Arch Sci (2007) 7:1–20
DOI 10.1007/s10502-007-9054-6

ORIGINAL PAPER

# Introduction: archiving research data

**Peter Doorn · Heiko Tjalsma**

Published online: 26 September 2007


**Abstract** This article is a general introduction into the special issue of Archival Science on "archiving research data". It summarizes the different contributions and gives an overview of the main issues in this special field of archiving. One of the leading questions is how and why research data archives differ from public record offices. In the past, the developments in these two worlds have been rather separate. There are however signs that they are converging in the digital world. In particular, this can be seen in the areas of metadata and Internet dissemination as these are strongly influenced by the rapid changes in information technology. These changes have also led to important new developments in the infrastructure of research data to which special attention is paid. New concepts such as collaboratories, data curation, Open Access and the Open Archives Initiative are discussed.

**Keywords** Research data · Research data archives · Digital preservation

## Introduction

This special issue of *Archival Science* is devoted to how permanent access to research data can be provided. Like all processes in the world of archives and libraries, the archiving of research data is influenced by the dynamics of modern information technology and is thus undergoing rapid changes. The articles presented here highlight some of the main issues, developments and new methods in this special field of archiving.

In this introduction, we first summarize the various contributions and then introduce the theme "archiving research data" in broad lines. We also investigate how and why research data archives differ from public record offices. In this, we pay special attention to important new developments in the infrastructure of research data.

P. Doorn (✉) · H. Tjalsma
Data Archiving and Networked Services (DANS), P.O. Box 93067,
The Hague 2509 AB, The Netherlands
e-mail: peter.doorn@dans.knaw.nl

H. Tjalsma
e-mail: heiko.tjalsma@dans.knaw.nl

## Contributions

The article by Adams is on the use of data at NARA by both scholars and non-scholars. The composition of the group of users of electronic records has changed, especially since data first became available online. In the first years that NARA made electronic data available, the user group comprised mainly researchers with a background in the social sciences. Most of these users were able to handle the data themselves: they knew the methodology used in collecting, coding and storing the data. This situation changed when the data were made available online and the user community broadened and became more similar to the user group of the traditional paper archives. The use of records that are not online is increasing as a result of the references made to them on the Internet.

In her article, Corti focuses on qualitative data as a special category within the social sciences. The reuse of these data creates special challenges, because until now social science data archives have used exclusively quantitative data. A new methodology has to be developed to deal with these data. This new methodology should be based on a combination of social science methodology and traditional archival descriptions. Qualitative data could enrich social science research in many ways. An additional issue is the question, what is the best place for archiving and disseminating qualitative data—in research (social science) data archives or in the more traditional libraries and archives?

Rasmussen and Blank concentrate on the use of metadata, particularly in the social sciences. They look closely at the history of the "data description" in this discipline concerning basic, dictionary and context information. They underline the fact that from the beginning of data archives, data have been described in a rather standardized way. Especially systematic information on the methods used in collecting the data is of vital importance in the social sciences. The authors also show how this tradition of data description became strongly influenced by technological developments, particularly the Internet, which led to the development of the DDI metadata system. They also pay attention to such new developments as the Nesstar software, which is based on the DDI system but is more than merely a retrieval system, as it can also be used for statistical analysis by way of the Internet. In this way, the DDI is creating networking effects and is part of the new data infrastructure for the social sciences.

In their article, Vardigan and Whiteman show how the OAIS can be applied (mapped) to a data archive, specifically the ICPSR. They concentrate on the ingest stage (the intake of data) and the access stage. Although the OAIS has a technological background (it originates from NASA), the authors conclude that it can be used in a data archive without too many adaptations. They make an interesting point by looking at the designated community, an important notion in the framework of the OAIS, which is now broadening from a community of social scientists to embrace a much wider group of users who need relatively more support in interpreting the data. Data archives have to be adapted to this. There is a strong similarity here with the changes in the composition of the user groups at NARA.

The focus of the article by Balkestein and Tjalsma is on retro-archiving projects; these are rescue and salvage projects in which, long after their creation, electronic data files are collected, selected, formatted and described for long-term preservation. The authors describe two specific projects that were carried out by DANS and its predecessor, as during these projects a new method in retro-archiving was developed: the ADA method.

The article by Iacovina and Todd analyses the degree to which privacy legislation restricts the long-term preservation of identifiable personal data, particularly in the EU, the USA, Canada and Australia. Strict legislation can mean that the use of personal data is very

restricted. Even the web archiving projects of national libraries are affected by this, as their access to many websites that contain some form of personal information is impeded. This issue is complex and has contradictory elements. There is a fine balance between privacy protection and the preservation of personal data. According to the authors, it is only in Italy that researcher and archival ethics codes have been added to the privacy legislative framework.[1] Privacy needs to be integrated with freedom of information and archival regimes, in order to ensure a wider scope of interpretation of what is permitted in the further processing of personal information.

## Research data archives and public record offices

### Historical developments

"Digital" and "academic" are two key words that recur throughout this special issue. Archiving resources from the academic world has always been an activity that takes place at a distance from the activities of public record offices. University archives are usually maintained by the universities themselves, often in the university libraries. In addition, over the centuries university libraries have acquired the collections of professors and distinguished scholars or scientists. These collections consist of correspondence and other private archives and within them there is often no sharp division between private and university property. In some countries, national institutes preserve and disseminate the personal archives of scientists and scholars. An example is the British National Cataloguing Unit for the Archives of Contemporary Scientists (NCUACS), in Bath. Until recently, these archives were mostly restricted to traditional paper material.

When the first data archives were established in the early 1960s, the tradition of archiving university resources within the universities or at least the academic world was maintained. In the case of the data archives, their material consisted of survey data from the social sciences. The Roper Center in the USA became operational and accessible to the public only in 1957, even though it had been founded (by Elmo Roper, one of the founders of survey research) 12 years earlier, in 1945. In the beginning, it was not a data archive as conceived today. The centres that were established in Europe only slightly later, however, certainly were. The first of these was the Zentral Archiv für Empirische Sozialforschung in Cologne (1960); the second was the Steinmetz Archive in Amsterdam (1964). In the USA, the American Inter-university Consortium for Political and Social Research (ICPSR) was set up in 1962 (Scheuch 2003; Doorn 2004, p. 98). Most data archives sooner or later became affiliated with national research organizations or academies. The organizational framework varies considerably from country to country, and each has changed over time. The case of the Danish Data Archives becoming part of the Danish State Archives—albeit as an independent unit—remains quite exceptional in both the North American and the European context (Introduction Danish Data Archives).

