Consequently, one question is whether the issue of selecting data is even the most urgent one at the moment. Van Horik[72] concludes in his NCDD survey that there are only a few organisations active in the field of digital preservation in the Netherlands. The question then is: Is digital preservation as such not a far more urgent issue for most academic disciplines than selection, in view of the general situation in the Netherlands?

It is important to create awareness among stakeholders. At the moment, existing data archives or repositories are not much concerned with the issue of selection decisions. That might change, as some repository managers have indicated. Selection decisions could become an increasingly important topic as the collections grow, regardless of whether these decisions are to be taken at the point of data creation, at the time of ingest (transfer into a repository) or years later. Having a clearly defined data policy, as pointed out in section 4, would be of great help. The financial dimension might also become a key issue in the future. Practical guidelines may become an increasingly vital tool in managing the growing volume of research data.

[72] Van Horik 2009, pp. 37-39

7 Bibliography

All URLs were consulted April 2010

Adams M.O., 'Analyzing Archives and Finding Facts: Use and Users of Digital Data Records', Archival Science, Vol. 7, Number 1 / March, 2007, pp. 21-36, online DOI 10.1007/s10502-007-9056-4

Balkestein M. and H. Tjalsma, 'The ADA approach: retro-archiving data in an academic environment', Archival Science, Vol. 7, Number 1 / March, 2007, pp. 89-105, online DOI 10.1007/s10502-007-9053-7

Beagrie N. and D. Greenstein, A Strategic Policy Framework for Creating and Preserving Digital collections. Version 5.0 (London last updated July 2001), URL: http://ahds.ac.uk/strategic.pdf

Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age; National Academy of Sciences (2009) Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age, www.nap.edu/catalog/12615.html

'Data, data everywhere. A special report on managing information', The Economist, February 27th 2010 www.economist.com/specialreports/displayStory.cfm?story_id=15557443

Dictionary of computing and communications (2003) New York: McGraw-Hill

Doek, A., L. Heerma van Voss, K. Hofmeester, J. Kok and T. van der Werf-Davelaar (2009) IISH Guidelines for preserving research data: a framework for preserving collaborative data collections for future research, Amsterdam, International Institute of Social History

Euro-VO Data Centre Alliance (www.euro-vo.org/pub/dca/overview.html).

Gutmann M., K. Schürer, D. Donakowski and Hilary Beedham, The selection, appraisal, and retention of digital social science data, Data Science Journal, Volume 3, 30 December 2004

Harvey, R. Instalment on 'Appraisal and Selection', DCC Digital Curation Manual 2007 www.dcc.ac.uk/resources/briefing-papers/introduction-curation/appraisal-and-selection or www.dcc.ac.uk/webfm_send/121

Horik van, R. Nationale Verkenning Digitale Duurzaamheid. Inputnotitie sector wetenschap Nationale Coalitie voor Digitale Duurzaamheid 2009 www.ncdd.nl/documents/NCDDinputwetenschap2009.pdf

Jeurgens, K.J.P.F.M., A.C.V.M. Bongenaar en M.C. Windhorst (eds.), Gewaardeerd verleden. Bouwstenen voor een nieuwe waarderingsmethodiek voor archieven, Rapport van de Commissie Waardering en Selectie, Den Haag 2007) www.nationaalarchief.nl/images/3_14750.pdf

Key Perspectives Ltd. (2010) Data dimensions: disciplinary differences in research data sharing, reuse and long term viability, SCARP Synthesis Study, Digital Curation Center

Kuipers, T., and Van der Hoeven, J. (2009) PARSE, Insight: INSIGHT into issues of Permanent Access to the Records of Science in Europe, Survey report, D3.4

Lyon L., Chr. Rusbridge, C. Neilson and A. Whyte, Disciplinary Approaches to Sharing, Curation, Reuse and Preservation, JISC Final Report DCC SCARP, 2010 http://dcc.ac.uk/scarp/scarp-final-project-report.pdf

McKemmish S., 'Yesterday, Today and Tomorrow: A Continuum of Responsibility', Proceedings of the Records Management Association of Australia 14th National Convention, 15-17 Sept 1997, RMAA Perth 1997
www.infotech.monash.edu.au/research/groups/rcrg/publications/recordscontinuum-smckp2.html

Masanès J., Web Archiving, Berlin, Heidelberg, New York 2006

The Netherlands Code of Conduct for Scientific Practice, VSNU 2005
www.dans.knaw.nl/sites/default/files/file/archief/Gedragscode_Wetenschapsbeoefening_VSNU_UK.pdf

NSF Cyberinfrastructure Council (2005) NSF's Cyberinfrastructure Vision for 21st Century Discovery, Arlington National Science Foundation

PARSE. Insight: First insights into digital preservation of research output in Europe. interim insight report September 2009
www.parse-insight.eu/publications.php#d3-5

