Data Curation
in Arts and Media Research

About this publication

Data Curation in Arts and Media Research

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

info@surf.nl
www.surf.nl/en

Authors
Peter Verhaar – *Leiden University*
Mariya Mitova– *Leiden University and Brill Academic Publishers*
Paul Rutten – *Leiden University*
Adriaan van der Weel– *Leiden University*
Frederik Birnie– *Leiden University*
Abram Wagenaar– *HUB Uitgevers*
Joppe Gloerich– *Universiteit Leiden*

Editors
Gerhard-Jan Nauta – *DEN and Leiden University*
Rob Grim – *Tilburg University and 'The Open Data Foundation'*
Inge Angevaarde – *National Library of the Netherlands*
Heiko Tjalsma - *DANS*
Annelies van Nispen - *DEN*
Annemiek van der Kuil - *SURFfoundation*

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

21 June, 2010

Table of Contents

Management Summary 4
Samenvatting (Translation of the management summary in Dutch) 6
1 Introduction 8
1.1 Research data in the humanities 8
1.2 Aims and methodology 10
2 Description of the research data 11
2.1 Definitions 11
2.2 Primary sources 13
2.3 Primary resources 14
3 Motivations 18
4 Challenges 20
5 Workflow 23
5.1 Data curation lifecycle 23
5.2 Role of institutions 25
6 Conclusions 28
6.1 Future implications 28
6.2 Recommendations 30
Bibliography 32
Appendix A. Summary of interviews 34

Management Summary

This report, *Data Curation in Arts and Media Research*, presents the current state of data curation and data reuse practices within the fields of Comparative Arts, Art History and Media Studies. This report is written as part of the project Collectioneren van Data [Collecting Data], which is one of the work packages within the SURFshare programme 2007-2010, WP7: Data Curation and digital preservation.

Research was carried out on the basis of a literature review and a series of interviews among three stakeholder groups:

1. Researchers in the Comparative Arts, Art History and Media Studies.
2. Organisations active in the field of data curation.
3. Academic Publishers who have shown an interest in enhancing their publications with research data.

It was found that, in the course of a research project or study, a researcher may basically produce three classes of artefacts. Firstly, a researcher may produce digital surrogates of primary sources, in case cultural heritage institutions have not digitised the relevant materials yet. Secondly, a researcher may also produce research data. These can be highly heterogeneous. Examples include databases, spreadsheets, transcriptions or survey results. Thirdly, the results of analyses of these data are recorded in publications. This report focuses primarily on the curation of the first two types of products.

All researchers that have been consulted in this project acknowledged that transferring files to a data archive can produce clear benefits.

- Data curation by an external agent would release them from the burden of having to manage the files themselves.
- Open access publication of research data enables other researchers to reuse these data in their own projects. This may lead to larger, more comprehensive analyses and the reuse of data in ways which were not foreseen when they were originally created.
- When research data are archived systematically, researchers can verify the claims and the conclusions of publications.

The researchers acknowledged that the reuse of data is not yet common. An important obstacle is that it is currently very difficult to discover what datasets other scientists are actually producing. As for releasing data, most researchers state this can only happen after they have been discussed in a published text. This is because sources such as databases or models have often required much intellectual effort to produce, but earn no scholarly credits on their own. The value of humanistic studies is often related to the uniqueness or the originality of those ideas, which data sharing might undermine.

Due to the lack of fixed procedures or codes of conduct, researchers have generally used their own insights and their common sense during the design and the organisation of their datasets. It is highly unlikely that quality assurance criteria will be developed in the near future. Nevertheless, the researchers who had some experience with reusing data, stated that they were mostly capable of evaluating the scientific accuracy. The reuse of data produced by colleagues is largely based on trust.

The report also proposes a workflow for the curation of research data. The lifecycle starts when researchers create research data, as part of a specific research project. Metadata should be assigned during or shortly after the creation of the files, as high-quality metadata are indispensible for an effective curation of the data. The compulsory metadata fields must be negotiated with the projected data repository. The various objects which are created need to be evaluated, and a decision must be taken on which objects are worth preserving. Next, the selected data need to be ingested into the data archive. The infrastructure for the deposit of research data with a data centre should have a front office which is integrated as seamlessly as possible into the working

environment of the researchers. After the ingest into the Trusted Data Repository, the maintenance of the dataset is no longer a concern of the researcher. This responsibility shifts to professional data archivists. There may not always be a need to store research data indefinitely: under certain circumstances, the data may also be deleted from the archive, possibly based on usage statistics.

The number of institutions that can actually function as Trusted Data Repositories for research data produced in the Netherlands appears to be limited. Within most universities, the expertise and the resources which are needed to set up such dedicated data centres are lacking. According to a 2010 report of the *Nationale Coalitie Digitale Duurzaamheid*, the National Library (KB) in The Hague should focus primarily on sustainable access to textual publications, and DANS will be the most evident destination for research data. If different institutions will be responsible for the curation of data, it is also important to create a central portal through which researchers can discover relevant datasets. Service providers such as NARCIS or DRIVER may ultimately have an important function in allowing researchers to discover data.

Academic publishers can also have a function in the dissemination of datasets. The scholarly open access article may be enhanced with links to datasets or other resources that are referred to in the text. In this situation, the enhanced publication may be seen as a documentation to the dataset it accompanies. Academic publishers may assist scholarly authors in the selection and the certification of data. Nevertheless, peer review of data is not a task that publishers seem to be willing to take on, as this is considered to be the responsibility of the scientific community.

Although sharing data is certainly not yet common among scholars, the attitude towards data sharing was found to be positive. A growing willingness to share research data, combined with an increased use of digital instruments, will open up possibilities to pose and answer new research questions. It must be recognised that the actual reuse of data will require a standardisation of terminologies, ontologies and practices. Securing the semantic or intellectual interoperability of the various resources will clearly be a next major challenge. Currently, there are very few incentives for scholars to share their data, as they do not receive any rewards for resources other than textual publications. However, when more and more data are shared, as a results of an institutional mandate, for instance, the importance of research data in the overall assessment of scholarly impact is likely to expand. Peer review processes and other quality assessment procedures will no longer be based solely on the written record of the research in article or a book, but also on the review of the data, providing transparency in the process and the management of the workflow. In this sense, data curation will also have considerable political implications for the valuation of humanities scholarship in the future.

Samenvatting (Translation of the management summary in Dutch)

Dit rapport, *Data Curation in Arts and Media Research*, schetst een beeld van de huidige stand van zaken rond het beheer en het hergebruik van onderzoeksdata binnen de kunstgeschiedenis, de vergelijkende kunstwetenschap en de mediastudies. Het rapport is geschreven als onderdeel van het project *Collectioneren van Data*. Dit project valt binnen het SURFshare programme 2007-2010, WP7: *Data Curation and digital preservation*.

