The ADA approach: retro-archiving data in an academic environment

Marjan Balkestein · Heiko Tjalsma

Published online: 25 August 2007
© Springer Science+Business Media B.V. 2007

Abstract This article concentrates on the retro-archiving of older digital research data. The *ADA approach* was developed and used to retro-archive older data files, most of which were between 10 and 30 years old. The origin and main characteristics of the ADA approach are described in the second section of the article. The third section discusses two recent data-archiving pilot projects that were conducted in the Netherlands. The first of these projects, the ADA project, laid the foundation for the ADA approach, which was subsequently applied and tested again in the second project, *eDNA*, which focused on archaeological data. The final section of the article provides a comparison of the results of these two projects.

Keywords Data archives · Digital preservation · Institutional repositories · Archaeological archives · Social science data archives · Humanities and social science cyberinfrastructure

Introduction

This article outlines the *ADA approach* for retro-archiving older digital research data. Developed by the Netherlands Historical Data Archive (NHDA; now part of DANS, see the acknowledgments), the ADA approach is a generic method that is based on practical experiences with data archiving, and it is intended for use in an academic research environment.

The article begins with a sketch of the genesis, background and main characteristics of the ADA approach, within the context of the increasing concern for the long-term

M. Balkestein (✉) · H. Tjalsma
Data Archiving and Networked Services (DANS), P.O. Box 93067, The Hague 2509 AB,
The Netherlands
e-mail: Marjan.Balkestein@Dans.Knaw.Nl

H. Tjalsma
e-mail: Heiko.Tjalsma@Dans.Knaw.Nl

preservation of digital data. The second part of the article provides a detailed discussion of two data-archiving projects that were conducted in the Netherlands. The first of these projects, the original *ADA project*, laid the groundwork for the ADA approach. The approach was applied and re-tested in the second study, the *eDNA project*. The article concludes with a comparison of the results of these two projects and the consideration of the possible future of the ADA approach.¹

The ADA approach

Background and origin of the ADA approach

Our digital heritage is in danger. As explained in the introduction to this special issue, this awareness has reached the academic community as well, but not everywhere. Research data archives, data libraries and university libraries are certainly aware of the issue, as are some research-funding organisations.

These concerns have sparked several recent initiatives in the Netherlands. One of these initiatives involved a study that was conducted jointly by the University Library of Leiden and the NHDA, which investigated the policy implications of the long-term preservation of the digital collections of universities and research institutes. One of the main conclusions of this study, which was entitled *Digital academic heritage*, was that the number of digital resources (e.g. research data files, academic electronic publications) is increasing rapidly, but universities and other research institutes lack any clear preservation policy. From a number of interviews with key figures (e.g. scholars, policy makers, executive managers of universities and national research organisations, archivists and science historians), the authors concluded that this problem is perceived as urgent and acute. Another important conclusion was that the appraisal criteria for preservation (i.e. decisions concerning which data should be preserved for posterity) can be established only by researchers within the discipline, with the exception of criteria that are of a more general historical nature. From the perspective of the history of science or cultural history, (science) historians and possibly other experts should be involved as well. Information from the interviews also identified central national depositories as the preferred means of guaranteeing enduring care for research data and electronic publications (Mostert et al. 1998).

As a follow-up to the earlier *Digital academic heritage* study, the ADA project was launched in 2000. The project lasted for more than 3 years; it was conducted on a far larger scale than the earlier study had been and, more importantly, it had a completely different character. The main activity within the ADA project was an experiment with the long-term preservation of digital data *in practice*. This was a typical pilot project. The ADA approach was formulated largely based on the practical lessons that were learned in this study. The following section describes the characteristic features of the ADA model.

¹ ADA is a Dutch acronym that stands for *Archiving Digital Academic Heritage*. Any correspondence between this acronym and Lady Augusta Ada Byron (1815–1852), Countess of Lovelace and one of the first ‘computer women’, is purely coincidental. *eDNA* is the acronym for *e-Depot Nederlandse Archeologie* (Repository for Dutch Archaeology). It does not in any way refer to Dame Edna (1955–), also known as Barry Humphries.

Main characteristics of the ADA approach

The ADA approach is designed for retro-archiving (i.e. retrospective archiving). The preservation of digital records for posterity demands that preservation planning activities be initiated as early as possible, preferably at the time that digital resources are created. The concept of ‘pro-active archiving’, however, has only recently begun to reach the research world.² From a strictly legal perspective, Netherlands law requires that all documents that are produced by state-funded universities and research be treated in the same way that administrative records are treated. While public record offices are charged with the task of pro-actively preserving the latter category of documents, such tasks are usually not performed for research data, largely because public record offices generally do not consider scientific data files selectable for preservation (Rijksarchiefinspectie 2003).

Issues of long-term preservation receive insufficient attention, even in disciplines where considerable digitisation of analogue originals (i.e. those that were not born digitally) is already taking place to increase the possibility of studying those sources digitally, as in literary studies (texts) or art history (images). This is also the case with information that is published on the Web.