Research Information Network, To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. Report commissioned by the Research Information Network (RIN), 2008

Shepherd E. and G. Yeo, *Managing records. A handbook of principles and practice,* London 2003

www.slavevoyages.org

Treloar. A, D. Groenewegen and C. Harboe-Ree, The Data Curation Continuum: Managing Data Objects in Institutional Repositories, D-Lib Magazine, Vol. 13 Numer 9/10 September/October 2007
www.dlib.org/dlib/september07/treloar/09treloar.html

UNESCO Guidelines for the preservation of digital heritage, UNESCO, Information Society Division 2003 (National Library of Australia) http://portal.unesco.org/ci/en/ev.php-URL_ID=13271&URL_DO=DO_TOPIC&URL_SECTION=201.html

Van gegevens verzekerd. Kennis over Volksgezondheid in Nederland, Advies Raad voor Gezondheidsonderzoek, Den Haag oktober 2008 http://www.gezondheidsraad.nl/nl/adviezen/van-gegevens-verzekerd-kennis-over-de-volksgezondheid-nederland-nu-en-de-toekomst

Voorbrood C., Data – Voer voor psychologen? Archivering, beschikbaarstelling en hergebruik van onderzoeksdata in de psychologie, Den Haag 2010, DANS studies in digital archiving 4

APPENDIX 1 – List of people interviewed

ISRIC: Ir. Niels Batjes, International Soil Reference Information Centre (ISRIC – World Soil Information)

OpenEarth: Dr. Gerben de Boer, Open Source, Deltares

ESA: Vincenzo Beruti, ESA-ESRIN Earth Observation Ground Segment Department, ESA – European Space Agency, Directorate of Earth Observation Programmes

EDNA: Drs. Milco Wansleeben, project leader for the e-depot Netherlands Archaeology EDNA

Jeurgens: Prof. dr. Charles Jeurgens, National Archives of the Netherlands and professor of archival science, University of Leiden

APPENDIX 2 – Interview Questions

Project Collectioneren van data

Vragen voor interviews opslag datasets

1. Algemeen

Wat is de aanleiding geweest voor oprichting van de organisatie?

Door welke instituten is de oprichting ondersteund?
Ondersteunen deze nu nog de organisatie?

2. Onderzoekers: leveranciers

Wie zijn de leveranciers van datasets?

3. Onderzoekers: gebruikers

Wie zijn de gebruikers van de datasets?

Zijn er beperkingen verbonden aan soort gebruiker (particulier, overheid, academisch, zakelijk) en soort gebruik (onderzoek, beleid, implementatie)?

4. Datasets

Wat voor type datasets wordt opgenomen?

Wat is de aard van de datasets?
- observaties
- experimenten
- veldmetingen
- interviews
- oudere (eventueel gedigitaliseerde) gegevens

Worden de geselecteerde datasets opgenomen voor de eeuwigheid?

Wat is de bewaarstrategie: hoe en door wie wordt de bewaartermijn bepaald?
- wettelijke eisen
- eisen gesteld door financier
- privacy aspecten

Zijn hier (internationale) afspraken over?

Hoe wordt omgegaan met de technische houdbaarheid van de datasets (beschikbare hard- en software, kwetsbaarheid voor storingen)?

Wordt er gebruik gemaakt van een standaard format voor opslag of verschilt deze al naar gelang soort data?

Worden de aangeleverde datasets door uw organisatie bewerkt voordat ze worden opgeslagen?
- controle
- correctie
- conversie

Worden de datasets gedocumenteerd en volgens welke standaard?
Doen de producenten van de datasets dit zelf?

Wat zijn eventuele redenen om aangeboden datasets te weigeren voor opname in het datacentrum?

5. Selectie

Wat zijn de redenen om de datasets op te slaan?
- wettelijke eisen
- bewijsvoering en verantwoording
- hergebruik door onderzoekers (‘data sharing’)
- bij contract vastgelegd met financier
- collectief geheugen (erfgoed)
- uniciteit (geen opslag elders)
- …

Hoe wordt bepaald of aangeleverde datasets waardevol zijn voor langdurige opslag en toegankelijk maken?

Door wie wordt bepaald of aangeleverde datasets waardevol zijn voor langdurige opslag en toegankelijk maken?

Hoe en door wie wordt de bewaartermijn bepaald?

6. Financiële aspecten

Hoe wordt de organisatie gefinancierd?

Betalen de producenten van datasets een bijdrage of worden ze juist beloond?

En de gebruikers, betalen die een bijdrage?

7. Communicatie

Doet uw organisatie actief aan acquisitie of komen de producenten van datasets uit eigen beweging naar uw organisatie?