Het onderzoek is uitgevoerd op basis van een literatuurstudie en een reeks van interviews met vertegenwoordigers van drie groepen van belanghebbenden:

1. onderzoekers op het gebied van de kunstgeschiedenis, de vergelijkende kunstwetenschap en de mediastudies.
2. Instellingen die actief zijn op het terrein van het beheer van wetenschappelijke data.
3. Wetenschappelijke uitgeverijen die belangstelling hebben voor het publiceren van wetenschappelijke teksten in samenhang met onderzoeksdata.

Er is vastgesteld dat wetenschappers tijdens hun onderzoeksprojecten drie soorten producten aanmaken. Onderzoek in de genoemde disciplines richt zich vaak op objecten die worden beheerd door erfgoedinstellingen zoals musea, bibliotheken, of archieven. Wanneer een bepaald object nog niet is gedigitaliseerd door de beherende instelling kunnen onderzoekers er voor kiezen om zelf digitale reproducties te vervaardigen. Op de tweede plaats maken onderzoekers onderzoeksdata aan. Het gaat hierbij om databases, spreadsheets, of transcripties van enquêteresultaten. In dit soort bestanden worden over het algemeen kwalitatieve of kwantitatieve data over de onderzoeksobjecten vastgelegd. Aan de hand van dit soort data worden analyses uitgevoerd die uiteindelijk worden vastgelegd in publicaties.

De onderzoekers die in het kader van dit rapport zijn geïnterviewd onderstreepten dat het delen van onderzoeksdata duidelijke voordelen kan bieden.

- Wanneer een externe partij de data beheert hoeven wetenschappers hier zelf geen tijd en moeite meer in te investeren.
- Wanneer de data vrij toegankelijk zijn kunnen andere onderzoekers deze data hergebruiken. Op termijn kunnen er hierdoor grotere en meer omvattende onderzoeksprojecten worden opgezet. Data kunnen ook worden gebruikt op manieren die nog niet werden voorzien op het moment dat de data werden aangemaakt.
- Wanneer onderzoeksdata systematisch worden gearchiveerd kunnen onderzoekers de beweringen en de conclusies die in publicaties voorkomen ook beter verifiëren.

Momenteel is het binnen de kunstwetenschappen en de mediastudies nog niet heel gebruikelijk om onderzoeksdata van collega's te hergebruiken. Onderzoekers missen een goed overzicht van de databestanden die door anderen zijn aangemaakt. Onderzoekers zijn meestal ook niet bereid om hun data te delen voordat zij over deze data hebben gepubliceerd. Ze worden over het algemeen beoordeeld op basis van hun publicaties, en het beschikbaar stellen van databestanden levert op zichzelf geen duidelijke beloningen op. De waarde van geesteswetenschappelijk onderzoek hangt meestal samen met de uniciteit of de originaliteit van ideeën, en wanneer data te vroeg openbaar worden gemaakt ontstaat er het risico dat een collega met deze ideeën aan de haal gaat.

Er zijn geen vaste procedures of richtlijnen voor aanmaken van onderzoeksdata. De structuur van databases hangt in veel gevallen sterk af van de specifieke onderzoeksvragen die worden gesteld. Het is niet waarschijnlijk dat er op korte termijn duidelijke richtlijnen zullen ontstaan binnen de kunstwetenschappen. De geïnterviewde onderzoekers gaven echter aan dat zij zich over het algemeen in staat voelden om de wetenschappelijke waarde van data van collega's te beoordelen. Het hergebruik van data is vaak ook gebaseerd op vertrouwen.

In dit rapport wordt eveneens een voorstel gedaan voor een procedure voor het beheer van onderzoeksdata. De levenscyclus van data gaat van start wanneer deze worden aangemaakt als onderdeel van een onderzoeksproject. Het verdient aanbeveling om metadata-beschrijvingen zo spoedig mogelijk na de aanmaak van de bestanden aan te maken. Het ligt voor de hand om deze

taak onder te brengen bij de wetenschapper die verantwoordelijk is voor de data. Met de instelling die de data uiteindelijk zal beheren moet overeenstemming worden bereikt over welke kenmerken moeten worden vastgelegd. De verschillende databestanden die worden aangemaakt moeten vervolgens worden geëvalueerd, en er moet daarbij een beslissing worden genomen over welke bestanden zullen worden gearchiveerd. De geselecteerde bestanden moeten hierna worden overgedragen naar een datacentrum. Het aanleveren van de data moet idealiter kunnen plaatsvinden via een mechanisme dat zo naadloos mogelijk aansluit op de directe werkomgeving van de onderzoeker. Na de invoer in het digitale archief ligt de verantwoordelijkheid voor het beheer van de bestanden bij het datacentrum. De bestanden hoeven in veel gevallen niet tot in de eeuwigheid te worden bewaard. De beslissing om data wel of niet te bewaren mogelijk worden gebaseerd op gebruiksstatistieken.

Het aantal instellingen in Nederland dat daadwerkelijk de rol van databeheerder op zich kan nemen is vrij beperkt. Binnen universiteiten ontbreken vaak de expertise en de technische middelen om een datacentrum in te richten. In een rapport van de *Nationale Coalitie Digitale Duurzaamheid* dat begin 2010 is verschenen wordt aangegeven dat de KB in Den Haag zich hoofdzakelijk zou moeten richten op het archiveren van publicaties, en dat DANS de meest aangewezen bestemming is voor onderzoeksdata. Wanneer verschillende instellingen een verantwoordelijkheid dragen bij het archiveren van data is het wel van belang om er voor te zorgen dat er een centrale portal is waarop onderzoekers een overzicht kunnen vinden van gearchiveerde databestanden. Diensten zoals NARCIS of DRIVER zouden op termijn een belangrijke functie kunnen gaan vervullen bij het bieden van toegang tot onderzoeksdata. Ook voor wetenschappelijke uitgeverijen is hier mogelijk een rol weggelegd. Publicaties kunnen bijvoorbeeld worden verrijkt met databestanden, en in dat geval kan de wetenschappelijke tekst worden gezien als documentatie bij deze data. De uitgeverijen die in het kader van dit project zijn geïnterviewd gaven aan dat zij niet bereid waren om ook faciliteiten te ontwikkelen voor het controleren van de wetenschappelijke kwaliteit van onderzoeksdata. Dit werd gezien als een verantwoordelijkheid van de wetenschapper zelf.

In algemene zin kon er worden vastgesteld dat onderzoekers een positieve houding hebben ten aanzien van het delen van onderzoeksdata. Wanneer grote hoeveelheden data beschikbaar worden gesteld via het internet biedt dit duidelijk nieuwe wetenschappelijke mogelijkheden. Het is echter ook duidelijk dat concreet hergebruik van data wordt belemmerd door het feit dat terminologieën en onderzoekspraktijken vaak nog sterk verschillen. Ook wanneer onderzoekers hun onderzoeksdata op grote schaal zouden uitwisselen bestaat er nog de uitdaging dat deze data vaak in semantisch opzicht niet interoperabel zijn.