Practices in the academic world cannot be expected to change soon. Significant increases in awareness of the need for preservation are no guarantee that researchers will become actively engaged in preservation activities during their research. The pressure to publish allows little time for sorting and documenting data or for ensuring their long-term preservation (e.g. by converting them into software-independent formats). Carrying out such data conversion as a standard activity at the end of research projects would already represent progress. In the current ‘rescue-after-creation’ situation (Ross 2000, p. 13), the preservation of data files that were created in the past should probably receive first priority. The ADA approach is therefore not concerned primarily with activities, policies or guidelines for the future preservation of digitally created data that are now coming into existence. It is limited to safeguarding files that were created a number of years ago, and those that may be contained in antiquated media or software. Such tasks sometimes require painstaking documentary reconstruction.

The ADA approach is an alternative working method for data archives. It can be described as a form of digital preservation *services*. This organisational model differs from the usual practice of data archives, in which individual researchers or research groups *deposit* data files in all kinds of formats. All of the labour-intensive data-archiving tasks are performed in the data archive. The concept behind the new approach is that ADA projects can be carried out within research institutes or faculty departments. This is not a rule however, as part or even all of the activities could be performed at the data archive as well. Due to the close involvement of the institute where the data were created, however, it is usually preferable to work at the research institute.

Based on our experiences in the first ADA project, we formulated the ADA model, which consists of seven distinct phases (see Fig. 1). This is the full model. Depending on the context of a given project, it may not be necessary to perform all stages completely. Particularly in the later stages of the ADA approach, it is not necessary to perform all of the activities in the same order or at the same pace. Projects can be subdivided into parts that vary in priority or speed, similar to the System Development Model (SDM). Financial considerations and institutional policies are among the possible reasons for varying the order and priority of the steps in the ADA model. As discussed further in the description of

² The editorial of this issue provides additional information on this point.

Fig. 1 MODEL OF THE ADA APPROACH**1. Global description**

The general purpose of this stage is to develop an overview of the size and nature (e.g., format, media) of the task ahead. The information gathered in this stage is necessary for drafting a general project plan. This exploratory information includes a description of the infrastructure, particularly with regard to the applied information technology, now and in the past. A broad outline of the organisational context of the research unit where these data were created should also be included. The history of the technical infrastructure can suggest the types of file formats or media that can be expected. The global description should also include information about any major research projects that may have been executed in the past.

2. Definition of material

This stage involves conducting an inventory of the main systems (both hardware and software) that have been used. The definition should include information about the major format groups and media that have been in use. This step involves a broad initial selection of the (groups of) formats or media that are to be included or excluded.

3. Selection and documentation of projects

This step involves the selection of projects to be archived. The selection is based on an inventory that consists largely of project information. The inventory should indicate the systems (media, software) that were used in these projects, and it should provide a broad indication of the total size, number and level of homogeneity of the digital files. The goal of this step is to determine which projects are important enough to preserve, in the interests of research or the history of science.

4. Medium conversion and documentation at the level of data clusters

This stage involves the description of the archival files of the projects that have been selected for (possible) preservation, at the level of groups or *data clusters*. These are distinguished according to either file format and media (e.g., all jpeg files, all files in a certain directory in a selected area on a server or on a tape or floppy disc) or (partial) projects. Medium conversion must take place, particularly for antiquated media (e.g., mainframe tapes). Reading mainframe tapes and other outdated media may require the services of such specialised institutes as computer museums or computer firms.

5. Documentation at the file level and conversion to standard formats

Data files are selected for further processing according to the data cluster descriptions. The selected data clusters are described in more detail at the level of the data files that they contain. Specific metadata schemes for each discipline (e.g., the DDI or 'Data Documentation Initiative' for the social sciences; see the article by Rasmussen and Blank elsewhere in this journal) are used to describe the data files. Version checking is performed in this stage, and data are converted to standard formats. If data conversion and migration strategies are used in the archive, data files are converted to such software-independent formats as ASCII, XML and SPSS portable files.

6. Documentation at the data level

This step involves the deepest level of data description (on the variable level of data files or individual record fields), again using DDI or equivalents. Tasks in this step include describing the content and structure of the data. This stage also involves data cleaning and the collection of codebooks, additional information (e.g., project reports, project journals or logbooks) and related (internal) publications.

7. Accessibility and preservation

Several options are possible in this step. Data files can be made accessible by the data archive, the research institute itself, others or some combination of these parties. Precautions for long-term preservation should be taken, particularly with regard to the storage of data and metadata.

the *eDNA* project, the ADA model is flexible and can be adapted to various settings in many ways. Data archives can be engaged in the process to varying degrees and for a variety of services. Possibilities include providing in-house consultancy, giving courses on technical, metadata or management and policy issues and performing full data archiving (including documentation, storage and access services).

Feasibility and target group

Although the services of the ADA model are targeted towards the research world, the approach may also be suitable for other audiences. While the value of retrospective

archiving in project form using the ADA approach is obvious, its feasibility in practice is less clear. One pertinent question is whether research institutes would want to apply the approach. We addressed this issue in the first ADA project, as the funding organisation for the pilot project was obviously interested in its feasibility. The limited market research consisted primarily of a number of interviews with key figures in research institutes and university departments in the humanities and social sciences. These were either scientific directors or managers of research institutes or university employees who were responsible for the use of IT applications within an institute. Despite its limitations, this research provided a good overview of how IT has been managed in the last several decades, and how it is currently managed in this part of the research world. This market research took place in 2003.