Heeft de organisatie een marketing strategie zoals
- invloed op zoekmachines om de trefkans te verhogen
- presentatie op beurzen en congressen
- verspreiding van foldermateriaal
- direct mail en telefonisch contact met potentiële klanten
- …

8. Aanvullende vragen

APPENDIX 3 – Full text of interviews

Verslag: gesprek met Ir. Niels Batjes, ISRIC
Datum: 3 februari 2010
Plaats: ISRIC kantoor, WUR Campus

Algemeen

In 1964, heeft de internationale bodemkundige (ISSS) vereniging voorgesteld om een 'International Soil Museum' op te richten ter karakterisering van de belangrijkste bodemtypes van de FAO-Unesco Wereldbodemkaart op schaal 1: 5000 000. Dit voorstel is daarna goedgekeurd door de 'UNESCO General Council'. In 1966, wordt het Internationaal Bodemkundig Museum in Nederland gevestigd, met financiering van de Nederlandse overheid (OC&W). Vanaf 1989 is ISRIC het World Data Centre (WDC) for Soils van de ICSU (International Council for Science).

Het ISRIC is een onafhankelijk instituut met een strategische alliantie met de Wageningen University & Research Centre (WUR) sinds 2002; voorheen gelieerd aan het International Institute for Geo-Information Science and Earth Observation (ITC), Enschede

Het ISRIC omvat 3 hoofdfuncties
- Museum:expositie van bodemmonolieten (profielen), met een onderwijstaak
- World Data Centre: opslag, analyse en wereldwijd toegankelijk maken van bodeminformatie
- Toegepast onderzoek

Het museum en het WDC (samen ca. 60% van het budget) worden gefinancierd uit openbare middelen (Ministerie van OC&W). Het toegepast onderzoek (ca. 40% van het budget) wordt gefinancierd uit projectgelden die via tenderprocedures beschikbaar worden gesteld door o.a. de EU.

Behalve het verzamelen, opslaan, analyseren en toegankelijk maken van bodemgegevens hoort ook het scannen en digitaal beschikbaar maken van bodemgerelateerde rapporten en kaarten tot de taken van het ISRIC. In veel, extern-gefinancierde projecten vindt een uitwisseling plaats van expertise enerzijds en bodemgegevens anderzijds.

De bodemdata zijn verkregen door observatie, experimenten, metingen en analyses. Ook oudere data worden opgenomen in de collecties.

Sinds de beëindiging van het National Soil Reference Collections (NASREC) project neemt het ISRIC zelf geen bodemmonsters meer op grote schaal; wel nog op *ad hoc* basis binnen de lopende projecten. De nadruk ligt nu op het toegankelijk maken van digitale bodeminformatie (zie: http://www.isric.org).

Openbaarheid van gegevens

Voor wetenschappelijke doeleinden kunnen de bodemgegevens vrij gebruikt worden conform de WDC criteria. In afgeleide publicaties hoort de bron te worden vermeld. Via bibliografische databases zoals Scopus, Science Direct en Web of Science wordt nagegaan welke bestanden worden gebruikt en waarvoor om beter op de wensen van de gebruikers in te kunnen spelen.

Voor commerciële toepassingen is toestemming van de Directie vereist; het is niet toegestaan de gegevens van het WDC te presenteren als ‘een nieuw product’.
In sommige gevallen rust er een embargo op bodemgegevens tot het onderzoek (project), waar ze bij horen, is gepubliceerd.

De datasets worden via de ISRIC website en het WDC portaal van de Global Change Metada Directory (GCMD) van de NASA beschikbaar gesteld.

Opslagformaat

De primaire bodemgegevens worden ‘geharmoniseerd’ volgens de internationaal vastgestelde FAO-classificatie van bodemsoorten. Metadata worden aangeleverd in het DIF formaat van GCMD-NASA en volgen de WDC standaarden; hierin wordt ook de ‘data lineage’ uitvoerig beschreven.

Bij het ontbreken van metadata wordt contact opgenomen met de aanleverende wetenschapper/organisatie.

De ruimtelijke bodemgegevens worden opgeslagen als GIS bestanden, hoofdzakelijk in het ESRI ArcGIS formaat, en worden voorzien van een rapportage. Het ISRIC bestudeert de mogelijkheden om gebruik te maken van Open Access GIS software.
Er wordt gewerkt aan standaardisering volgens ISO afspraken en FAO richtlijnen; zo ook voor datamodellen (SoterML, SoilML, GeoSciML, enz.) .

Oudere sets van bodeminformatie die in een ander formaat zijn vastgelegd (bijvoorbeeld dBase, Excel, Access) worden omgezet in het nieuwe formaat.