Op dit moment zijn er weinig motivaties voor onderzoekers om hun data te delen, omdat deze bestanden op zichzelf niet ‘meetellen’ bij beoordelingen van wetenschappelijke productiviteit. Wanneer data beschikbaar komen wordt het hele onderzoekstraject meer transparant, en naarmate er meer databestanden on-line worden geplaatst kan er worden verwacht dat het belang van deze data bij het meten van wetenschappelijke impact zal toenemen. Peer review processen en andere kwaliteitsmetingen zullen niet meer alleen meer zijn gebaseerd op het eindresultaat van het onderzoek, de publicatie, maar ook op de onderzoeksdata die tijdens een studie zijn aangemaakt. Op deze manier kan een systematisch databeheer ook aanzienlijke politieke consequenties hebben bij de evaluatie van wetenschappelijk onderzoek binnen de geesteswetenschappen.

1 Introduction

1.1 Research data in the humanities

In her seminal study *Scholarship in the Digital Age*, Christine Borgman argues that research data form "the foundation of scholarship" (p. 115). They are often the first concrete results of academic research and comprise the vital ingredients for subsequent publications. Data are considered to be a key element in the "chain of evidence" that underlies scholarly research in all disciplines (Marlo Welshons et al., 18). In many cases, research data are not only relevant to the researchers who had originally produced them. They can also be highly serviceable to colleagues who are working on related questions. Open access to research data can clearly contribute to the advancement of scholarship, as academics who reuse the resources that have been produced by others can progress more quickly to the discovery stage of the research. "Education, scholarship and research all require the sharing of data and the communication of results" (ibid, 22). In The Fourth Paradigm, Hey et al. (2009) note that research in fields such as high-energy physics, astronomy and genomics is increasingly powered by digital research instruments and grid-enabled applications, which currently produce data at rates which clearly outstrip the possibilities to analyse them. In such data-intensive fields, scientists increasingly rely on sophisticated search tools and visualisation techniques which enable them to trace patterns and to make new discoveries on the basis of massive sets of research data. A new type of methodology thus appears to be emerging, in which discoveries are mainly made by mining existing datasets.

Importance of digital research data

The importance of digital research data is clearly growing, but co-ordinated approaches to the preservation of research data are crucially lacking. Whereas the 2003 Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities stated that open access contributions may also include "original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material",[1] efforts to implement open access to research results have predominantly focused on publications. An inventory made in 2008 among 192 repositories in 22 European countries indicated that only 8.4% of these contained primary datasets (Van der Graaf, 2009), and that, more generally, the repositories that contain non-textual materials clearly form a minority. Nevertheless, there are some notable exceptions. The three technical universities in the Netherlands (TU Delft, TU Eindhoven and the University of Twente) have jointly established a data centre in order to provide long-term storage for technical-science data.[2] Secondly, Data Archiving and Networking Services (DANS) is an important institute under the auspices of Royal Netherlands Academy of Arts and Sciences (KNAW) which aims to preserve and to provide access to research data in the arts and humanities and social sciences. When the number of deposited datasets is compared to the number of research projects that are registered by the KNAW, it becomes clear that these data centres only cover a small section of all the data that are actually produced. Unfortunately, for a variety of reasons, many of these resources ultimately get lost.

Return on public investment

The 2004 OECD *Declaration on Access to Research Data from Public Funding*[3] highlights the usefulness of making underlying research data available for reuse "beyond the original project for which they were gathered, in other fields and in industry" (p. 9). According to the report, an important outcome of the availability and reuse of research data from publicly funded projects is the increased return on public investment through reducing the duplication of data collecting efforts. Investigating the way OECD requirements can be implemented within the Finnish academic environment, Kuula and Borg (2008) demonstrate that the possibility of reusing data after it has been collected is a definite advantage to the scientific community nationally and internationally. Besides increasing the efficiency of public funds, continued access to research data improves the productivity and quality of research (p. 24).

1. http://oa.mpg.de/openaccess-berlin/berlindeclaration.html
2. www.datacentrum.3tu.nl/
3. www.oecd.org/dataoecd/9/61/38500813.pdf

Exploring the needs of researchers

Developing facilities for the curation of data is complicated by the fact that different disciplines differ widely in terms of research methods and the type of data that are produced. A recent report produced by the SCARP project in the United Kingdom concluded that "researchers' attitudes and practice with regard to the creation, sharing, reuse and long term care of data are closely linked to the discipline in which they work". For this reason, "a generic approach to data curation will not be sufficient to cope with the different data-related needs and expectations of researchers working in different disciplines other than at a superficial level" (Key Perspectives, 2010, p. 2). These findings suggest that the key to enhancing the Dutch infrastructure for data curation lies in firstly exploring the needs of researchers working in fields across the entire academic spectrum, and, consequently, to develop facilities and guidelines which can adequately address these domain-specific needs.

Disciplines differ

Disciplines may differ in terms of methodologies and culture, but also in terms of the volumes of the data that are produced. Borgman (2007) observes that "the predicted data deluge is already a reality in many fields" (p. 115), but it is also clear that this problem of abundance is not equally poignant in all disciplines. The terabytes of data that are generated by astronomers or physicists often form a shrill contrast to the relatively modest datasets that are produced by humanities scholars or social scientists. The SCARP project found that, in comparison, "researchers in the arts and humanities do not publish a great deal of research data" (Key Perspectives, 2010, p. 5), even though, as Welshons et al. (2006) point out, humanists have always relied on collections of information for their research, and that assembling these collections has been considered part of their scholarship (p. 7). This difference can be explained in part by the nature of the research. Humanists study phenomena that are related to language, history, philosophy, religion and culture (Ibid), and these phenomena can normally not be recorded by automated measuring devices. Results in humanistic research are often based on critical interpretation rather than on standardised objective observations. Furthermore, there are barely any established practices or protocols which can govern hermeneutic analyses.

The heterogeneity of humanities research materials

An additional difference between the humanities and the sciences is the scope of data that humanists can use as the basis of their scholarship, which is as broad as it is diverse. Harley points out that "[w]hile much (but far from all) data within the physical and biological sciences are relatively more comparable and can be deposited into common databases, no such 'common denominator' exists for social and humanistic data, since data types, sources, and collecting practices can vary so widely". In *Our Cultural Commonwealth* Welshons et al. describe the subject of humanities research as "vastly more messy and idiosyncratic" than that of the sciences (Welshons et al., p. 7). Brockman (2007) also talks about the heterogeneity of humanists' research materials coming from a wide range of sources (p. 2).