The interviews revealed a high degree of interest in and concern for the long-term preservation of digital data within these institutes. In particular, there was a clear demand for expert advice. To date, however, this general awareness has not led to the formulation of clear institutional policies regarding the data preservation, and such policies are obviously not being applied. Some institutes do have preservation policies, however, and they are sometimes linked to access policies, usually through their websites. This is especially true of humanities institutes, in which almost all of the information that is collected is published in the form of source editions. Many institutes appear to have serious doubts concerning the quality and safety of their policies in this regard. Institutes that work with large amounts of (historical) texts are particularly likely to attach the highest priority to digitising older texts, which currently exist only on paper. For these institutes, it was clear that preservation issues are, at best, of only secondary importance. Preservation thus receives insufficient attention, and the risks of data loss are certainly underestimated.

Since digital preservation apparently does not always have the highest priority, the demand for the ADA approach in practical cases is unclear. Awareness of the need for digital preservation policies is somewhat abstract for most research institutes. Even though research institutes are not extremely active in this area, there does appear to be a clear demand for at least some ADA activities, under the condition that data archives are available to provide professional assistance. There is a great need for best practices within the field of information, particularly with regard to preservation strategy, storage and documentation standards. Research institutes are also interested in making inventories of projects or databases in order to obtain a better idea of the magnitude of the digital preservation problem. Some institutes also mentioned a desire for a national centre in which research files could be stored, maintained and made accessible. The market study did not produce enough information to provide a clear answer to the question of whether institutes or faculties would be willing to spend money on full data-archiving projects (including appraisal and storage).

A number of developments that have taken place in the Netherlands since 2003 are worthy of note. One promising initiative is the national project DARE (Digital Academic Repositories), which involves the creation of digital repositories containing digital scientific and scholarly publications, as well as (to a lesser degree) research data. The section about the eDNA project discusses these repositories in more detail. The recent creation of DANS (Data Archiving and Networked Services) as the national research-data archive by the two national research organisations, the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organisation for Scientific Research (NWO), is a second initiative.

The DANS organisation is structured as a network with a strong core. In collaboration with research groups, DANS establishes thematic development programmes in which research data in particular domains are archived and made accessible. This leaves a large

part of the responsibility with the researchers. In the case of archaeology, researchers within the discipline took the initiative upon themselves, leading to the *eDNA* project.

Cost models

An important point that was raised in the market research involved the costs of the ADA approach. As various experts have concluded earlier, this is a complicated issue. In their 2001 handbook, *The preservation management of digital material* in 2001, Maggie Jones and Neil Beagrie observe that ‘costs for both technical and organisational infrastructure are still not well defined’; this observation is still valid. According to Jones and Beagrie, one of the main problems is the impossibility of separating preservation costs from other costs that are made by the same institute (e.g. for data accessibility or for digitisation). The costs of *long-term* preservation are particularly difficult to establish, as Chapman noted in a more recent article (Jones and Beagrie 2001, pp. 27–28 and 79–82; Woodyard-Robinson 2006; Chapman 2004). The only cost models that are currently available are for storage. Cost is an important element in the discussion between those who favour migration strategies and those who favour emulation. Both sides maintain that their method will be less expensive in the long term (Oltmans and Kol 2005).

As the discussion above clearly shows, the ADA approach offers a number of advantages in this respect. Since it is custom-made, the ADA approach allows collaboration partners (the customers) to decide what they do or do not want. For this reason, the initial stages are designed to provide an overview of both the expected quantities and the risks. The model also allows flexibility with regard to who will actually perform most of the activities (e.g. personnel from the data archive or staff from the research institute). In addition, the institutional ADA approach is obviously more cost effective than archiving each dataset one-by-one in contact with individual researchers.

The ADA approach includes a model for reducing financial risks by clarifying what can be expected in subsequent steps and what the costs are likely to be. It is clear that any large-scale project will require money, regardless of who will carry out the actual work. Preservation can be seen as an investment for both the institute and the academic discipline.

Two pilot projects

In the following section we will take a closer look into the ADA project at the Meertens Institute and the *eDNA* project in the domain of archaeology.

The ADA project at the Meertens Institute

The Meertens Institute

The first ADA project involved archiving the digital research-data files of the Meertens Institute. The Meertens Institute is part of the KNAW. The institute’s primary field of study concerns local and regional diversity in language and culture in the Netherlands. Its website states that the institute focuses on ‘the structural, dialectological and sociolinguistic study of language variation within Dutch in the Netherlands’ (Meertens Institute).

Since its establishment, the institute has been strongly involved in collecting data and documents concerning the Dutch language and culture, which had long exceeded the usual tasks of conducting research and publishing scholarly articles. As a result, the institute possesses a substantial library, with numerous collections and a documentation system that consists largely of databases. The topics that are covered by the books and databases include dialects, first and family names, place names, folk culture, ancient customs, feasts, rituals, fairy tales and songs.