The need to create these data archives arose from the fact that computerized data were increasingly being used in research carried out in the social sciences. The data deposited by Elmo Roper in 1945 were on IBM punched cards. One of the main reasons to preserve these computer data was to allow other researchers to check them, just as it necessary for historians to be able to check the sources quoted in historical publications. As the example

[1] In the Netherlands, a researcher ethics code that is in compliance with the privacy laws has recently been established (The Netherlands Code of Conduct for using personal data in academic research 2005).

of the Roper Center as a forerunner shows, these data archives were conceived as survey archives. Right from the very beginning, however, the secondary analysis of the data was another important motive. Elmo Roper was aware that his survey data were, to quote Erwin Scheuch, "vastly underutilized" and had "great historical value". These three motives—verification, secondary use and long-term historical value—were as important then as they are today (Scheuch 2003; Doorn 2004, p 98).

The data archives were followed by electronic text archives in the late 1970s and early 1980s, and from the 1980s on by historical data archives. In 1996, an archaeological data archive was established in the UK (Wise 1997). The text archives originated from the textual, literary and linguistics research communities. One of the first was the Oxford Text Archive, which was set up in 1976 by Lou Burnard. Text archives contain a variety of electronic texts and linguistic corpora. They are organized in the same way as data archives and are either connected with a university, as often is the case in the USA, or with national academic bodies (Doorn 1996, pp. 361–365; Oxford Text Archive). There are fewer historical data archives than social science data archives. They are embedded in rather different ways, although most are now attached to a social science data archive, sometimes (as in the UK) as an independent unit (Doorn 2004, pp. 99–100; Tjalsma 2004, pp. 112–114). The Archaeology Data Service in the UK was established jointly by several British universities and the Council for British Archaeology (Archaeology Data Service).

Two things are important to stress here. First, all these institutes originated in the academic world; their development was almost totally unconnected with public record offices. Second, these various data archives in the social sciences and humanities—collectively indicated as "research data archives"—were the very first institutes to handle electronic material. This might explain why there was surprisingly little contact between research data archives and public record offices until the latter started to handle electronic data. Nowadays, this divergence is even more remarkable than it was in the 1970s or 1980s. After all, European public record offices have effectively become involved in maintaining digital data—reluctantly, it seems, as they started to do so later than the National Archives and Record Administration (NARA) in the USA, which was the first public record office to start archiving electronic records (Ambacher 2003; Doorn 2004; Adams in this issue).

## Digital preservation

There is, at last, widespread awareness that our digital heritage is in danger. In the last 20–30 years, computer technology has brought us a digital and almost totally new information world, one that has unforeseen possibilities. Here, it is probably sufficient to mention only the Internet. Information retrieval and storage have changed in such a dramatic way that the effect can be compared with the effect the introduction of the printing press had centuries ago. Moreover, we have certainly not yet seen the end of the developments. There is, however, one dimension that so far seems to have been overlooked in this revolution: the persistence of information. The preservation of digital material does not have a logical place in modern information technology, which is very much focused on the rapid and easy retrieval of data, regardless of their format or physical distance. Most Internet applications deal with (and in fact are designed to handle) the most recent data, not those of yesterday or the day before.

Consequently, those whose interest, mission or job it is to keep information available for a longer period, for whatever reason or to fulfil whatever obligation, have to overcome

various obstacles. These obstacles are well-known and are identified in the professional literature (including such journals as *Archival Science*). We shall therefore confine ourselves to summarizing the situation.

A major obstacle in providing permanent access to digital data is the rapid obsolescence of both hardware and software (media, platforms, operating systems). In addition, it is becoming increasingly clear that the long-term preservation of digital material implies more than just solving these technical problems. There are a wide range of other issues that are far more difficult to resolve. These issues concern, for example, metadata (information *on* the data), management, property (who controls or is responsible for the data), authenticity, privacy protection and appraisal. The whole range of issues is to a large extent familiar to the "traditional" paper archivists and is therefore neither new nor characteristic of the preservation of digital information alone. Nevertheless, in digital preservation all these issues acquire new and unexpected features, in both a technical and an intellectual sense (Dollar 1999, pp. 11–43; Jones and Beagrie 2001, pp. 19–20). In 2000, Seamus Ross distinguished some other factors that are "putting digital memory at risk", such as fear of legal action, data protection directives (in particular that of the EU) and protection of intellectual property rights. He also noted viruses and encryption as potential dangers. The issues he mentioned are all typical "new" factors related to the nature of digital information and especially the Internet (Ross 2000, pp. 22–23).

This is a relatively new area of concern and research, and few information technologists are interested in it. Preserving digital information over a long period is not part of their "natural habitat". As they do not have the drive to tackle these issues, the solution must come from the information providers themselves, that is, the archivists and librarians. They must take the lead – an understanding that is not yet shared by everybody. For most archivists and librarians, the majority of whom do not have a technological background, information technology is a world to which they need to get accustomed. Although things are gradually changing for the better, this helps to explain why the whole adaptation process is taking place rather slowly.

There is certainly a general (and, at times, vague) understanding of all these issues. Articles are published and conferences and workshops are held on the subject, in recent years at an increasing frequency. Nevertheless, there are doubts about the overall effectiveness of all these awareness-raising activities, however praiseworthy in themselves they may be. The question is whether the general concerns on this point are of significance outside a relatively small group of library, archive and information specialists. Even within this professional group, convinced as most of its members may be of the threats, there is a lot of uncertainty. There is often a passive attitude of "wait and see", despite an increasing number of pilot projects as well attempts to construct infrastructural facilities (Schürer 2001). This can be explained by the fact that there are no proven solutions to the problem of the long-term preservation of digital data.

In the world of academic research, however, there is awareness of the problem. The most concerned are those who work in research data archives, data libraries or university libraries. Furthermore, the organizations that fund research—namely either universities and research institutes, or national research bodies—are alarmed. Investments in data collection and digitization are wasted if the databases that are created become neglected after a number of years. Also researchers are worried, particularly about how safe their intellectual creations are. How long will the articles they publish in electronic scientific journals survive? And what about the data that are at the basis of these publications?

Added to these feelings of unease is the deep division of opinion on what the main strategy in digital preservation should be. Experts strongly disagree with each other here.

A debate has been going on for some years now between those who support emulation and those who support conversion or migration as a strategy for preserving digital material. In the first strategy, the basic idea is that the original operating system and/or the original software can be emulated (reproduced) in newer software environments, so that the original digital documents do not need to be changed and one can always read them again. It might only be necessary to copy them to new media (media renewal). In the latter strategy, following the definitions given by Charles Dollar (1999, pp. 29–30) the electronic records are exported to other, up-to-date software programs or to neutral, software-independent formats (ASCII, UNICODE, XML) without loss of content or structure. The discussion of the pros and cons of these strategies has intensified since David Bearman published a very critical reaction in 1999 to a report issued in the same year by the American information scientist Jeff Rothenberg, at that time a consultant to the RAND Corporation. In the report, Rothenberg declares himself to be an outspoken advocate of emulation. Bearman on the contrary is a pronounced opponent of this strategy. In his own words: "Electronic records that are not moved out of obsolete hardware and software environments are very likely to die with them." (Rothenberg 1999; Bearman 1999). A special form of emulation – one that might be described as a pragmatic middle way between emulation and conversion – is the concept of the "universal virtual computer", as developed by Raymond Lorie from IBM Research (Lorie 2002).