Importance of digital data within the humanities

Nevertheless, there are clear signs that the importance of the digital medium is expanding within the humanities as well. The DARIAH project, for instance, aims to "provide a coordinated technical infrastructure for supporting the preservation of cultural heritage in Europe, and will enable dramatically improved access to research material for the humanities".[4] Similarly, the CLARIN[5] project aims to implement an international infrastructure for linguistic research by providing uniform access to the contents of distributed digital archives and to the various language and speech processing tools that have been developed. In the Netherlands, the *E-depot for Archaeology*[6] has been established to ensure that unique primary data on archaeological excavations in the form of images and GIS files can be archived and made available to other scholars. Dutch archaeologists are legally bound to deposit their research data in this archive. Whereas it is generally agreed that humanities scholars do not produce as many digital data as researchers in other fields, the examples that have been cited clearly illustrate the manner in which humanities researchers can benefit from a co-ordinated effort to secure the long-term accessibility of relevant digital resources.

4. www.dariah.eu
5. www.clarin.eu
6. www.edna.nl

1.2 Aims and methodology

This report is written for the project *Collectioneren van Data* [Collecting Data], which is part of the SURFshare programme 2007-2010, WP7: *Data Curation and digital preservation*.[7] It provides recommendations for a national policy regarding data curation. In line with findings by Lyons (2010), the scope of this study has been confined to a specific field of research. The recommendations in this report apply primarily to the research data that are produced by researchers in the Comparative Arts, Art History and Media Studies. The main aim is to investigate existing research practices, the resulting products of these practices, their preservation and curation, and the attitudes of these particular scholarly communities towards data preservation and reuse.

In a first stage of the research project, a conceptual framework was developed on the basis of a literature study. In a second stage of the project, ten semi-structured interviews were conducted with representatives of the various stakeholders in the field. Three main stakeholder groups have been identified:

4. Researchers in the Comparative Arts, Art History and Media Studies. For the purpose of this study, five meetings were organised with scholars who actively produce or reuse digital research data.

5. Organisations active in the field of data curation. Interviews were held with employees of the *KB, National Library of the Netherlands,* in The Hague,[8] the *Netherlands Institute for Art History*[9] and *DANS*.[10]

6. Academic Publishers who have shown an interest in enhancing their publications with research data. The potential advantages of data curation have been discussed with representatives of *Koninklijke Brill* in Leiden[11] and *Amsterdam University Press.*[12]

These three stakeholder groups also form the main target audience of this report. The findings of this report will also be relevant to national funding bodies such as NWO and SURFfoundation.

The results of the literature review and the interviews have been combined to produce a clearer picture of the present state of data curation and data reuse practices within the fields of Art History and Media Studies. The following section gives an overview of the fields from the humanities that are the specific focus of this investigation, based on existing literature, theoretical constructs, and descriptions of university programmes. The report tries to raise awareness of the current difficulties and recommends actions that may help to solve today's most important bottlenecks.

7 www.surffoundation.nl/surfshare
8. www.kb.nl
9. http://website.rkd.nl
10. www.dans.knaw.nl
11. www.brill.nl
12. www.aup.nl

2 Description of the research data

2.1 Definitions

In this chapter, a general description is given of the disciplines that are investigated in this project, and of the data that are produced in these fields. The term “data” is a very generic term, which can be used to refer to different types of artefacts, depending on the stage in the research process. For this reason, this section will firstly provide a number of descriptions of the terms that will be used in this text.

Data may be classified based on the way it is used in research as input and the level of analysis it has undergone. Some researchers would prefer to use products that already formulate some sort of findings or report patterns derived from data, while other researchers would rather run their own analyses on raw, unprocessed data. Borgman (2007) suggests that different stages or levels of data need to distinguished, but also concedes that “[d]ata levels are more ambiguous in the qualitative areas of the humanities and social sciences” (p. 121). This report will make a basic distinction between the terms “source” and “resource”.

Sources and resources
Sources are all the materials that are available to scholars and which enable them to do their research. They are created by others, outside the research process of the researcher or research team itself, and form the input for scholarly studies. Resources, on the other hand, are the “data, documents, collections, or services that meet some data or information need” (ibid, p. 122). They are the artefacts which are created by the scholars themselves in the course of their study. In this context, the term ‘resource’ does evidently not refer to research funds or personnel. Sources can be divided into primary sources and secondary sources. Primary sources are understood as the phenomena or the physical or born-digital objects which are investigated, and secondary sources are the scholarly works which interpret primary sources or which place these in a certain context. These two terms are usually conflated in the case of historical or philological research that takes place on the basis of older scholarly texts.

A similar distinction can be made between primary resources and secondary resources. In this report, the term “primary resource” will be considered to be synonymous with “research data”, and the term “secondary resource” will be taken to refer to scholarly publications. Examples of primary resources include databases, research annotations, models or visualisations of primary sources produced by the scholars themselves. These scholarly artefacts usually form the basis of secondary resources, which are “reports of research, whether publications or interim forms” (ibid, p. 122), also written by the scholar. At a risk that the terminology may be further complicated, it will be added that many authors also make a distinction between primary and secondary research data. Primary data are essentially the uninterpreted facts. They are the direct result of the recording or observation process. Such raw data may not be directly suitable for analysis because they may contain all kinds of errors, misspellings or other faults. When data are normalised, they are still primary but not raw. In a subsequent phase of the research cycle data may be systematised, coded or ordered in some way or another to make them more fit for analysis. These types of data are referred to as secondary datasets (Tjalsma et al., 2010, p. 10). This description of the various products of the scholarly process is largely compatible with the distinctions which are made by Doek et al. (2009). The report that was produced by the IISH, *IISH Guidelines for preserving research data: a framework for preserving collaborative data collections for future research*, firstly distinguishes ‘research data collections’ and ‘source-corresponding datasets’. These types of resources are both considered to be primary data. The IISH report also distinguishes ‘processed datasets’ and ‘datasets for analysis’. The latter two types of research data are both taken to be secondary data.

Figure 1 provides a visual summary of the various terms that were discussed in this section.

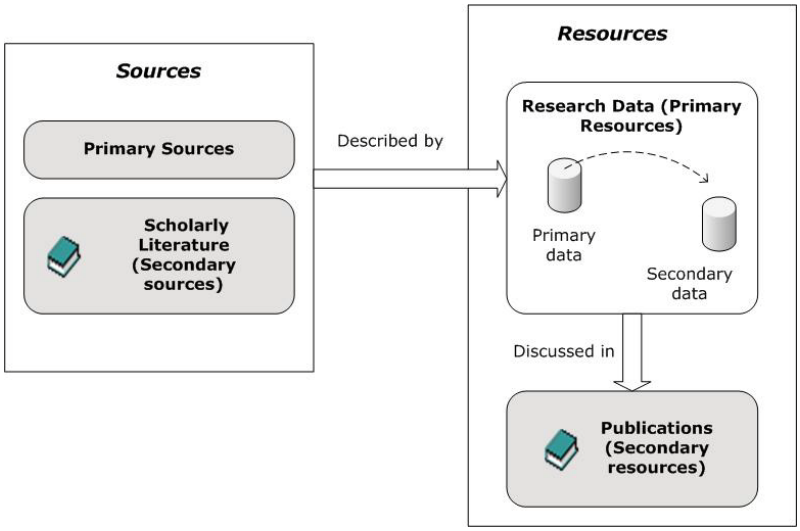

Figure 1. Sources and resources in scholarly research.