For some time, the institute had intended to formulate a digital-preservation policy. This intention, combined with the desire to have an inventory of all digital files within the institute (old or new), led the Meertens Institute to respond enthusiastically when it was invited to participate in the ADA project. Many of their digital holdings were kept unsystematically in diskette boxes in a ‘cupboard’, but nobody knew exactly what the cupboard contained.

Stages of the project

The ADA project was a pioneering project. As is often the case with such initiatives, reality differed considerably from our initial expectations. We had originally divided the project plans into three major stages: making global descriptions (inventories) of the data, appraising them and providing long-term preservation (data archiving). These stages were based on the existing routines of the data archive. We soon discovered that this would not work in practice. The magnitude of the project was a major unforeseen element. The inventory revealed an unexpectedly large number of data files in different forms and on varying media (see Table 1). In addition to the diskettes that were located both inside and outside of the cupboard, we discovered a number of other (primarily backup) media, including CD-ROMs, large mainframe tapes and other cartridges (e.g. SyQuest removable hard-disk cartridges).

As estimated at the end of the project, the various media together contained 18,500 data files. The sheer volume of data files was a major challenge, which necessitated several adaptations to the original project plan. The three stages of the project, which were initially intended as consecutive, had to be executed in a much more flexible—and *iterative*—manner than had originally been expected. The phases should not be conceived as completely separate project steps. An important point that soon became evident was that it is far more efficient to conduct an initial appraisal at the level of large data clusters before

Table 1 List of media found in the ADA project

Description of medium	Initial number	Not readable	Processed
diskettes 3½"	1497	36	1461
diskettes 5½"	68	14	54
<i>diskettes total</i>	<i>1565</i>	<i>50</i>	<i>1515</i>
magnetic tapes	4	1	3
CD-ROMs	23		23
SyQuest <i>backup</i> media	6		6
DAT medium	1	1	
Total	1599	52	1547

attempting to provide even the broadest description of the data files, which can be appraised later in the process. Other adaptations involved limiting the extent to which the data would eventually be described.

It was not always easy to create the data clusters (i.e. to distinguish and demarcate groups of data files from each other), particularly on the large backup media, which often contained copies of the hard disks from a number of personal computers. In most cases, the information that was available on research projects that had been conducted in the past was insufficient for making accountable decisions. It was therefore not possible to formulate and use consistent criteria for creating the data clusters. Naming systems sometimes had a role in the process, as did the creators and the provenance of files. Data were classified into a number of categories. The most important distinction was between the various types of manufactured *software* and the digital *data* that had been created by the researchers of the institute. Most of the items in the first category (which included numerous copies of well-known software programs or operating systems) were excluded from the data-archiving project. Raw research data constituted by far the most important sub-category of the latter category, in addition to a small number of other data material (e.g. annotated digital texts). The value of engaging the assistance of the original creators (or of those who had worked closely with the project and, in particular, the data) became quite clear in this phase.

Appraisal and selection

After appraisal, we were able to reduce the total number of data files from the estimated 18,500 to 900 files, which were ultimately archived. The original total of 18,500 was estimated but never exactly calculated, as many digital files had been deselected at an early stage. Since a large number of data clusters could be eliminated very easily and quickly, it was unnecessary to investigate a substantial percentage of the data files, as they were simply system, program or other files that contained no relevant research data or information. Only 11.3 % (in bytes) of all the files were data files.

The range of time that was considered was not that large (see Fig. 2), as could probably be expected from an institute in the humanities. Significant numbers of digital data files had not begun to emerge in this field until the 1980s. Prior to that time, the humanities had made very little use of ICT (see for history: Boonstra et al. 2004, pp. 25–28). The institute used primarily Mac computers.

After the earlier, later and duplicate versions of files had been filtered out, the role of the institute and the researchers proved indispensable during the actual appraisal process. The process was carried out according to a preservation policy that had been formulated in broad lines by the institute. This policy was especially important for data that had been created much earlier by researchers who were no longer at the institute and whose projects were finished. In such cases, the documentation department played a crucial role in final appraisal decisions or in mediating such decisions with researchers who were still present.

The participation of a specialised institute was needed for the media-conversion phase of the project. Without the facilities of the Computer Museum of the University of Amsterdam, we would not have been able to read the data that were contained on old media (e.g. magnetic tapes from the mainframe or Sy-Quest and other antiquated backup media) or to recover damaged floppy disks (Computer Museum of the University of Amsterdam). These tasks required outdated and therefore very rare hardware.

Fig. 2 Overview of the creation years of data files

Data archiving

Most of the project time was consumed by work in the first two (original) stages of inventory and appraisal. The last stage, in which the selected data files were actually archived, took considerably less time. After completion of the selection, there were approximately 900 data files, which had to be converted into software-independent formats. Most of these data files were either database files (largely FileMaker) or text files (largely WriteNow), and almost all had been created on a Mac platform. There were two main activities in this phase: (1) converting the data to ASCII files and (2) describing the data. All texts were eventually saved in Unicode, in order to make them readable for both Mac and Windows platforms.