Kwaliteit

De kwaliteit van de bodemgegevens wordt altijd gecontroleerd volgens gestandaardiseerde procedures. Gegevens van dubieuze kwaliteit worden niet opgeslagen. In het geval dat inconsistenties worden opgemerkt, worden de gegevens geretourneerd naar de afzender ter verbetering.
De gegevens worden geconverteerd naar een standaardformaat voor opslag; deze conversie kan zowel door de leverancier als door het ISRIC worden uitgevoerd.

De opgeslagen bodemgegevens worden voorzien van een tijdsaanduiding; de gegevens van één bepaalde locatie kunnen variëren op verschillende tijdstippen door o.a. de volgende oorzaken:

- Fouten in monstername en analyse
- Variaties in analysetechnieken
- Veranderingen in de omgeving (voorbeeld: pre-Tchernobyl gegevens)

Gebruikers

De gebruikers zijn vooral Universiteiten (wetenschappers, studenten), nationale onderzoeksinstituten en publieke diensten. De gebruikers worden geregistreerd, zo kan het ISRIC de gebruikersstatistieken bijhouden, trends volgen, en daarop in spelen.

Bewaarstrategie en technische betrouwbaarheid

Er worden geen wettelijke eisen gesteld aan de opslag van gegevens. Het ISRIC moet voldoen aan de eisen zoals vastgelegd in het mandaat en de WDC eisen.

Privacy aspecten zijn niet van toepassing (public domain); wel moet er worden voldaan aan de 'Data use and citation' eisen (zie: www.isric.org).

De analoge en digitale bodemgegevens worden in principe opgeslagen voor de eeuwigheid. Om verlies door storingen te voorkomen en om de ontwikkeling van digitale opslagtechnieken te volgen, worden regelmatig back-up bestanden gemaakt door de WUR.
Veel van de gegevens (monolieten, monsters, rapporten, kaarten) zijn uniek en behoren daardoor tot de referentie/WDC collectie.
De bodemmonsters (orde van grootte 1 kg), behorend bij de referentie monolieten, worden opgeslagen in een magazijn annex werkplaats; geprepareerde monolieten worden in het World Soil Museum geëxposeerd. Onderzoekers kunnen op verzoek, tegen vergoeding van de nominale kosten, conform de WDC voorwaarden, een monster krijgen (orde van grootte enkele grammen).

Communicatie

De trefkans in zoekmachines waaronder Google is groot, zonder dat het ISRIC zich daarvoor actief inzet (anders dan het bijhouden van de website en beheren van de collecties/bestanden).
Presentatie op beurzen is niet van toepassing. Wel vindt verspreiding van het gedachtegoed plaats op congressen door mondelinge presentaties, posterpresentaties, verspreiding van foldermateriaal en persoonlijk contact met vakgenoten.
Marketing via direct mail en telefonisch contact vindt niet plaats.

Verslag: gesprek met Dr. Gerben de Boer, Deltares
Datum: 9 februari 2010
Plaats: TU Delft Library

Algemeen

OpenEarth is een open source initiatief van Deltares voor de inzameling, opslag en verspreiding van datasets op het gebied van oceanologie en kustbeheer. Het is een samenwerkingsverband tussen wetenschappers van o.a. Deltares, TU Delft, Universiteit Twente, UNESCO-IHE en IMARES (Wageningen), vooralsnog zonder formele status.

OpenEarth is nauw verbonden aan het Nederlandse Centrum voor Kustonderzoek NCK.
Ook zijn er nauwe banden met de organisatie Building with Nature.
Financiering voor diverse projecten is vooral afkomstig van EU fondsen.

Leveranciers

OpenEarth ontvangt datasets van de volgende personen/instanties:
- Rijkswaterstaat, diverse diensten
- NASA en NOAA
- KNMI (in beperkte mate omdat tot nu toe datasets alleen tegen betaling verkrijgbaar waren, bovendien is het format moeilijk toegankelijk)
- Onderzoekers van de onder 'Algemeen' vermelde organisaties Deltares, TU Delft, Universiteit Twente, UNESCO-IHE en IMARES.

Rijkswaterstaat benadrukt dat de beschikbaar gestelde datasets niet mogen worden gebruikt om evt. claims te ondersteunen.

Gebruikers

De gebruikers van de beheerde datasets zijn wetenschappers die zich met zee- en kustonderzoek bezig houden. Onder andere van de hierboven genoemde instanties, maar gebruik staat in principe open voor iedere wetenschapper of belangstellende in zee- en kustbeheer, vanwege het karakter van openbaarheid.

Datasets

De datasets in beheer bij OpenEarth betreft o.a. de volgende soorten data: zeestromingen, getijdenstromingen, bodemdata, KNMI meteorologische gegevens, chemische en biologische waterkwaliteit
Veel datasets zijn verkregen aan de hand van satellietwaarnemingen.