Kuula and Borg (2008) describe research data in juxtaposition to final research publications (p. 28), and base this discussion on inherent features relating to their level of processing, readiness for use, etc. Figure 2 summarises this classification:

Research publication	Research data
Information transformed into results	Information not transformed into results
Use requires basic software and instruments and their command	Use often requires special software and instruments and their command
Self-explanatory	Requires additional information and documentation if not archived
Should not include sensitive information	May include sensitive and confidential information
Use does not require permission	Use often requires permission
Ownership and copyright often clear	Ownership and copyright often unclear
Openly accessed by the scientific community for a fee or for free	Several degrees of openness (from completely open to closed)
Understood as scientific output (mentioned in the CV)	At the moment not understood as scientific output/merit even if the data were published (usually not mentioned in the CV)
Ready to be used by others as such	Use requires processing

Figure 2. Differences between research data and publications

Any guidelines for the preservation and curation of research data should therefore specify if the recommendations apply to primary or secondary sources or to primary or secondary resources or to a combination thereof, and what these refer to specifically.

2.2 Primary sources

Art History, the Comparative Arts and Media Studies are generally considered to be part of the Humanities, although Media Studies is sometimes classified under the Social Sciences as well. Unlike other fields in the humanities, such as literary studies or philology, Art History predominantly investigates the visual or auditory aspects of human artefacts. Studies focus on moving or still images (paintings, movies, photographs, drawings), music, three-dimensional objects (sculptures, installations, architecture), or born-digital materials such as video art, internet movies or computer games. The field of Media Studies has been defined in many different ways in existing literature and in overviews of related university programmes. Even the name of the discipline varies from Media Studies to multimedia, new media, hypermedia, etc. (Rockwell and Mactavish, 2004, p. 108). A formal definition of multimedia has been given by Rockwell and Mactavish in *A Companion to Digital Humanities*, as a discipline studying works that are "computer-based rhetorical artefact[s] in which multiple media are integrated into an interactive whole" (ibid). Media Studies scholarship may entail the reflection on computer-based, multimedia works such as web hypermedia, computer games and digital art through the critical study of the history, definitions and different types of media and their contents.

Improving digital access to primary sources

In general terms, it may be argued that the primary sources of scholars who focus on the arts are formed by objects which are created by human beings for artistic purposes. These sources are often held by institutions such as museums, archives or libraries. Historically, these sources could only be investigated by visiting these institutions physically. Most cultural heritage institutions have recognised that digital access to their collections is essential for the emergence and the further development of scholarship and have digitised large sections of their holdings. In the last few decades the availability of primary sources in digital form has been increasing at an ever accelerating speed. National digitisation programmes, such as *Metamorfoze*[13] in the Netherlands, help institutions to scan collections of historical interest and ultimately allow researchers to do part of their research via the internet. In addition, a consortium of the KB, university libraries and KNAW institutes is currently looking for major funding for a project called 'Libratory'. The initiative should lead to the digitisation of all Dutch books before 1840. These should be made available to researchers in a sophisticated online work environment equipped with an array of advanced analytical tools. Libratory has been called a "particle accelerator for the humanities". This image, chiefly intended to speak to the imagination of those who have to decide on the project's funding, is apt for at least one reason: the investment involved is (comparatively speaking) huge. An investment of that order can only be made if the project is supported by a large number of institutions and individuals. There are also various international initiatives to improve and to centralise the access to these growing volumes of digital content. One notable example is the *Europeana* portal,[14] which currently offers digital access to over a million objects from the Rijksmuseum in Amsterdam, the British Library in London and the Louvre in Paris, among many other institutions.

Audiovisual media

Researchers in the field of Media Studies often focus on the analysis of moving images and audiovisual media. In the Netherlands, there are several institutes which aim to preserve this audiovisual cultural heritage. A very prominent example is the Netherlands *Institute for Sound and Vision*, which archives more than 700,000 hours of television, radio, music and film. New material is added on a daily basis. The *EYE Film Institute Netherlands* is a relatively new organisation, responsible for the preservation of the Dutch cinematographic heritage. The Institute owns more than 37,000 films, 500,000 photographs and 60,000 film posters.[15] Both the *Institute for Sound and Vision* and the *EYE Film Institute* participate in the collaborative digitisation project *Images for the Future*. The aim of this digitisation project is to make the collections of the participating institutions "available to education and to the public as broadly as possible".[16] This project, which

13. www.metamorfoze.nl
14. http://www.europeana.eu
15. www.eyefilm.nl/eye
16. www.beeldenvoordetoekomst.nl

has been underway for some years now, provides an enormous amount of digitised research materials.

Artistic research
Within disciplines such as Comparative Arts and Media Studies, the focus is usually not only on cultural artefacts, but also on the wider historical, social and cultural context in which these artefacts are produced, distributed or consumed. Primary sources needed to answer these types of questions can usually be found in public archival institutions or in the private archives of artists. The NWO-funded project *Cinema Context*, which is led by professor Karel Dibbets, focuses on the reception of movies in the Netherlands, and accumulates information on the locations of movie theatres and on the movies that were shown in these theatres. Such data are mostly derived from local archives and from historical newspapers and magazines. The field of artistic research, in the form that is practiced by researchers such as professor Henk Borgdorff, also needs to be mentioned in this context. This line of research views art not only as an expressive practice, but also as a field of scholarly contemplation and research. Artistic creation is, in a sense, highly comparable to the way in which architects and engineers commit themselves to construction works. Both engineered and artistically created objects are the result of a research process covering trial, error and falsification. Artistic research distinguishes itself from the more traditional forms of art historical research by the discursive context in which art researchers reflect on their specific activities and processes, and subsequently share these reflections with peers. In this field of research, the documentation that is maintained by the artists themselves also produces important primary sources.

Performing arts
Although artistic expressions such as theatre or dance performances appear to be “in defiant opposition to the computer’s rapacious tendency to translate everything into disembodied digital data” (Saltz, 2004, p. 121), scholars who focus on the performing arts have also found various ways to procure digital access to their primary sources. Scholars interested in drama or ballet increasingly make use of photographs or video recordings of performances, and even create 3-D computer models of theatres and digital visualisations of staging strategies (ibid, p. 123).