It is obvious that we used a conversion/migration strategy. This was primarily for practical reasons and because of the fact that, as data-archive workers, we were accustomed to this method. Testing an emulation strategy was not part of the project, and we had neither the time nor the resources to do so. The entire project was already under pressure to be completed on time, given the unexpectedly large amount of data that were found.

In the project we were able to add technical metadata (e.g. information on the size, type, platform and the creation dates of the data files) to the data. Although we added a limited quantity of contextual metadata (file names and very summary descriptions of groups of files), it will be necessary for the Meertens Institute to complete this task. In the course of the project, it was decided that issues of data access exceeded the scope of the project.

Conclusions of the ADA project

The execution of this data-archiving project proved to be a valuable experience for us. As noted above, we encountered many unexpected challenges. We obviously discovered that we had made mistakes, both during and after the project, or that it would be necessary to return to earlier phases because something had been overlooked. Our experiences with this

project eventually led to the formulation of the ADA approach (see Fig. 1). To the best of our ability, we incorporated the lessons that we learned (including the mistakes) into the model.

The iterative character of the process was one of the most important lessons. We therefore concluded that it is far more efficient to make an initial selection at the research project level, making a subsequent selection at the level of large data clusters and making a final selection at the data file level. Only after this process of progressive selection is it possible to begin the detailed description and conversion of the data files. This procedure differs from our original intention to follow the consecutive, linear stages of making global descriptions (inventories) of all the data, appraising them and providing long-term preservation (data-archiving). We therefore extended the number of project stages in the model from three to seven. Even under a strict planning schedule, iterations will inevitably arise from time to time during the course of a project, as the necessity of returning to data clusters or projects that have already been deselected is always a possibility. This should be avoided as much as possible, however, as such iterations decrease efficiency.

Related to this point, we discovered that it is indeed necessary to make a good initial estimate of the size of the project (i.e. quantity of data, variation of data, variation of media) before actually beginning. This step requires an overview of the institute's IT history. Close communication with the institute and its involvement in the process are also vitally important. The participation of the data producers is even more essential, as the institute that owns the data is usually the only source of information on the data. Without such information, data reconstruction is often impossible. Although the degree of detail in accompanying metadata obviously varies, a certain amount of such information is necessary. This requires at least a minimal level of cooperation and communication with the institute.

As indicated earlier, the ADA approach also offers considerable flexibility in what an institute can do in various types of data-archiving projects. Our experiences in this project taught us that typical activities in the final stages of a project (e.g. data retrieval and data access) are almost completely in the hands of the institute that owns the data.

We concluded that neither the necessity of reconstructing and reading data nor technical issues of preservation (archiving) pose any great impediments to retro-archiving activities. The tasks of finding the right versions, evaluating, appraising and subsequently documenting the data pose a much greater problem and can form a bottleneck in this type of project.

The *eDNA* project: Repository for Dutch Archaeology

Background of eDNA

From January until July 2004, the Faculty of Archaeology at Leiden University conducted a joint pilot project with the Netherlands Historical Data Archive (*eDNA* Project). The aim of the project was to begin archiving archaeological datasets that had been created by the staff of this faculty, in order to save valuable archaeological data for future research. Although the results of archaeological excavations and research are increasingly becoming available only in digital form, there were no facilities for the preservation of Dutch archaeological data. Once an archaeological project is finished, data files often disappear into a drawer somewhere, thereby running the risk of becoming obsolete. The need and

urgency to start preserving such digital information was therefore felt strongly among the faculty and within the archaeological field in general.

The Leiden project was intended to make an inventory of datasets that had been created by researchers of the faculty and to start preserving them for the long term. The project was thus a retro-archiving project: digital datasets that had been created in the past were to be archived retrospectively. This proved to be quite a difficult task. For the most part, everything had to be developed from scratch. There was no suitable metadata scheme for archaeological data files, no tools for documenting these files, no structural storage solution and no data archivists with experience with archaeological data (particularly for data created with GIS and CAD software). In addition, the number of datasets and data files was also quite large. It was therefore already obvious at the time of the pilot that much more work would be needed than would be possible within the limited time and possibilities of the project.

Since the need to initiate some sort of action regarding the preservation of archaeological data was also felt strongly outside the University of Leiden, a joint proposal for a larger retro-archiving project was formulated. The proposal was submitted to SURF, which was developing an academic infrastructure for research results within the framework of the Digital Academic Repositories (DARE) programme (Darenet). DARE is a joint initiative of the Dutch universities, the Royal Library (KB, the National Library of the Netherlands), KNAW and the NWO. The initiative aims to make publications and research data digitally accessible to the academic community by developing repositories for academic research (DARE Repositories). The *eDNA* project (*Repository for Dutch Archaeology*) began in August 2004 and continued until February 2006.