There are also other strategies (Jones and Beagrie 2001, pp. 102–114). One is backwards compatibility (the automatic upgrading to newer versions of software and operating systems), a possibility offered by a growing number of software packages. Another is hardware preservation. Yet others use permanent identifiers (for web pages), print out digital documents on paper or perform "digital archaeology" (Ross and Gow 1999), which is rescuing files from obsolete digital environments when the need has risen to do so. The last-mentioned option is not really a preservation strategy.

The question which strategy to choose is in fact the most fundamental one in digital preservation. However, so far the debate on this issue has been mostly of a theoretical nature. The solution will therefore evade us for quite some time to come. The implications for large-scale environments (such as libraries or archives) that arise from choosing one of these strategies are still uncertain. On the other hand, both practical experiments and research projects have been going on in this field. For example, in the national deposit library of the Netherlands—the Royal Library (KB) in The Hague—experiments are taking place with the universal virtual computer concept (Digital preservation at the Royal Library of the Netherlands; Oltmans and Van Wijngaarden 2004), while research data archives have built up a lot of experience in using conversion methods.

## Separate developments

So far the common elements in digital preservation have not helped to bridge the gap between research data archives and public record offices. Research data archives developed in relative isolation, and continue to do so even though public record offices have also moved into the digital age. The professional organizations have always been apart from each other. For public record offices, there is the International Council on Archives (ICA); for data archives, various international organizations were established in the 1960s. 1976 saw the founding of CESSDA (Council of European Social Science Data Archives), originally in reaction to a perceived threat of the USA asking European polling organizations to deposit their data in the USA and thus creating a "world data archive". CESSDA

is now the main international organization of data archives, for mutual consultation, deliberation and initiating projects. IASSIST (International Association for Social Science Information Services and Technology) is the international professional organization for individual data archivists. It organizes annual conferences. Text archives and historical data archives do not have such organizations, although there are some rather loose and hardly formalized connections with scholarly international organizations in their field, such as the Association for Literary and Linguistic Computing (ALLC), the Association for Computers and the Humanities (ACH) and the Association for History and Computing (AHC) (Scheuch 2003, p. 392).

The same isolated development can be seen if one looks at two metadata systems in which both communities have invested much, again totally autonomous from each other. The International Council on Archives (ICA) played an important role in developing ISAD-G for public record offices. The Study Description Scheme (SDS) and, later, the Data Documentation Initiative (DDI) were created for data archives. These two systems, however, are not comparable with each other. ISAD-G was originally a description system designed for paper archives. It places heavy emphasis on the field of provenance: from which part of the administration do these specific records come? ISAD-G—which is strongly focused on administrative records and is not designed to deal with digital data—is hardly known let alone used by research data archives. The DDI on the other hand is designed for electronic research data. However, it is in principle possible to make the ISAD-G suitable for electronic files, as Shepherd and Smith concluded in 2000 (Scheuch 2003, p. 393; Shepherd and Smith 2000; Van Ballegooie and Duff 2006; ISAD-G; DDI; Rasmussen and Grant in this issue).

The divergence between the two different types of archives is perhaps not that surprising. In the digital world it would seem that all digital data and documents are the same and that consequently the preservation problems are all the same. MS Word files are MS Word files, irrespective of their content, although the institutional context differs. There is a diversity in approach, urgency and priorities that has to do with the different missions of the institutes involved. For our purpose here, we can categorize them into public record offices, research data archives, depository and academic libraries, and other memory institutions. "Other memory institutions" form a special category, including institutes that archive photographic, sound, moving images, and many other kinds of collections, for instance statistics in statistical offices (van Horik 2005, p. 15). Most of the material in public record offices and in national depository libraries is there because of legal obligations.[2] This is not the case with data archives and the other memory institutions, where collections mostly consist of voluntary gifts or contractual (permanent) loans.

In a research data archive, data are kept for the reasons indicated above, namely verification and reuse. For scholars, the form in which these data are kept is of secondary importance; the only important thing is to have the original data. This explains why, generally speaking, the conversion strategy is mostly followed in data archives. For libraries, particularly depository or academic libraries, that want to preserve the now digital cultural and intellectual heritage of the nation, it is far more important to remain as close as possible to the electronic original. It is essential to preserve not only the content but also the form. Thus, deposit libraries and other memory institutions are interested in testing the emulation strategy in which documents keep their original *look and feel*. For public record

[2] Formally speaking with the exception of the national deposit library of the Netherlands (the Royal Library): there is no legal obligation to deposit a copy of a book in this deposit library. In practice, however, there is no difference from other countries.

offices, it seems that an important issue is the authenticity of the digital documents, as archival records have a legal dimension. This creates problems, because the terms "authenticity" and "original document" work out differently in relation to digital documents than they do in connection with paper documents (Dollar 1999, p. 26; Authenticity in a Digital Environment 2000; Duranti 2005).

## The records continuum in the research world

The concept of "electronic records" illustrates the separation between public record offices and research data archives. This concept was developed by public record archivists as a reaction to the social science data files, which are described by Terry Cook as "one-time, one-shot statistical data files". However, these data files were the only working models available to the first public record archivists who, in the 1970s and 1980s, started thinking about how to archive electronic files. What these one-time files lacked was the element of change over time, of business transactions carried out within information systems; in short, the process dimension. Missing was "the full context and functionality over decades and centuries" (Cook), described by its creator, as well as the principle of provenance. Consequently, content, structure and context are the attributes of electronic records. Electronic records "do not exist "of course" as physical entities" (Dollar), meaning that in the electronic age one has to think of virtual archives, containing virtual records (Cook 1999, pp. 58–61; Dollar 1999, pp. 22–26; Upward 2005).

Dollar looked back over the last 30 years and distinguished three stages in the way archivists in the USA conceived the archiving of electronic files:

1. Rescue and salvage ("a stopgap measure that focused on the back end of the records" life cycle").
2. Focus on information systems.
3. Functional analysis and record making/record keeping systems.

The last-mentioned is the most mature form of preserving electronic records. This is a useful periodization, both for public records offices and research data archives, even if in Europe the timing might deviate from that of Dollar for the USA (Dollar 2003, pp. 140–143; 2004).

Archivists would like to keep these electronic records as much as possible in their original context with their original structure in "record keeping systems", and not have them mixed up with document management systems. That means keeping them in the information systems in which they were created. The "records continuum" concept is built upon this approach. The basic idea is that records can function both actively in the organization in which they were created and passively as part of an archive. There is no question of transferring or capturing records from an active environment into a passive archive (McKemmish 1999; Shepherd and Yeo 2003, pp. 9–10, 122–123; Doorn 2004, pp. 100–102; Dollar 1999, pp. 23–26; Dollar 2004, pp. 23–26).

The records continuum concept suggests that electronic archivists should be involved in the early design stages of new information systems (McKemmish 1999, pp. 195–210). This is an early understanding in electronic archiving that is difficult to materialize. In practice, archivists try to rescue the records some time after their creation, the "salvage and rescue" stage (Ross 2000, p. 13).