Digital reproductions
It may be said that, generally, all humanities scholars can benefit immensely from having their primary sources available in a digital form. When cultural heritage items in disparate locations are made available digitally, this enables scholars who study them to juxtapose and compare sources held in different locations in one environment, on “the Web as the carrier of multimedia research” (Greenhalgh, 2004, p. 31). A number of researchers interviewed had mentioned that, in some cases, the digital reproductions which were provided by cultural heritage institutions were insufficient for their research, because of low scanning resolutions. Alternatively, it is frequently the case that the relevant objects are simply not yet available in a digital form. Priorities of cultural heritage institutions do not always coincide with those of academic researchers. At the VU University in Amsterdam and at Leiden University, the departments of Art History have both set up their own image repositories. Such repositories contain scans of slides with reproductions or scans from museum catalogues or other books with facsimiles of art objects. It is probably inevitable that researchers are sometimes forced to make their own digital reproductions, because the “volume and variety of content are vast”, and because “the frequency and conditions of reuse are difficult to anticipate” (Borgman, 2007, p. 216). Providing digital access to objects of humanistic study is a never-ending task, since “almost any document, physical artefact, or record of human activity can be used to study culture” (ibid).

2.3 Primary resources

In general, the objective of research that focuses on the arts is to understand the manner in which the materials are produced, to interpret the contents of the objects, or to investigate the reception of these objects.

Descriptions
Academic work in the fields of Art History and Media Studies often begins by making descriptions of

the relevant cultural artefacts. A number of respondents in this project had explained that the results of such initial analyses are usually recorded on paper, and sometimes in a digital text document. Such first annotations are usually not structured or standardised in any way. Nevertheless, a number of researchers also record descriptions and interpretations of primary sources in database programs, spreadsheets or reference management systems. Greenhalgh (2004) explains that "[t]he fact that primary data for art historians are visual – images, still and perhaps moving – did not prevent an early interest in the ordering and interrogation of text data arranged into databases" (p. 31). The data that are aggregated may be of a quantitative nature, when the physical dimensions or the years of creation are relevant, but descriptions are usually interpretative and qualitative. Computer applications are used "to order, sort, interrogate, and analyze data about artworks, preferably with images" (ibid, p. 31). Similarly, "attempts by performance scholars to tap the power of computers relied on computers' ability to crunch textual and numeric data, such as dramatic dialogue, critical texts about performance, names, locations and dates" (Saltz, 2004, p. 122).

Automatically extracted and user-generated metadata

In some cases, information can be derived from other databases such as library catalogues or from finding aids created by museums or archival institutions. The audiovisual archive of the Dutch *Institute for Sound and Vision* contains more material than its cataloguing department can ever describe, and, for this reason, this institute has developed sophisticated technologies that can automatically extract descriptive information from the digital files. The *Institute for Sound and Vision* has also found a way to involve non-staff members in the cataloguing process. The website *waisda*[17] presents the task of assigning descriptive tags in the form of a game, and thus stimulates end-users to do part of the work which was traditionally carried out by staff only. Such automated or user-generated generated metadata strongly improve access, and thus produce exciting prospects for the humanities. For scholars, it becomes easier to explore and analyse content from many different sources, from audiovisual to written texts, using advanced analytical tools and instruments as well as the metadata added to these primary sources by collecting institutes such as academic libraries and audio-visual archives. Nevertheless, it has also been acknowledged that the descriptive information provided by cultural heritage institutions often lack the information which is needed to answer specific research questions. The formal metadata may be used as a basis, but it usually needs to be expanded to make them more suitable for academic research. The interpretative, analytical information normally needs to be added by researchers themselves.

Databases and spreadsheets

In art historical research, database applications are "seen as able to provide intelligent answers to researchers' questions" (Greenhalgh, p. 33). The data that are recorded in spreadsheets or databases can be manipulated and transformed into formats that enable researchers to answer their research questions. Nevertheless, it is not common for humanities scholars to produce secondary data in addition to the primary data. Scholars may create databases or spreadsheets in which the records can be queried, sorted or annotated, but the results of these queries or manipulations are rarely stored as separate files. Since researchers generally do not derive new datasets from existing datasets, the distinction between primary data and secondary data will be emphasised in this report.

A clear example of a database that is created to support research is the extensive resource that is created as part of the *Cinema Context* project. This database enables researchers to trace the reception of a particular movie. Other examples include the various databases that are created by the Netherlands Institute for Art History (RKD).[18] The RKD is not only a documentation centre but also a research institute which investigates the biographies of artists and of art historians. Much of the research focuses on the ascription of works of art to specific creators, and the results of these studies are often captured in online resources such as RKDArtists&, RKDImages and RKDPortraits.

Lists and classifications

Researchers who describe and analyse cultural artefacts often consult or produce lists of personal names and geographic terms. A number of researchers have indicated that they make use of online

17. www.waisda.nl. the website derives its basic idea from the ESP game which was developed earlier by Luis von Ahn, see: <www.espgame.org/gwap/>
18. http://english.rkd.nl

resources such as the GeoNames website[19] or the Internet Movie Database.[20] The IconClass taxonomy or the *Arts and Archaeology Thesaurus* (AAT) are also useful resources for researchers. Researchers appreciate if they can simply build on the work that is carried out by others, and if they can rely on external databases as being authoritative sources on certain topics. Nevertheless, because of unfamiliarity with or unavailability of globally registered and authorised lists, researchers often produce their own lists and classifications.

Multimedia

In addition to databases, researchers who focus on the arts may also create many other types of research data. They may make sound recordings of interviews, create transcriptions of such interviews or of voice-overs of TV programmes. One respondent also noted that collections of hyperlinks could also be seen as research data. In the case of video games research, the emphasis often lies on the way in which users play the games. In this case, primary data may consist of survey results captured in Excel or of SPSS files, or of specific scenarios of game play which are recorded within the consoles that are used to play these games. There may also be video recordings of the people who play the computer games. The video files can be used to analyse the reactions to the games. Other derivative scholarly products resulting from Art History research include iconographic catalogues, controlled vocabularies and authority files.

The craft of multimedia itself can also be considered as scholarship in this field. As Rockwell and Mactavish observe, conceptualisation of the field can be achieved through the "use of multimedia to think and to communicate thought," which is a popular approach for emerging and dynamically developing disciplines such as Media Studies (ibid, 117). Notably, the two modes of scholarly exploration of the field of Media Studies, "thinking about" and "thinking with," have their own "traditions of discourse, forms of organization, tools, and outcomes." In other words, the products of Media Studies scholarship can range from traditional paper-based publications such as scholarly articles and monographs to complex, interactive, multiple-media artefacts. A similar attitude can be found in the way Henk Borgdorff implements a dynamic discourse of interpretation and perpetual revision of artistic research, published via the emerging *(e-)Journal of Artistic Research* (JAR).[21] His field of research is by definition characterised by multi-interpretability, and the Research Catalogue that will be set up to bolster the JAR will be open to discussion. Social interaction and the publication of artistic research as primary and secondary resources will create depth in an extensive and heterogeneous database.