The project was a joint initiative of the archaeological faculties and departments of all Dutch universities, the Netherlands State Service for Archaeological Investigations (*Rijksdienst voor het Oudheidkundig Bodemonderzoek*, or ROB) and the NHDA.³

The main aims of the project were as follows:

1. to create awareness in the archaeological field for the problem that data and digital media will become obsolete and inaccessible, thereby becoming unavailable for (future) research, if no proper measures are taken for preservation;
2. to demonstrate the value of an archival service for archaeological data by preserving a number of datasets from each of the participating institutes and making them permanently accessible through one or more academic repositories;
3. to explore the organisational and financial possibility of developing a structural service in the Netherlands for the long-term preservation of archaeological data.

DARE repositories and Dublin Core

The *eDNA* project was guided by both the ADA approach to retro-archiving, as discussed above, and by the DARE requirements. DARE focuses strongly on the accessibility of academic output, primarily the publications of researchers. It therefore stimulates the construction of institutional repositories, in which researchers can store and provide access to ('publish') their research. It was decided to build a central repository especially for archaeological data (*eDNA* Repository for archaeological data), while storing some

³ After completion of the project, DANS and the ROB decided to continue *eDNA* jointly as a structural service for archaeologists.

datasets in the existing institutional repositories as well. For the central *eDNA* repository, we used i-Tor repository technology, which was developed by the NIWI, and which has been used for a number of other DARE institutional repositories. All DARE repositories follow the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), which enables the accessibility of digital collections in a low-maintenance manner, and allows the interchange of metadata between different archives or institutes (OAI-MPH).

An additional requirement of DARE is that academic output that is stored in DARE repositories should be accompanied by a Dublin Core description. Dublin Core (DC) is a metadata standard for digital sources on the Web (Dublin Core Metadata Initiative). It consists of fifteen elements that can be used to provide a general description of a digital source (in the case of *eDNA*, a dataset from an archaeological project). Each dataset in the *eDNA* repository is accompanied by a Dublin Core metadata description in XML. The DC description provides a general overview of the dataset, including a summary of the archaeological project in which the dataset was produced. One DC description covers an entire dataset and encompasses many separate files.

Each DC description can be harvested by other service providers. By harvesting the Dublin Core descriptions from all other DARE repositories, DARE itself acts as a service provider that can be accessed through the DAREnet search facility. This facility can be seen as the central catalogue of information from all DARE repositories (Darenet search service).

eDNA and the ADA approach

The main objective of the *eDNA* project was to increase awareness regarding the problem of digital preservation of archaeological data by starting to safeguard data files on a small scale and by allowing local archaeologists to perform data-archiving activities themselves. The ADA model provided valuable guidelines for the workflow procedures during the project. The flexibility of the ADA approach, in which the results of one step are used to guide the next step, proved particularly useful.

The same data-archiving activities were carried out at each of the six participating institutes. One local archaeologist was appointed to carry out a limited retro-archiving project, with help from other staff members of the institute and from the central project organisation. The idea behind this was that this could stimulate the dissemination of existing and newly acquired knowledge about data archiving among and between the participating institutions.

It was agreed that the following activities would be carried out by each of the participants. These steps largely followed the ADA approach, although some of the original aspects were divided or reordered. A short summary of the *eDNA* steps is provided below, accompanied by some remarks about our experiences. This discussion is followed by conclusions about some of the major differences between the ADA steps and the approach that was followed in the *eDNA* project.

eDNA retro-archiving steps

Global description of the IT history

This step proved particularly useful during the pilot project in Leiden, as it provided a good indication of which archaeological projects had been executed in the past and what hardware and software had been used. At some of the other faculties, however, it was

rather difficult to recover the IT history, as the staff members who had the necessary knowledge had already left the institute. Consequently, this information was not always acquired at the beginning of the project. In general, however, global description is an important initial step for a retro-archiving project, as it can provide an impression of the scale of the job ahead and a background framework for subsequent steps.

Inventory and documentation of projects

Different from the ADA approach we concentrated on excavations, research projects and the people involved, and did not pay much attention to technical aspects. Interviews were held with staff members at each of the participating institutes in order to acquire detailed information on (the availability of) datasets from excavations and surveys. Uniform Dublin Core metadata descriptions of the different projects were subsequently created, and they were eventually made available in the *eDNA* repository.

Selection of datasets

Information from the interviews and Dublin Core descriptions allowed us to make a good selection of datasets that could be archived during the *eDNA* project. Since the time in which the project(s) had to be carried out was limited (about 4 months per institute), and because the number of datasets and files was often quite large, we knew beforehand that each participant would be limited to archiving only a small selection of datasets. This selection of datasets can be regarded as a representative sample and a showcase for the entire project.

Medium conversion

All of the files from each of the selected datasets were collected and copied into a single directory (medium conversion), in order to allow efficient file management. The files had originally been stored on a variety of media, including old magnetic tapes, 5½" floppy disks (double and high density), 3½" disks, ZIP disks and CD-ROMs. Surprisingly, we generally had no problems reading the media; the few exceptions included a mainframe tape and one or two old floppy disks.