At first sight, the notion of electronic records seems to be less relevant to research data archives, electronic libraries and the other memory institutions. "Proactive" archiving

seems far more difficult to implement in the far from bureaucratic, often chaotic or anarchistic research environments than it is in the administrative world. There are, however, clear signs of a trend towards moving earlier into the life cycle of research data files. This is not without problems. Beagrie and Greenstein concluded in their 1998 study that one of the main problems in preserving digital documents is that there are different stakeholders during the life cycle of an electronic document. The data creators are often not the ones who attend to the 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 organizations that are set up to take care of the long-term preservation (e.g. data archives) have hardly any influence over the creation of the data. Their operations often have the character of rescue and salvage work (Beagrie and Greenstein 2001).

Beagrie and Greenstein 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 their study, Beagrie and Greenstein recommend making these funding organizations aware of their potential key position for digital preservation. Cooperation in the field of digital preservation between the different memory institutes involved should be encouraged (Beagrie and Greenstein 2001, pp. 1–3).

Another handicap in this regard is that for all digital archives the words of Maggie Jones and Neil Beagrie are still valid: "... costs for both technical and organizational infrastructure are still not well defined." Although the fact that there are different stakeholders plays an important role here, it is difficult if not impossible to separate preservation costs from general infrastructural IT costs like access to data (Jones and Beagrie 2001, p. 28; Ashley 2000).

## Metadata

There are more overlaps between these different memory institutions. For example, the Open Archival Information System (OAIS) reference model is used in a growing number of data archives, both libraries and public record offices. This is a very general reference model for archiving various materials, ranging from electronic publications to data. It includes all the processes needed but only in the form of a framework. Further specifications are necessary to make the execution of the many archiving tasks possible for a variety of archival organizations. This has been done in the past for depository libraries, for example in the NEDLIB project. Orientation towards the OAIS is now taking place in data archives, too (NEDLIB project; Vardigan and Whiteman in this issue).

In this respect it is especially interesting to look at the UK, where the British Data Archive (UKDA) and the British National Archives are working together to see whether they – two quite different organizations—can both use the OAIS. The background to this initiative is that the UKDA was designated in January 2005 as a "legal place of deposit for the National Archives". The general conclusion of the assessment of how far the UKDA and the National Archives are "OAIS-compliant" was that they are certainly compliant in broad lines but that it is very time-consuming to map all the details of the functional model. Two interesting points were made. One is the importance in the OAIS model of "the strong link between the user community and the way the material in the archive should be described and preserved". It was concluded that it is difficult to limit user groups or communities as narrow as the OAIS does. The UKDA and the National Archives are

simply not able to identify clearly described and homogeneous user communities. Another important (and unexpected) finding was that the use of the OAIS as such is very helpful as a means of communication between these two archives with their very different organizational backgrounds (Beedham et al. 2005, pp. 4–8, 81–84).

There are clear trends of convergence not only in organizational matters but also in the area of metadata. We mentioned the different metadata schemes stemming from the two archiving worlds. Digital material and particularly the Internet have led to the development of newer, very popular metadata schemes, which originated outside the world of public record offices and research data archives. To a certain degree, they can already be considered as standards.

The Dublin Core metadata set has seen an enormously wide application for all types of information put on the web. The set consists basically of fifteen fields. Originally it was intended to provide metadata for electronic publications on the web, but can in fact be used for metadata for all kinds of digital files and objects. This is both its strength and its weakness. This summary information is mostly sufficient for discovering and finding data on the web, but for researchers who are in the process of deciding whether or not these data can be useful for them, these metadata do not offer enough detailed information.

Like OAIS, METS—which originated in digital libraries and archives—is a framework system, but what it "wraps" can be very different underlying metadata. METS is specifically designed for documenting and keeping together material that has been digitized and is scattered around in different files. The conclusion of the UKDA/National Archives assessment was that it was not suitable for born digital records, which can be described by the existing archival metadata systems (Beedham et al. 2005, pp. 79–80).

## The memory institutions: approaching each other?

Memory institutions are gradually converging. This was not the case in the beginning, when the research data archives were developed within the academic world, rather isolated from other archival worlds. Especially the Internet has had an important equalizing effect in making accessible very diverse collections held by memory institutions with different backgrounds and missions. Metadata schemes constructed for the Internet, as well as the overall organizational reference framework OAIS, can be used by all these various institutes. This also applies to such other aspects as copyright or privacy issues, or the emergence of persistent identifiers for websites. Even in the acquisition phases, the same trends are observable along the concept of the records continuum, leading to archiving from their creation on. The memory institutions have become aware of how common many of the different digital challenges are. In the UK, this has led to the foundation of the Digital Preservation Coalition (DPC), in which various memory institutions are working together (Digital Preservation Coalition). In the Netherlands, DANS and the Royal Library have taken the initiative to create a National Coalition for Digital Longevity.

The records continuum, on the other hand, works out differently for the memory institutions. It is precisely on this point that the different contexts of these institutions can be seen. The situation of research data archives is fundamentally different from that of public record offices: the latter work in clearly defined legal contexts, whereas the context of research data archives varies from country to country in both legal and practical terms. In most cases, however, research data archives operate from within the academic world and are totally dependent on what the academic world itself wants. It is useless to maintain research data archives if academia does not see the need for them, especially now that

research data archives have become aware that it is important to become involved in the creation stages of an electronic data file and to remain involved throughout its life.

These two developments—placing the responsibility for preserving research data for the long term in the hands of the researchers themselves and the need to provide data with lifelong care—have led to new ideas on the data infrastructure.

## New infrastructures for research data

In the light of the rapid developments in science and in information technology, the demands on data archiving are constantly changing. The following is a concise overview of the current trends.

Many research groups and academic institutions, including university libraries, have started their own data activities on the web. Although not all of these activities will turn out to be permanent, some will evolve into de facto electronic archives. Institutional repositories, open archives, virtual collaboratories, data curation centres, e-science, data hubs, cyberinfrastructures, research infrastructures and other terms have been coined to indicate the proliferation of new initiatives. It is clear that there is no longer a monopoly on providing data services for research, unlike in the time of mainframe computers. In fact, any researcher with a PC that is connected to the web can now be a data supplier. These developments in the emerging grid pose new challenges to existing data archives, which are confronted with fundamental questions about their functioning and tasks. How can the supply of data best meet the demand? What services do research communities need? How can these services be organized effectively? How can tasks be divided, and how can one interconnect the disparate and distributed data activities?

Although considerable results have been achieved by research data archives in recent decades, the renewal and improvement of data infrastructure in several countries and internationally is high on the agenda scientific organizations. New initiatives and innovations in data archiving seek to:

- Upgrade the technology to achieve higher efficiency, for example by promoting self-archiving, simplifying administrative procedures and minimizing manual work in creating metadata.
- Link up scattered activities, for example by creating portals, networks or unified service desks in which the distributed resources can be found more easily and more transparency is created.
- Stimulate open access to publicly funded data (administrative or registration data), for example in collaboration with statistical agencies.