Openbaarheid van gegevens

De opstelling van OpenEarth is ondubbelzinnig: alle onderzoeksdata uit zee- en kustonderzoek moeten vrij beschikbaar zijn. Een uitzondering betreft data die horen bij nog te publiceren onderzoek, voor deze data wil men graag een unieke code ontwikkelen (DOI) waarmee de data wel kunnen worden gepubliceerd. Als ze voor ander onderzoek hergebruikt worden, kunnen de oorspronkelijke data worden geciteerd.

Er wordt gestreefd naar wederzijdse openheid: onderzoekers kunnen onbeperkt gebruik maken van de producten van OpenEarth (met vermelding van de bron) mits de resultaten uit het hiermee uitgevoerde nieuwe onderzoek ook openbaar worden gemaakt.

Naast openbaarheid van datasets, is ook de beschikbaarheid van vrij toegankelijke software belangrijk voor OpenEarth.
Een veelgebruikt softwarepakket is Matlab, dit wordt binnen TU Delft kosteloos verspreid maar is buiten de campus alleen commercieel verkrijgbaar en daarom voor OpenEarth minder interessant. Er is nu vooral belangstelling voor het softwarepakket Python.

Opslagformaat

Standaardisering van opslagmethoden vindt plaats m.b.v. cdf-cf (climate forcast). De World Meteorolgical Organisation (WMO) heeft deze ontwikkeld en beschikbaar gesteld. Verder wordt gebruik gemaakt van o.a. Fortran. Het gebruik van ASCI formaten wordt niet aangemoedigd door OpenEarth.

Voor metadata standaarden wordt aangesloten bij INSPIRE, een Europese afspraak op het gebied van geodata. Dit t.b.v. de financiering door de EU, hoewel de praktische bruikbaarheid hiervan volgens OpenEarth beperkt is. INSPIRE werkt met xml-bestanden die voor OpenEarth wetenschappers geen meerwaarde hebben.

Kwaliteit

De beoordeling van de kwaliteit van de gegevens wordt volledig overgelaten aan de leveranciers. De gegevens worden niet door OpenEarth gecontroleerd of gecorrigeerd. Ook voor de omzetting naar de standaard formaten doet OpenEarth een beroep op de leverancier. Daar waar de benodigde kennis hiervoor ontbreekt, biedt OpenEarth cursussen aan.
De kwaliteit van datasets hangt nauw samen met de kwaliteit van meetapparatuur en de kennis die de onderzoeker heeft van (de beperkingen en storingen van) de apparatuur.
Er wordt geen gebruik gemaakt van ISO of andere internationaal erkende kwaliteitsafspraken.

Answers concerning the selection of data by ESA re EO data: Earth Observation data, by Vincenzo Beruti, ESA-ESRIN, Frascati, Italy

Answers indicated in italics

1. Do you apply appraisal criteria regarding the ESA data and which ones? Or do you simply preserve everything?

   We do not apply appraisal criteria and decided to preserve anything, provided the data integrity is still maintained at a quality satisfactory for the data use after successful technologies migration. Operationally this is cheaper than applying selective criteria.

2. The case study mentions that ESA conducted a public survey on 'current and envisaged exploitation' of environmental data? Is the outcome of this survey a factor, or even leading in the selection of ESA data for long-term preservation?

   The survey is mainly focused on the interest of users in historical data more than aiming at a selection of the data. The survey was not oriented towards selecting types of data to be preserved but towards understand awareness and interests of users in future data preservation overall. It is difficult to predict the needs of environment data use for the future. All data will be useful in principle.

3. Are technical (data format etc.) factors or the presence of sufficient metadata (context documentation) decisive in selecting data for long-term preservation?

   Metadata and data access are fundamental elements in data availability awareness. Metadata must evolve with user requirements, as well as the data access and product deliveries.

4. Is it right to assume from the EO case study that you mainly preserve data for a wide multi-disciplinary research community and not in the first place (or not at all) for your own organisation?

   Correct, data are for users/customers all over the world in general, by far most of them outside our organisation.

5. Are you applying a cost model? Do researchers have to pay for using the ESA data and is this a factor when selecting data?

   The current trend, especially for scientific application, is to deliver data free of charge or in specific cases at cost of reproduction.

6. Is ESA itself making the final decisions for selecting or not selecting data for long-term preservation or this is done in cooperation with external parties/researchers?

   For the ESA data, it is an ESA decision, due to the importance of the data for humankind. ESA is funded by public funds. ESA, via the LTDP program, coordinates and aims to take a similar approach for all other parties involved in EO data holding from their mission, including possibly commercial partners.

7. Are you preserving raw data or data which have been processed, if later processing takes place anyway?

   ESA preserves the data at the lower level (raw data, in our case satellite telemetry data). In general ESA delivers to users products extracted from the raw data following data processing. Product generation capability is an integral part of the data access and therefore of the data preservation process.