Appraisal and selection of files

Together with the researchers, a selection of the files to be archived was made for each dataset. The selected files contained primary source material and additional information (e.g. reports and codebooks) that are necessary for a good understanding of the archaeological project. With the help of tools that generate lists of files, grouped according to format, date and size, the researchers selected files that they considered important for archiving. The files were also subjected to a global check for integrity, completeness and quality.

The selected files were then physically grouped into clusters, according to file format (e.g. databases, AutoCAD maps, images, text documents) or content (e.g. field registration files, maps, photographs, reports).

Documentation at the file level

Since no metadata standards were available for archaeological files, we had already developed our own metadata scheme during the pilot project in Leiden, and we tested that scheme in this step. The system, which is known provisionally as ADDI (Archaeological Dataset Documentation Initiative), consists of a set of descriptive elements based on three metadata standards: the DDI (Data Documentation Initiative) metadata scheme for Social Science datasets, the FGDC (the Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata) and Dublin Core (Data Documentation Initiative; Federal Geographic Data Committee; Dublin Core Metadata Initiative). Metadata were stored in a simple application from which XML metadata files were generated. The XML descriptions were eventually stored in the *eDNA* repository, together with the data files. With the documentation tool, it is also possible to describe groups of files, if they share common descriptive elements.

Conversion of files in archival formats

For the preservation of the data files, we followed the conversion/migration strategy used by the NHDA. When possible, databases, spreadsheets, texts and geospatial files were converted into ASCII. Images were stored in TIFF or JPEG format, and publications were stored in both ASCII and PDF (portable file format). Although it was also decided to preserve all the files in their original format, they were not made available to end-users in the *eDNA* repository.

Documentation at the data level

The assistance of the researchers was vital for the creation of codebooks (description of variables and data-labels). The meaning of variables and codes in data files was not always clear, and it was often necessary to approach the creator of a file to retrieve their exact meaning. The codebooks were published as separate files, in ASCII format, and stored in the repository with the data files and the other metadata.

Publication of datasets and metadata in the eDNA repository

As previously indicated, it was decided to build a new repository for storing all of the archived files and metadata. For each dataset, a single XML metadata file describing both the project and the separate files is stored in the repository, together with the data files and the codebooks. File documentation, which can be viewed and downloaded, is linked to the data files and codebooks.

In principle, all data files can be freely downloaded from the repository. A number of files, however, are still under embargo, and can be consulted only with permission from the depositor.

Backup storage at NHDA

Since repositories are neither intended nor equipped for long-term preservation, it was decided to store a backup copy of each archived dataset and metadata at the NHDA (now DANS). The copies of the original files are also stored here.

Characteristics of the *eDNA* project

As is the case with any retro-archiving project, the *eDNA* project has a number of unique characteristics. First, considerable attention was paid to making an *inventory* of all archaeological projects that had resulted in digital data, with the help of all the archaeologists at the various institutions. In compliance with a requirement from DARE, a *Dublin Core description* was made for each project. These descriptions could be supplemented later with file descriptions, should it be decided to archive a particular project. The explicit *project selection* that we applied is another unique characteristic of the *eDNA* project. From the beginning, it had been agreed to archive only a limited number of datasets at each participating institute, which could serve as an example (a showcase) for demonstrating the importance of a digital archive for archaeology.

The important phase that followed the selection of datasets required the management of large quantities of data files. In this phase, considerable attention was paid to the *appraisal and selection* of the files. The number of files that were initially acquired from the researchers was sometimes enormous. In one case, we received more than 20,000 different files for one dataset. For the documentation of the data files, we used our own archaeological metadata description scheme, ADDI.

The last characteristic of the project that we would like to mention is the storage of the data files and metadata in a *repository* that was especially developed for archaeological datasets. For each project and its data files, one *XML metadata description* was generated and stored in the repository, together with the data files and the codebooks. The repository provides the Dutch archaeologists with a central facility for storing and accessing archived datasets.

Final considerations

As indicated above, the two projects that are described here took place in different disciplines and in different institutional contexts. The type of research, the method of data collection and, consequently, the data themselves all had unique features. Comparing the experiences in the two projects shows both the similarities and the differences in approach that can occur when applying the ADA model.

One of the main differences between the projects was that the *eDNA* project did not involve any data-archiving services performed at another institute, as was the case with the ADA project. Instead, the ADA model provided a set of guidelines for local archaeologists, who carried out retro-archiving activities in a structured way within their own institutions.

Another difference was that the *eDNA* project was able to make effective use of the lessons that had been learned in the earlier ADA project. In particular, the *eDNA* project used the same stepwise approach that had been followed in the ADA project.

The ADA project involved a large quantity of unknown data files that were kept in a cupboard in one of the rooms of the Meertens Institute. The quantity of files and variety of media was the major challenge of this project. The *eDNA* project, on the other hand, began by examining the types of past archaeological projects that had resulted in digital data. The initial focus was therefore not on the data themselves, but on the excavations, research projects and the people who had been involved. Only after all of the archaeological projects had been listed and uniformly described with the help of Dublin Core, and after a final selection had been made of datasets that would be archived, did the attention shift towards the data files themselves. This was also one of the most important

recommendations of the ADA project: try to make an estimate of the magnitude of the project, and try to keep it manageable.