### Central versus decentral organization

In several countries, new data archives have been established and/or existing ones have been upgraded. There is also a trend to integrate data facilities into wider networks. For example, in 2003 the UK's Social Science Data Archive, which had been founded in 1967, became part of the wider Economic and Social Data Service (ESDS)—a national data service that provides access to and support for an extensive range of key economic and social data, both quantitative and qualitative, spanning many disciplines and themes. In the arts and humanities, data centres such as the History Data Service and the Oxford Text

Archive were integrated into the Arts and Humanities Data Service as early as 1995. The AHDS also has data centres on archaeology, the visual arts and the performing arts (Economic and Social Data Service; Arts and Humanities Data Service). There are developments in science as well in the field of data sharing (Large-scale Data Sharing 2005).

In the Netherlands, Data Archiving and Networked Services (DANS) was established by the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organization for Scientific Research (NWO) in 2005. It continues the tasks of the former Steinmetz archive for the social sciences, the Netherlands Historical Data Archive and the Scientific Statistical Agency. DANS is tasked to become the national organization responsible for storing and providing permanent access to research data for the humanities and the social sciences. To this end, DANS collaborates with researchers and encourages them to work in partnership with one another. DANS operates as a network, with a centre responsible for organizing the data infrastructure.

DANS aims to create new data archives in areas where such facilities do not yet exist. To this end it sets up topical development programmes (TOPs) in collaboration with research teams, on the initiative of the academic world. One initiative of this kind was taken in the field of archaeology, where an Electronic Depot of Netherlands Archaeology is now emerging (EDNA Project; Electronic Depot of Netherlands Archaeology). In the social sciences, work is in progress to link up existing longitudinal surveys and to enrich them with administrative data from official registrations. Obviously, collaboration with Statistics Netherlands (Centraal Bureau voor de Statistiek; CBS) is intensified to supply remote and on-site access to the Social Statistics Database, a conglomeration of government records and sample surveys used to compile inter alia the 2001 census.

As more thematic data archives emerge, the emphasis will shift to the network of trusted data repositories. This will help bring about better guarantees of involvement by data suppliers, and the better coordination of supply and demand, although central facilities and a central organization will still be required. Thanks to the Internet and the upcoming generation of data grids, it is now possible to establish virtual links between distributed stored data sources. A new electronic archiving system, DANS EASY, will make the archiving and dissemination of data more efficient (DANS EASY).

## Towards national and international data strategies

The increasing scale and internationalization of research requires more attention for international data services. For instance, in the humanities and social sciences the investments in long-term, international data collection and digitization projects need guarantees for preservation and permanent access that surpass the responsibility of national data services. Currently, the European activities of such centres are funded on a project basis and carried out as voluntary activities. Robust, truly pan-European and other international data infrastructures for the humanities and social sciences hardly exist.

In specific areas of research interests, and mainly within the EU, North America and Australia, considerable efforts have been made to develop and promote comparable international data sets and to establish better arrangements for sharing data resources between countries. Despite the growing need for social scientists to adopt a global approach to research in areas such as trade, health, security, social infrastructure development, climate change and the transmission of disease, in many parts of the world the data required to address such problems are not readily available.

In the natural and biomedical sciences, the development of data for research purposes is shaped directly by the research issues that are pursued. For the social sciences and humanities a more eclectic approach is characteristic, in which data resources are developed without a specific view of or agreement upon the future research agenda. Several countries are now working towards such research agendas, of which the UK's National Strategy for Data Resources for the Social Sciences is an excellent example (National Strategy for Data Resources for the Social Sciences). Key elements in a national data strategy are the identification of future research needs and the establishment of mechanisms that allow stakeholders to consider the potential gains from cooperation in planning the data resources required to meet these needs.

## Cyberinfrastructure and research infrastructure

The National Science Foundation (NSF) in the USA has launched the "cyberinfrastructure initiative" to facilitate the development of new applications, to allow applications to interoperate across institutions and disciplines, to ensure that data and software acquired at great expense are preserved and easily available, and to empower enhanced collaboration over distance, time and disciplines. In its current vision document, the NSF states that: "Cyberinfrastructure integrates hardware for computing, data and networks, digitally-enabled sensors, observatories and experimental facilities, and an interoperable suite of software and middleware services and tools." (NSF Cyberinfrastructure 2006). The cyberinfrastructure is intended not only for the hard sciences and technical engineering, but also for the humanities and social sciences (Our Cultural Commonwealth 2006). It includes investments to create, disseminate and preserve scientific data, information and knowledge.

In Europe, too, there is increased awareness of the creation of new "research infrastructures" for the European research area. The European Strategy Forum on Research Infrastructures (ESFRI) published the first European roadmap for new, large-scale research infrastructures in October 2006. This roadmap identifies 35 large-scale infrastructure projects that are at various stages of development and lie in seven key research areas, including the social sciences and the humanities. All proposals for new research infrastructure in this field concern the creation, archiving and dissemination of data resources. Moreover, European networks of funding organizations for the humanities (HERA-NET) and social sciences (NORFACE) are developing new strategies for data infrastructures.

## Data curation

Data infrastructure refers to more than just data archiving. It includes care for the data from the moment of their creation onwards. The term "data curation" was introduced in the UK a few years ago. The Digital Curation Centre in the UK holds that:

> ... digital curation is about maintaining and adding value to a trusted body of digital information for current and future use. The digital archiving and preservation community now looks beyond the preservation, cataloguing and cross referencing of static digital objects such as documents.

The ambition of the DCC is to manage data "before it was created, right through to the end of its "life cycle"' (Digital Curation Centre; Rusbridge et al. 2005).

## Institutional repositories

Probably the first but certainly the most famous repository was created by Paul Ginsparg in 1994 as an archive of pre-prints of scientific articles. On 11 September 2006, arXiv.org provided open access to 383,624 e-prints in physics, mathematics, computer science and quantitative biology. The Scholarly Publishing & Academic Resources Coalition (SPARC) defines "institutional repository" as digital collections that capture and preserve the intellectual output of a single or multi-university community (Crow 2002). The content of repositories does not have to be limited to publications (electronic texts): they can also contain other digital objects, such as video files or databases (Heery and Anderson 2005).

In the Netherlands, the higher education and research partnership organization for network services and information and communications technology (SURF) has been very successful in stimulating the creation of academic institutional repositories, which are accessible through the DAREnet portal (DARE = Digital Academic Repositories). The repositories all comply with the OAI protocol and use Dublin Core as the minimum standard for metadata. Three repositories are primarily oriented towards data in the fields of archaeology, education research and hydrology, respectively. In 2006, an international partnership called DRIVER (Digital Repository Infrastructure Vision for European Research) started to build a large-scale public infrastructure for research repositories across Europe.