Interview met drs. Milco Wansleeben, projectleider EDNA, m.b.t. selectie bij EDNA

Inleiding

EDNA staat voor e-depot Nederlandse archeologie (zie www.dans.knaw.nl/content/categorieen/projecten/edna-het-e-depot-voor-de-nederlandse-archeologie)

In EDNA, ondergebracht bij DANS in samenwerking met de RCE (Rijksdienst voor het Cultureel Erfgoed), zijn de digitale bestanden opgeslagen met onderzoeksgegevens van Nederlandse archeologen. Het zijn bestanden met primaire archeologische gegevens van opgravingen, regionale verkenningen en materiaalstudies. Het gaat daarbij om reeds afgeronde en gepubliceerde onderzoeksresultaten, waarvan de auteur(s) hun basisgegevens toegankelijk hebben gemaakt voor andere wetenschappers. Het e-depot zorgt voor de duurzame archivering en ontsluiting van alle digitale documentatie van het archeologisch onderzoek.

Er is een principe afspraak gemaakt dat archeologische onderzoeksrapporten als pdf-document worden aangeleverd bij de RCE. De digitale databestanden worden bij DANS in het archiveringssysteem EASY gedeponeerd. Deze data kunnen zich eventueel in pdf-documenten, maar veelal in databases, GIS-, CAD- en meetbestanden bevinden.

Sinds het inwerkingtreden van het verdrag van Malta (1992) worden in Nederland de meeste archeologische opgravingen door commerciële opgravingsbedrijven uitgevoerd, anders dan vóór die tijd. Op dit moment zijn 'in de wetenschap' ca. 100 personen werkzaam tegen 1000 in de commerciële opgravingsbedrijven.

Volgens de kwaliteitsnorm voor de Nederlandse archeologie (KNA 3.1) is deponering van zowel onderzoeksrapporten als data bij EDNA verplicht. Dat geldt zowel voor de wetenschappelijke archeologen als voor de grote groep archeologen die momenteel in de commerciële (Malta) archeologie werkzaam is.

Type data

Wat betreft de aard van de data: bij een opgraving wordt alles wat uit de grond komt of op het opgravingsterrein wordt waargenomen vastgelegd. Deze informatie is, letterlijk, uniek: een archeologische opgraving vernietigt altijd zichzelf, met uitzondering van opgegraven voorwerpen, zoals botresten, scherven etc.,en is daardoor niet herhaalbaar. Daarom wordt alles bewaard van een opgraving. Deze opgravingsdocumentatie is (nu nog) deels digital born materiaal, deels gedigitaliseerd analoog materiaal. Digital born materiaal wordt bewaard; gedigitaliseerd materiaal wat van oorsprong analoog is, zou eventueel niet bewaard hoeven te worden, maar bij EDNA gebeurt dat wel. Hier speelt een preserveringsmotief mee en een hergebruiksvoordeel.

Het systeem ARCHIS wordt gebruikt bij voorstudies. Dit is een landelijke database van de RCE, waarvoor een meldingsplicht bestaat bij onderzoek in de grond. ARCHIS bevat niet meer dan samenvattingen. Voor het tot een daadwerkelijke opgraving komt worden de volgende stadia doorlopen:

1. inventariserend bureau-onderzoek --> stoppen of doorgaan met:
2. inventariserend veld-onderzoek --> stoppen of doorgaan met:
3. opgraving

In al deze drie stadia worden rapporten opgesteld. Bij commerciële opgravingen, op grond van de Malta verplichting, wordt een basisrapportage opgesteld, die alleen ruwe data bevat. Wanneer deze data, geheel of gedeeltelijk, voor wetenschappelijk onderzoek gebruikt worden, is er vervolgens sprake van bewerkte data. Beide soorten data worden bewaard door EDNA.

Deponering van rapporten en data is nu verplicht. Deze worden op dit moment zonder selectie in EDNA opgenomen. Iets anders ligt dat bij het retro-archiveren van data van in het verleden

uitgevoerde opgravingen. Er is niet genoeg capaciteit bij EDNA om deze allemaal te verwerken. Daarom worden hier selectiebeslissingen genomen door de projectleider van EDNA, waarbij praktische overwegingen (de door EDNA te besteden tijd) afgewogen worden tegen het archeologische belang van de opgraving. Vooral de opgravingen die uitgevoerd worden in het kader van grote infrastructurele werken (Betuwelijn, Maaswerken) genereren veel data en kosten daardoor veel (verwerkings)tijd. Milco Wansleeben is van mening dat er eigenlijk geselecteerd zou moeten worden op basis van de Nederlandse Onderzoeksagenda voor Archeologie (NOaA).