Data selection proved an important task in both the ADA and the *eDNA* projects. The selection of files from each dataset was particularly time consuming in the *eDNA* project. Both projects generated the conclusion that it is vital to determine good selection criteria at the start of a retro-archiving project, as it is neither possible nor useful to archive everything. It is necessary to decide—preferably with the help of the creator of the data file— which files should be considered important for future research.

Another conclusion from both projects was that the problems of long-term data archiving are certainly not primarily technical. Despite all of the obstacles that we encountered in these two projects, we think that the ADA approach is useful for retro-archiving projects, on the condition that is applied flexibly.

The project descriptions show that one of the major issues in the long-term preservation of digital data—the archiving strategy—has hardly been addressed. In both projects, it was decided early on to use the conversion strategy, as this was the accustomed strategy of the data archives. There was simply no opportunity to experiment with emulation. One of the problems with comparing the two strategies is that such tests require large-scale environments, whether within or outside the framework of a project. Opponents of the emulation strategy argue that, unlike the conversion strategy, which is used by many data archives and other institutions around the world, there have been no large-scale experiments with emulation (Tjalsma 2000).

The ADA approach is well suited for retro-archiving the research in the data programmes of DANS. Retro-archiving digital data will apparently remain one of the most important activities in digital archiving for some time to come. The future of the ADA approach is promising, therefore, particularly within the academic environment.

Acknowledgements Both authors are grateful to *eDNA* project leader Milco Wansleebe, archaeologist and staff member of the Faculty of Archaeology of the University of Leiden, for his help and editorial comments when writing this article. The ADA project was funded by iWI/SURF. At the time of the study, iWI was the research and development branch of the national SURF organisation, the higher education and research partnership organisation for network services and information and communications technology (ICT) in the Netherlands. The ADA project was executed by the Netherlands Institute for Scientific Information Services (NIWI), an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW). Until 1 July 2005 (and therefore during the time of the ADA project), the NHDA was part of NIWI. Since that time, the NHDA has been incorporated into DANS.

References

- Boonstra O, Breure L, Doorn P (2004) Past, present and future of historical information science. NIWI-KNAW, Amsterdam
- Chapman S (2004) Counting the costs of digital preservation: is repository storage affordable? *J of Digit Inf* 4/2: 1–15: <http://journals.tdl.org/jodi/article/view/jodi-113/99>. Consulted December 2006
- Computer Museum of the University of Amsterdam. <http://www.science.uva.nl/faculteit/museum>. Consulted December 2006
- DARE Repositories. <http://www.darenet.nl/en/page/language.view/repositories>. Consulted December 2006
- Darenet. <http://www.darenet.nl/en/page/language.view/dare.start>. Consulted December 2006
- Darenet search service. <http://www.darenet.nl/en/page/language.view/search.page>. Consulted December 2006
- Data Documentation Initiative. <http://www.icpsr.umich.edu/DDI>. Consulted December 2006
- Dublin Core Metadata Initiative. <http://dublincore.org>. Consulted December 2006
- eDNA* Project. <http://www.edna.leidenuniv.nl>. Consulted December 2006
- eDNA* Repository for archaeological data. <http://edna.itor.org/en>. Consulted December 2006

- Federal Geographic Data Committee. <http://www.fgdc.gov>. Consulted December 2006
- Jones M, Beagrie N (2001) Preservation management of digital materials. A handbook. The British Library, London: <http://www.dpconline.org/graphics/handbook/index.html>. Consulted December 2006
- Meertens Institute. <http://www.meertens.knaw.nl>. Consulted December 2006
- Mostert P et al (1998) Digitaal academisch erfgoed. Beleidsaspecten in verband met het behoud van wetenschappelijke digitale informatie [Digital academic heritage: Policy implications of long-term preservation of scientific or scholarly digital information]. SURF/IWI, Utrecht: <http://www.surf.nl/publicaties/index2.php?oid=35>. Consulted March 2006
- OAI-MPH. <http://www.openarchives.org>. Consulted December 2006
- Oltmans E, Kol N (2005) A comparison between migration and emulation in terms of costs. RLG DigiNews 9: http://www.rlg.org/en/page.php?Page_ID=20571. Consulted December 2006
- Rijksarchiefinspectie (2003) Informatie als kapitaal. Doorlichting archiefbeheer universiteiten: Verzamelrapport [National Archival Inspectorate: Information as capital. Screening the archival management of the universities]. Rijksarchiefinspectie, Den Haag
- Ross S (2000) Changing trains at Wigan: digital preservation and the future of scholarship. The British Library, London: <http://www.bl.uk/services/npo/pdf/wigan.pdf>. Consulted December 2006
- Tjalsma HD (2000) Reviews of two reports by Jeff Rothenberg on emulation. Hist Comput 12:374–377
- Woodyard-Robinson D (2006) Costs and business modelling. In: Jones M, Beagrie N Preservation management of digital materials. A handbook, Internet version, The Digital Preservation Coalition: <http://www.dpconline.org/graphics/inststrat/costs.html>. Consulted December 2006