## Open Access and the Open Archives Initiative

The primary purpose of digital research archives and repositories is not to preserve materials but to provide access to what is preserved. A common understanding has emerged that the access to research output (publications and data) should be as open as possible, taking into account the legal framework of intellectual property right and privacy regulations. The Budapest Open Access Initiative of 2002 formulates the principle as follows (Budapest Open Access Initiative 2002):

> ... by "open access" to literature, we mean its free availability on the public Internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the Internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and consulted.[3]

In January 2004, the governments of the OECD countries adopted the Declaration on Access to Research Data from Public Funding, in which open access is also the primary aim. Although open access strives for free access without barriers, this does not mean that all research data and publications will be up for grabs. New licences have been developed that offer an alternative to full copyright. For instance, Creative Commons licenses provide a flexible range of protections and freedoms for authors, artists and educators (Creative Commons). They have built upon the "all rights reserved" concept of traditional copyright

[3] Related are the Bethesda Statement on Open Access Publishing (2003) and the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003).

to offer a voluntary "some rights reserved" approach. Offering work under a Creative Commons license does not mean giving up copyright. The goal of Science Commons is to promote innovation in science by lowering the legal and technical costs of sharing and reusing scientific work, and removing unnecessary obstacles to scientific collaboration by creating voluntary legal regimes for research and development.

The primary aim of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is to facilitate access to electronic files in institutional interoperable repositories for metadata sharing, publishing and archiving. The protocol of the OAI originated in the e-print world, where the term "archive" is mainly used to indicate a repository for electronic publications (although the content can also be data, images, video, etc.). Metadata from any system as used by a community can be incorporated, but to reach a basic level of interoperability it is necessary to specify a least common denominator, for which an XML scheme of "unqualified" Dublin Core metadata elements is used. The institute disseminating the data on the web is in OAI terms called the "data provider". The metadata can in principle be accessed by any user, but typically a "service provider" harvests the information and provides it and additional services to the end users (OAI-MPH).

## Collaboratories and journals with a data availability policy

The term "collaboratory" was coined in 1989 by William Wulf to describe a "center without walls in which the nation's researchers can perform their research without regard to geographical location—interacting with colleagues, accessing instrumentation, sharing data and computational resources, and accessing information in digital libraries" (Wulf 1989, 1993). Although there are many newer definitions, research into over a hundred collaboratory-like efforts led to the identification of four key activities for research collaboratories: (a) distributed research centres, (b) shared instruments, (c) shared, community-wide data systems and (d) an open system for community contributions of materials.

In many collaboratories, data, publications and computer-supported research tools are presented together in a working environment on the web, allowing research communities, small or large, to work together in a virtual way. Collaboratories originated in the sphere of science and technology, where there are now many collaboratories. The social sciences and humanities are following this trend, although on a somewhat smaller scale (Ross 2000, pp. 7–8; National Collaboratories 1993; Human Genome; Collaboratories in science; Collaboratories example Humanities).

An increasing number of journals have formulated a "data availability policy" (DAP), which allows readers to check the validity of an author's claims on the basis of the data sets that were used. Prior to publication, the data, programs and other details of the computations sufficient to permit replication are submitted to the journal in order to enhance external peer review. Usually, such materials are put on the Internet as a "digital appendix" to the paper journal. However, journals are not data archives, and it is not known what journals will do with respect to metadata and documentation, data formats, long-term storage, access to non-subscribers to the journal, etc.

van Zanden and de Moor describe the complementarity between collaboratories, DAP journals and data archives. They see three essential problems related to data availability for research: (1) convincing collectors of data to share what they see as "their" data with others, (2) guaranteeing the quality of the data and (3) preserving and disseminating the data. DAP journals have an advantage by making it obligatory for a researcher to submit

the underlying data if he/she wants to publish an article. Collaboratories offer good environments for quality control by peer review, and data archives have long-term storage and access as their main business (van Zanden and de Moor 2006).

## E-science and the grid

The term e-science is predominantly used in the UK, where the Department of Trade and Industry's official definition is widely supported: "science increasingly done through distributed global collaborations enabled by the Internet, using very large data collections, tera-scale computing resources and high performance visualization." Of course, e-science originated in the hard sciences, but "e-social science" and "e-humanities" have also emerged. In the UK, the National Centre for e-Social Science (NCeSS) is funded by the Economic and Social Research Council (ESRC) to investigate how innovative and powerful computer-based infrastructure and tools developed over the past 5 years under the UK e-Science programme can benefit the social science research community. E-social science refers to the use of grid infrastructure and tools within the social sciences. In the Netherlands, the Virtual Knowledge Studio was set up by the Royal Netherlands Academy of Arts and Sciences (KNAW) as an e-science programme for the humanities and social sciences (VKS).

The Grid is the architecture proposed to make a reality of such visions for e-science. Foster and Kesselman define the Grid as an enabler for virtual organizations: "An infrastructure that enables flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources" (Foster and Kesselman 1999). Resource in this context includes computational systems, data storage and specialized experimental facilities (see: http://www.nesc.ac.uk/nesc/define.html). The computational grid is the next generation computing infrastructure to support the growing need for computational based science. This involves the utilization of widely distributed computing resources, storage facilities and networks owned by different organizations but used by individuals who are not necessarily a member of those organizations. A descriptive way to explain computational grids is by analogy with the electric power grid. The latter provides us with instant access to power, which we use in many different ways without any thought as to the source of that power. A computational grid is expected to function in a similar manner. The end user will have no knowledge of what resource they used to process their data and, in some cases, will not know where the actual data came from. Their only interest will be in the results they can obtain by using the resource. Today, computational grids are being created to provide accessible, dependable, consistent and affordable access to high performance computers and to databases and people across the world. It is anticipated that these new grids will become as influential and pervasive as their electrical counterpart.

## Trusted digital repositories

As data are increasingly stored in many different places, it is particularly important to ensure that they meet minimum standards of quality, traceability, accessibility and usability. Various organizations in a number of countries have projects under way that seek to specify criteria for "trusted digital repositories". For instance, the RLG (Research Libraries Group) and NARA (National Archives and Records Administration) in the USA aim to formulate requirements that will ultimately lead to criteria and procedures for the

certification of digital repositories (Audit Checklist 2005). One of the difficulties is that digital archiving is not rocket science. Most institutions have only a limited experience, and because of the rapid technical developments, the guidelines and best practices of today may soon become outdated. The digital preservation initiative NESTOR (Network of Expertise in Long-term Storage of Digital Resources) in Germany, and DANS and the Royal Library in the Netherlands are also working on guidelines for trusted digital repositories. To ensure the broadest possible acceptance of its seal of approval, DANS wants to minimize the extent of mandatory requirements and to maximize the use of recommended guidelines, preferred standards and good practices. This data seal of approval is intended to eliminate or lower barriers to data access for research purposes. It is also intended to bring about greater efficiency in data management by minimizing the amount of manual work involved in the documentation, management and making available of data, for both researchers/data producers and data managers (Ross and McHugh 2006).