Selectie

Afgezien van deze retro-archiveringsactiviteiten selecteert EDNA op dit moment niet. Daar zijn twee uitzonderingen op:

1. hele kleine datasets die in een bepaalde vorm (pdf bijvoorbeeld) bewaard worden, maar nog elders zijn opgeslagen en gemakkelijk herbruikbaar zijn.
2. niet-originele data, dat wil zeggen data die uit een andere bron (Archis) zijn overgenomen, worden niet bewaard.

Samenvattend

Op dit moment speelt selectie nauwelijks een rol bij EDNA. Het is zeker niet uitgesloten dat dit in de toekomst gaat veranderen. Als de binnenkomende hoeveelheden blijven toenemen, kan er een capaciteitsprobleem gaan optreden. Qua techniek (opslag) is dat waarschijnlijk gemakkelijker op te vangen dan wat mankracht betreft. Digitale foto's die zeer veel opslagruimte vragen zouden gecomprimeerd kunnen worden. Uiteindelijk zal bepalend zijn voor de capaciteit van EDNA hoeveel geld daarvoor beschikbaar komt en blijft. Dat zou in de toekomst er op neer kunnen komen hoeveel geld het archeologische veld hier zelf voor over heeft.

Februari 2010

Interview met prof. dr. Charles Jeurgens, verbonden aan het Nationaal Archief en hoogleraar archivistiek Universiteit Leiden, 17 Februari 2010

In dit gesprek stonden twee hoofdvragen centraal:

1. Wat zijn de ervaringen in de praktijk tot nu toe met betrekking tot de aanbevelingen van het rapport 'Gewaardeerd verleden'? Welke elementen zouden daarvan ook bruikbaar kunnen zijn voor waardering en selectie van onderzoeksdata?
2. Zijn er nog andere waarderingscriteria dan hergebruik en controle? In het bijzonder algemene (wetenschaps)historische waarderingscriteria, niet aan één vakgebied gebonden, en wie dat dan moet bepalen.

Ad 1)

Charles Jeurgens vertelde wat er na verschijning van het rapport 'Gewaardeerd verleden' gebeurd was. Het rapport, opgesteld door de onder zijn voorzitterschap staande 'Commissie Waardering en Selectie', is in september 2007 verschenen. In de loop van 2008 is hem gevraagd de in het rapport voorgestelde waarderingsmethodiek in de praktijk te brengen bij het Nationaal Archief. Het ging daarbij vooral om de HMA + (de Historisch-Maatschappelijke Analyse, zie pagina's 25-26 en 44-46 Gewaardeerd Verleden) en de hot spots. De HMA is een methode waarbij d.m.v. interviews met een of meer deskundigen op bepaalde beleidsterreinen getracht wordt inzicht te krijgen in de historische en maatschappelijke ontwikkelingen binnen een specifiek beleidsterrein. Het gaat er daarbij om een terrein, speciaal de spelers daarin, zodanig in beeld te krijgen dat er waarderingsuitspraken gedaan kunnen worden. Vaststellen van 'hot spots' kunnen daarbij een belangrijke rol spelen: dit zijn de meest opmerkelijke interacties tussen overheid en burger, tussen (georganiseerde) burgers onderling of tussen instituties en burgersals actie en reactie op gebeurtenissen of ontwikkelingen. Dit kan zich uiten in ofwel het ontstaan van wetgeving ofwel het optreden van actoren op bepaalde terreinen. Een voorbeeld is de brand in Volendam en de gevolgen daarvan. Het gaat hierbij uitdrukkelijk niet om de incidenten op zichzelf.

Met betrekking tot de hot spots is nu een stroomschema ontwikkeld. Het idee is om de gekozen hot spots te blijven volgen, te monitoren. Er worden twee sporen gevolgd. Het ene is op de toekomst gericht, waarbij volgens de methode van Gewaardeerd Verleden gewaardeerd zal gaan worden, het andere is om voor de (retro-)archivering van een hot spot uit te gaan. Hiermee wordt nu bij milieuzaken geëxperimenteerd. In plaats van een HMA + wordt over een trendontwikkeling gesproken. Deze laatste is voor milieuzaken voor de periode 1975 – 2005 voltooid. Dat zal ook gaan gebeuren voor de huidige periode (2005-nu). De analyse voor de periode 1975 – 2005 is zojuist voltooid; nu moeten de hiervoor relevante archief)delen) geïdentificeerd worden in de organisatie. Dit is een risico-analyse vanuit het perspectief van de overheid. Wat zijn de risico's in politiek, financieel, juridisch opzicht, maar ook dat van de recht- en bewijszoekende burger, indien bepaalde stukken na een aantal jaren niet meer voorhanden blijken te zijn?