Primary and secondary data

In conclusion, there are essentially three classes of artefacts that can be curated. Firstly, there are digital surrogates of primary sources. These reproductions are mostly maintained by cultural heritage institutions, but, when such digital sources are unsuitable or unavailable, researchers may also have produced their own digital surrogates. Primary sources may either be analogue or born-digital. Secondly, researchers also produce research data. These data can be highly heterogeneous. They often consist of databases with descriptions, annotations or interpretations of cultural artefacts, but may also record details about the wider social or cultural context in which these objects are produced or consumed. Thirdly, the results of analyses are recorded in publications. Figure 2 visualises the various products which are generally created or consulted by scholars that are active in Arts History, Comparative Arts and Media Studies. Since digital surrogates can be either managed by cultural heritage institutions or by the researchers themselves, they are included twice in the diagram.

19. www.geonames.org
20. www.imdb.com
21. www.jar-online.net/call/call.html

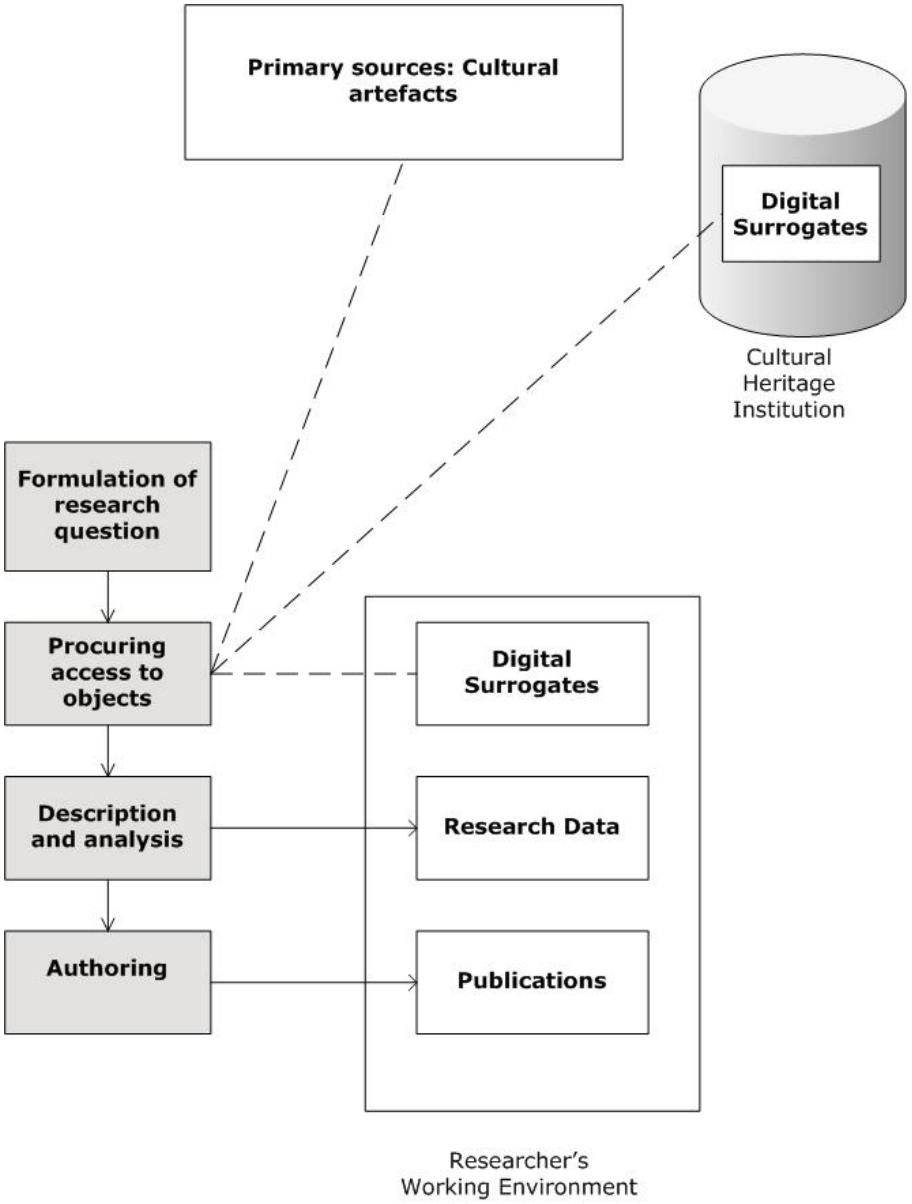

Figure 2. Scholarly products.

3 Motivations

Potential advantages of sharing data
Research data predominantly originate within the context of a particular research project, and primarily serve the needs of individual researchers. To realise the full potential of a well-functioning infrastructure for open access to research data, researchers need to be encouraged to focus not only on their own needs, but also on those of a wider scholarly community. At the moment, it is still common practice for many researchers to store the data themselves after the conclusion of the research project. Data are stored on university networks, on DVDs, or on a home computer. These data are obviously not available for a wider audience. Although other researchers may benefit from these data, they are not shared and they may simply be deleted at some stage, since researchers rarely take measures to secure the long-term availability. The situation can be improved by making researchers aware of the potential advantages of sharing data. Such a co-operative environment can only be realised if there are sufficient incentives for researchers to share their data and if there are not too many practical barriers.

Data archives and Trusted Data Repositories
All researchers that have been consulted in this project acknowledged that transferring files to a data archive can produce clear benefits. A very practical advantage would be the fact that this would release them from the burden of having to care for the data themselves. Research data, especially if they consist of video recordings or high resolution images, may require much storage space. Researchers are mostly aware of the fact that Trusted Data Repositories also have a strong expertise in mitigating the risks of digital obsolescence. This is important because, in a number of cases, materials are available only on carriers which are rapidly becoming obsolete, such as VHS, or WordPerfect text files.

Re-use of data
The researchers who were interviewed also acknowledged that when data are archived, this enables researchers to reuse these data in their own projects. If research data, either raw or processed, are made available in open access, the wider research community can then also build on this work. Colleagues who are active in related fields do not need to duplicate work that has already taken place, and, consequently, they can use their limited time more efficiently. When new data can be combined with older data, this also has the effect that larger, more comprehensive analyses can be carried out. It was also mentioned during the interviews that other researchers may reuse data in ways which were not foreseen when they were originally created. If several publications by various authors are based on a single dataset, this implies a very effective usage of this resource. The creation of high-quality datasets are often very costly, and sharing data is thus a way of getting the most value out of them.