## Web archiving

A new field for all organizations involved is that of web archiving. There are several challenges in web archiving, in particular when archiving scientific/academic collaboratories. So far, most experiments in this field have been carried out by national libraries; only a few have been performed by national archives. One of the ways to capture websites is to take snapshots of a specific part or even the whole national Internet from time to time, for example every month or twice a year. This is known as the bulk method. Another way is to download websites more selectively. The latter method provides better, tailor-made information in which the frequency of capturing reflects the rhythm of the changes in the website (every week, every month), although it is more time-consuming. The snapshot method was applied by "the" first Internet archive, which was set up by Brewster Kahle in 1996; the websites that were captured are now made available by the Wayback Machine. There seems to be a trend for depository libraries to switch from snapshots of the whole domain (country) to selective archiving (Dollar 2001; Masanès 2002, 2005; Brown 2006; Internet Archive).

There are very few examples in the academic world. In the Netherlands, the Archipol websites archive is an interesting exception, where as part of the national documentation centre for Dutch political parties (part of the University of Groningen), websites of the national Dutch parties are being archived. This is a selective method in which the websites are downloaded once a month, or more frequently during elections, periods of political tension, etc. Once a year all the websites are integrally downloaded, as a snapshot. The websites are always downloaded by the archive itself. In Germany, a website providing information of a political nature (i.e. on the 2005 general elections in Germany) is now being kept by the German data archive in Cologne. This is not exactly the same design as the Archipol archive. It is remarkable, however, because one of the very first websites to be preserved in the USA was that of the 1996 presidential elections, kept in collaboration between the Wayback Machine (part of the Internet archive) and the Smithsonian Institution. Another interesting project in the academic sphere is the Occasio project at the International Institute for International History in Amsterdam. This collects web documents mostly from informal newsgroups, e.g. those related to the civil war in the former Yugoslavia (Archipol; General Elections Germany 2005 Website; Occasio Project).

For the academic world one of the most important questions in this regard is how and where collaboratories should be archived. Should they be archived by data archives or by

libraries (depository or university)? The question is whether these collaboratories should be treated as electronic publications—which they are—or as datasets, which they also are. Some question whether the scientists and scholars themselves will take care of this, or whether it will work only with collaboratories on the scale of the Human Genome Project.

**Acknowledgements** The authors would like to thank professor Seamus Ross (Glasgow, UK) for his valuable comments, in particular his literature suggestions

## References

### Unless otherwise indicated, all websites were consulted in December 2006

Ambacher BI (ed) (2003) Thirty years of electronic records. The Scarecrow Press, Lanham (MD) and Oxford
Archaeology Data Service: http://ads.ahds.ac.uk/project/general.html#consortium
Archipol. http://www.archipol.nl/english/index.html
Arts and Humanities Data Service. http://www.ahds.ac.uk/index.htm
Ashley K (2000) Digital archive costs: facts and fallacies. In: Proceedings of the DLM Forum on electronic records. European citizens and electronic information: the memory of the Information Society. Brussels, 18–19 October 1999. European Commission, Luxembourg, pp 121–126
An Audit Checklist for the Certification of Trusted Digital Repositories (2005) Draft (August 2005), RLG and NARA, Mountain View, CA: http://www.rlg.org/en/page.php?Page_ID=20769
Authenticity in a Digital Environment (2000) Report of the Council on Library and Information Resources, Washington: http://www.clir.org/pubs/abstract/pub92abst.html
Beagrie N, Greenstein D (2001) A strategic policy framework for creating and preserving digital collections. Version 5.0 (last updated July 2001), London: http://www.ahds.ac.uk/strategic.pdf
Bearman D (1999) Reality and chimeras in the preservation of electronic records. D-Lib Magazine 5 (4, April): doi:10.1045/april99-bearman
Beedham H et al (2005) Assessment of UKDA and TNA compliance with OAIS and METS standards. UK Data Archive: http://www.data-archive.ac.uk/news/publications/oaismets.pdf
British National Cataloguing Unit for the Archives of Contemporary Scientists (NCUACS): http://www.bath.ac.uk/ncuacs/home.htm
Brown A (2006) Archiving websites: a practical guide for information management professionals. Facet Publishing, London
Budapest Open Access Initiative (2002): http://www.soros.org/openaccess/read.shtml
Collaboratories example Humanities: http://www.lifecoursesincontext.nl/
Collaboratories in science: http://www.scienceofcollaboratories.org/Resources/colisting.php
Cook T (1999) What is past is prologue: a history of archival ideas since 1898, and the future paradigm shift. In: Horsman PJ, Ketelaar FCJ, Thomassen THPM (eds) Naar een nieuw paradigma in de archivistiek, Stichting archiefpublikaties, 's-Gravenhage, pp 58–61: http://journals.sfu.ca/archivar/index.php/archivaria/article/view/12175
Creative Commons: http://www.creativecommons.org
Crow R (2002) The case for institutional repositories: A SPARC position paper. The Scholarly Publishing & Academic Resources Coalition, Washington DC: http://www.arl.org/sparc/bm%7Edoc/ir_final_release_102.pdf
DANS EASY: http://easy.dans.knaw.nl/dms
DARE (Digital Academic Repositories): http://www.darenet.nl/en/page/language.view/search.page
DDI: http://www.icpsr.umich.edu/DDI/
Digital Curation Centre: http://www.dcc.ac.uk/FAQs/data-reuser#q3
Digital Preservation Coalition: http://www.dpconline.org
Digital preservation at the Royal Library of the Netherlands: http://www.kb.nl/hrd/dd/dd-en.html
Dollar CM (1999) Authentic electronic records: strategies for long-term access. Cohasset Associates, Chicago