Hoe kun je deze trends met wel of niet bijbehorende hot spots nu vaststellen? Deels worden vooral voor het signaleren van trends rapporten van het SCP gebruikt. Deels worden deze, en dat is meer bij het vinden van de hot spots, door het Nationaal Archief zelf én (ministeriële) archiefcommissies geformuleerd. VROM is daarbij al een pilot ministerie, VWS, Justitie, LNV en SZW gaan daar ook aan meedoen. Bij de rechtbanken bestaan er 'aanwijscommissies', samengesteld uit rechters, officieren van justitie en archivarissen, die oordelen op criteria als belang voor de jurisprudentie, maar ook rumoer in de samenleving. De archiefcommissies, zoals voorgesteld in Gewaardeerd Verleden, zijn in opkomst, maar zeker nog niet overal gevestigd. Voorbeeld voor de commissie was de reeds bestaande archiefcommissie bij Buitenlandse Zaken, die incidenten beoordeelt die belangrijk genoeg zijn om de documenten daarvan voor bewaring in aanmerking te laten komen. Het concept zou mogelijk interessant kunnen zijn voor het waarderen van onderzoeksdata. Ook daar kan zowel vanuit de specifieke wetenschappelijke discipline zelf maar ook vanuit andere disciplines (interdisciplinair) én de maatschappij en ook wetenschapsgeschiedenis interesse zijn voor onderzoeksdata.

Privacywetgeving belemmert ook archivering en vooral beschikbaarstelling. De grote bestanden bijvoorbeeld die bij de agentschappen van SZW ontstaan (UWV, SVB) zijn door de koppelbaarheid van persoonlijke gegevens van groot belang voor de wetenschap, mits daar behoedzaam mee wordt omgesprongen.

Een voorlopige conclusie is dat er nog zeer veel gebeuren moet bij de overheid. Er zijn niet alleen de ministeries, maar ook vele uitvoeringsinstanties onder ministeries (zoals bijvoorbeeld bij SZW) en overige ZBO's. Op gemeentelijk niveau leveren vooral de vele gemeenschappelijke regelingen problemen op. Bij een organisatie als TNO blijkt weinig oog voor het maatschappelijk gebeuren te zijn. Een knelpunt is op dit moment vooral de enorme achterstand bij de archivering van de overheid (800 km). Daarnaast zijn er de langzamerhand bekende problemen m.b.t. digitale bestanden: authenticiteit, betrouwbaarheid. Daar zijn geen eenduidige regels voor te geven. Het grootste probleem is de noodzaak om digitale bestanden direct te moeten waarderen op het moment dat ze gecreëerd worden, in plaats van later (zoals nog steeds bij papieren documenten).

Dit laatste geldt uiteraard ook voor wetenschappelijke data. Ook daar is de aanwezigheid van goede metadata, contextdocumentatie, van zo mogelijk doorslaggevend belang. Peer review is vaak een eerste kwaliteitsbeoordeling, zoals in iederr geval in de technische wetenschappen wordt toegepast.

Ad 2)
Geconstateerd werd dat m.b.t. het wetenschappelijke bedrijf het administratief-bureaucratische gedeelte (waarschijnlijk) wel redelijk bewaard wordt en dat onderzoeksdata hier en daar (zeker niet overal) ook nog wel bewaard worden. De koppeling daartussen is echter problematisch. Een zeer interessante bron van informatie rond het wetenschappelijk onderzoek, namelijk de wetenschappelijke communicatie, die tegenwoordig voornamelijk in de vorm van e-mails plaatsvindt, valt geheel tussen wal en schip. Illustratief is de recente ophef m.b.t. de waarde van sommige klimaatdata en de (in dit geval gehackte) communicatie daaromheen in de vorm van e-mails. Voor de genoemde koppelingen spelen een metadata een essentiële rol. E-mails zijn in elk opzicht een zeer problematische bron om te bewaren, ook vanuit privacy overwegingen (scheiden van privé- en zakelijke post). In de administratieve wereld vinden wel registraties plaats van zakelijke e-mails, maar in de wetenschap is er geen begin van registratie door de vrij strikte scheiding administratie – onderzoek(ers) op universiteiten.

Vanuit de archiefwereld gebeurt er bijzonder weinig op dit vlak, afgezien van een rapport van de Rijksarchiefinspectie van enkele jaren geleden.

Als een interessant en belangrijk onderzoeksonderwerp voor waardering en selectie van onderzoeksdata wordt beschouwd het waarnemen van veranderingen binnen het wetenschapsbedrijf. E-mails zouden daarbij een essentiële rol kunnen spelen. Mogelijk zou aansluiting gevonden kunnen worden bij onderzoek van instituten als het Rathenau instituut of de VKS (Virtual Knowledge Studio) die als onderzoeksterrein hebben het waarnemen van maatschappelijke en/of wetenschappelijke veranderingen onder invloed van de technologische ontwikkelingen. Hier zouden speerpunten uitgezocht kunnen worden.