Dollar CM (2001) Archival preservation of Smithsonian web resources: strategies, principles, and best practices. Smithsonian Institution Archives, Washington DC: http://www.si.edu/archives/archives/dollar%20report.html
Dollar CM (2003) An insider/outsider perspective on the electronic records program of the National Archives of the United States. In: Ambacher BI (ed) Thirty years of electronic records. The Scarecrow Press, Lanham (MD) and Oxford, pp 139–147
Dollar CM (2004) Trends in the archival acquisition and preservation of electronic records: 1970–2000. In: Doorn P, Garskova I, Tjalsma H (eds) Archives in cyberspace. electronic records in east and west. Moscow University Press, Moscow, pp 11–36
Doorn P (1996) Archives. In: Mullings C et al (eds) New technologies for the humanities. Bowker, London, pp 357–379
Doorn P (2004) Research data archives and public electronic record offices: What can we learn from each other? In: Doorn P, Garskova I, Tjalsma H (eds) Archives in cyberspace. Electronic records in east and west. Moscow University Press, Moscow, pp 96–111
Duranti L. (2005) Long-term preservation of accurate and authentic digital data: the InterPARES Project. Data Sci J 4:106–118: http://journals.eecs.qub.ac.uk/codata/journal/contents/4_05/4_05pdfs/DS426.pdf
Economic and Social Data Service: http://www.esds.ac.uk/
EDNA Project: http://www.edna.leidenuniv.nl/
Electronic Depot of Netherlands Archaeology: http://edna.itor.org/en/
Foster I, Kesselman C (1999) The Grid: blueprint for a new computing infrastructure. Morgan Kaufmann, San Francisco
General Elections Germany 2005 Website: http://www.gesis.org/information/sowinet/sowiplus/buwa2005/
Heery R, Anderson S (2005) Digital repositories review. Joint Information Systems Committee JISC Review, London: http://www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf
Human Genome: http://www.ornl.gov/sci/techresources/Human_Genome/genetics.shtml
Internet Archive: http://www.archive.org/web/web.php
Introduction Danish Data Archives: http://www.dda.dk/ddagb/introdoc/default.html
ISAD-G: http://www.icacds.org.uk/eng/history.htm
Jones M, Beagrie N (2001) Preservation management of digital materials. A handbook. The British Library, London: http://www.dpconline.org/graphics/handbook/index.html
Large-scale Data sharing in the life sciences: Data standards, incentives, barriers and funding models (The 'Joint Data Standards Study') (2005). Prepared by the Digital Archiving Consultancy Limited (DAC), the Bioinformatics Research Centre at the University of Glasgow (BRC), and the National e-Science Centre at Glasgow (NeSC): http://www.mrc.ac.uk/consumption/idcplg?IdcService=GET_FILE&dID=5027&dDocName=MRC002552
Lorie R (2002) The UVC, a method for preserving digital documents: proof of concept. IBM/KB Long-Term Preservation Study Report series 4, IBM Netherlands, Amsterdam: http://www.kb.nl/hrd/dd/dd_onderzoek/reports/4–uvc.pdf
McKemmish S (1999) Yesterday, today and tomorrow: a continuum of responsibility. In: Horsman PJ, Ketelaar FCJ, Thomassen THPM (eds) Naar een nieuw paradigma in de archivistiek, Stichting Archiefpublikaties, 's-Gravenhage, pp 195–210: http://www.sims.monash.edu.au/research/rcrg/publications/recordscontinuum/smckp2.html
Masanès J (2002) International initiatives in archiving the web. In: Den Hollander F, Voerman G (eds) Het web gevangen. Het archiveren van de websites van de Nederlandse politieke partijen. Universiteitsbibliotheek Groningen, Documentatiecentrum Nederlandse Politieke Partijen, Groningen, pp 19–24: http://www.archipol.nl/symposiumbundel.pdf
Masanès J (2005) Web archiving methods and approaches: a comparative study. Library Trends 54:72–90
National Collaboratories. Applying Information Technology for Scientific Research (1993). Washington DC: http://www.nap.edu/catalog/2109.html
National Strategy for Data Resources for the Social Sciences (UK). http://www2.warwick.ac.uk/fac/soc/nds/
NEDLIB project. http://nedlib.kb.nl/
NSF Cyberinfrastructure (2006) NSF Cyberinfrastructure Council, NSF"s cyberinfrastructure vision for 21st century discovery. CI Draft (Version 7.1; 20 July 2006):http://www.nsf.gov/dir/index.jsp?org=OCI
OAI-MPH: http://www.openarchives.org/
Occasio Project: http://www.iisg.nl/occasio
Oltmans E, Van Wijngaarden H (2004) Digital preservation in practice: The e-depot at the Koninklijke Bibliotheek. VINE - J Inf Knowl Manage Syst 34:21–26
Our Cultural Commonwealth: The final report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities & Social Sciences (2006). Report of The American Council of Learned Societies (ACLS): http://www.acls.org/cyberinfrastructure/

Oxford Text Archive: http://www.ota.ahds.ac.uk/
Ross S (2000) Changing trains at Wigan: digital preservation and the future of scholarship. National Preservation Office, The British Library, London: http://www.bl.uk/services/npo/pdf/wigan.pdf
Ross S, Gow A (1999) Digital archaeology: rescuing neglected and damaged data resources. British Library Research and Innovation Report 108, The British Library, London: http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/p2.pdf
Ross S, McHugh A (2006) The role of evidence in establishing trust in repositories. D-Lib Magazine 12 (7/8, July/August): http://www.dlib.org/dlib/july06/ross/07ross.html
Rothenberg J (1999) Avoiding technological quicksand: finding a viable technical foundation for digital preservation. A report to the Council on Library and Information Resources. Council on Library and Information Resources, Washington DC: http://www.clir.org/pubs/abstract/pub77.html
Rusbridge C et al (2005) The digital curation centre: a vision for digital curation. In: Proceedings from local to global: data interoperability – challenges and technologies, Forte Village Resort, Sardinia, Italy. pp 1–11: http://eprints.erpanet.org/archive/00000082/01/DCC_Vision.pdf
Scheuch EK (2003) History and visions in the development of data services for the social sciences. Int Social Sci J 55:385–400
Schürer K (2001) Better access to electronic information for the citizen. The relationship between public administration and archives services concerning electronic documents and records management. European Commission, Luxembourg
Shepherd E, Smith C (2000) The application of ISAD(G) to the description of archival datasets. J Soc Archivists 21:55–86
Shepherd E, Yeo G (2003) Managing records. A handbook of principles and practice. Facet, London
The Netherlands Code of Conduct for using personal data in academic research (2005) (available only in Dutch): http://www.knaw.nl/nieuws/pers_pdf/Gedragscode.pdf
Tjalsma H (2004) Long-term preservation of electronical historical data in the Netherlands. In: Doorn P, Garskova I, Tjalsma H (eds) Archives in cyberspace. Electronic records in east and west. Moscow University Press, Moscow, pp 112–127
Upward F (2005) Records continuum. In: McKemmish S, Piggott M, Reed B, Upward F (eds) Archives. Recordkeeping in society, Charles Sturt University, Wagga Wagga, pp 197–222
Van Ballegooie M, Duff W (2006) Archival metadata. In: Ross S, Day M (eds) DCC Digital curation manual. Digital Curation Centre, Glasgow: http://www.dcc.ac.uk/resource/curation-manual/chapters/archival-metadata/
van Horik R (2005) Permanent pixels. Building blocks for the longevity of digital surrogates of historical photographs. DANS, The Hague: http://www.knaw.nl/publicaties/pdf/20051103.pdf
van Zanden JL, de Moor T (2006) Do ut des: collaboratories as a "new" method for scholarly communication and cooperation for global and world history. Draft article/ in process of publication (August). http://www.iisg.nl/publications/do-ut-des.pdf
Wise A (1997) UK consortium opens archaeology data service. CSA Newslett 9(4): http://www.csanet.org/newsletter/feb97/nl029702.html
Wulf W (1989) The national collaboratory. In: Towards a national collaboratory. Unpublished report of a National Science Foundation invitational workshop, Rockefeller University, New York
Wulf W (1993) The collaboratory opportunity. Science 261 (13 August)