Peer review
Borgman (2007) also argues that, “[i]f the data are available, then a more rigorous review of the scholarship becomes possible” (p. 127). When research data are archived systematically, researchers can verify the claims and the conclusions of publications. Paper-based academic texts often make references to research data, but can mostly incorporate these in a condensed form only. Normally, such limitations no longer exist in a digital environment, and when the full data are made available to peers, this will make the scientific process more transparent and more reliable. All respondents agreed that the possibility to verify and to validate the contents of academic texts is a valid reason for archiving research data.

Education
Some respondents had noted that openly available research data may be used for educational purposes. Such resources could be used, for instance, to teach research methods. Tjalsma et al. (2010) also note that research data may be archived in some cases when they are seen as cultural heritage. These resources may be relevant for the study of the history of science (p. 9). However, this particular motivation was not mentioned by any of the researchers who were consulted.

Digital access to cultural heritage
Cultural heritage institutions often hold vast collections of digitized resources, which can become

input to humanistic scholarship. While making these freely available on institutional websites or in special repositories may discourage potential visitors from physically attending the institutions' exhibitions, or from directly compensating cultural heritage institutions for these usually unique and often expensive resources, theoretically there are some motivations for libraries, museums and archives to contribute to the data layer in the humanities. As Borgman argues, these institutions can use the visibility of and free access to their digitized resources to draw wider interest to themselves, their collections and activities, as well as attracting available public and private funds. The success of this strategy will largely depend on the wider interest of the public, i.e., the popularity such collections achieve. This leads Borgman to conclude that, at least as far as cultural heritage institutions are concerned, "the pressure to build the content layer for the humanities comes more from the public than from scholars, but scholarly interest is accelerating" (Borgman, p. 222).

Data Availability Policy

NWO, the most important funding body for scientific research in the Netherlands, increasingly demands that all the results of research that they have supported financially should become available to a wider public. Cinema Context is an NWO-funded project, and in the project proposal the researchers needed to explain how they were going to organise the digital preservation of their data. All data are currently stored in the data repository of the Digital Production Centre at the University of Amsterdam. Some journals have a policy to accept submitted publications only if the supporting datasets are made digitally available as well. Nevertheless, none of the respondents in this project had ever come across such a "Data Availability Policy". Kuula and Borg have found, however, that explicit requirements by funding bodies for the preservation and making available of research data, together with dedicated parts of the funding budget towards that goal, could be one effective way of making scholars mindful of the need for data preservation and sharability (p. 8).

4 Challenges

While researchers generally recognised that there is much value to be gained from sharing data, they also acknowledged that the reuse of data is not yet common. Some exchanges of files have taken place after individual requests, but only incidentally. An important obstacle appears to be the fact that it is currently very difficult to discover what datasets other scholars are actually producing. Researchers may learn about the existence of datasets during conferences, or because they are mentioned in publications, but there is certainly no central registry of relevant datasets. This observation is confirmed by Borgman, who indicated that "[s]earching for the existence of data, types of data, data from a specific collection, or data gathered by a specific research method in publications is difficult because of the lack of consistent bibliographic control" (p. 140). The NARCIS portal,[22] which currently harvests many descriptions of datasets from DANS,[23] was not widely known among the scholars that were interviewed.

Credits
Researchers were asked under what conditions they would be willing to share their data. Respondents indicated that they have no objection to making their supporting materials publicly available, but they added that the data can only be released after they have been discussed in a published text. Some researchers had indicated that more than one publication can be written about a single set of data. In that case, sharing the data needs to wait until all the potential publications have been completed. The background of this caveat is the fact that scholars rarely receive academic rewards for products other than publications. Resources such as databases or models have often required much intellectual effort to produce, but no credits are awarded for these products in themselves. These investments can be justified only if these materials enable researchers to produce books and articles with impact (Rutten and Adema, 2009). This confirms findings by Ballon and Westermann (2006), who explained that, for art historians "scholarship remains wedded to the printed page".[24] Similarly, in the final report of the SCARP project[25] it is explained that the "process of curating data is normally secondary to the cycle of publishing papers and applying for research funding". Furthermore, humanities research is often focused on the development of ideas. The value of humanistic studies is often related to the uniqueness or the originality of those ideas. If annotations are shared at an early stage, there is a risk that other scholars use those ideas in their own texts.

Copyrights
In Media Studies, the focus is often on films, broadcasting or on music. Researchers indicated that using reproductions as part of a publication often involves high costs, since copyright owners often demand high compensations for the use of these protected materials. Both commercial and publicly funded institutions claim compensations for the reuse of the materials they own, even if the reuse takes place in a not-for profit sector such as a scholarly publication. The result is that scholars often refrain from using reproductions in their publications for budgetary reasons.

Privacy
When data are based on surveys, interviews or ethnographical research, there is a risk of violating privacy laws. Surveys often contain data which can be traced back to individuals. The persons who are represented in the dataset must formally agree to make these data available. When there is no explicit permission, there is a serious risk of running into legal problems. Alternatively, researchers may publish datasets in which all personal details have been anonymised.

Standardisation of data
In Chapter 2 of this report, it was explained that researchers often create research annotations which are not particularly structured and which are mostly very subjective and interpretative in nature. It is generally agreed that such initial explorations need to remain private, since the ideas that are laid down in such documents are not yet fully developed. They are usually first steps in an

22. www.narcis.nl
23. www.dans.knaw.nl
24. Ballon, H., & Westermann, M. 2006. Art History and Its Publications in the Electronic Age. Connexions, September 20, 2006. http://cnx.org/content/col10376/1.1/ .
25. www.dcc.ac.uk/projects/scarp

overall evolution towards a more coherent and more consistent argumentation. As such, they will not be very useful to other researchers. Most researchers agreed that only those objects which are relatively 'objective' or 'factual' in nature, or which have reached a certain definitive state can reasonably be shared. Examples of the latter classes of resources include transcriptions, well-structured databases, images or video recordings.

Respondents also explained that there are currently no widely accepted disciplinary norms that govern the structure or the contents of research data. Datasets which are produced are mostly geared towards the needs of specific research questions. Because of a lack of fixed procedures or codes of conduct, researchers have generally used their own insights and their common sense during the design and the organisation of their datasets. Respondents considered it highly unlikely that such quality assurance criteria will be developed in the near future. More so than in other fields, results in the humanities tend to be based around individual researchers and it is unlikely that generic guidelines for the qualitative description of cultural artefacts can ever be devised. Nevertheless, the researchers who had some experience with reusing data stated that they were mostly capable of evaluating the scientific accuracy. Although there are no standardised procedures for checking the validity and the correctness of a database which contains both quantitative and qualitative data, a random sampling of records usually gives a very quick impression of the usability of the resource.

Tacit knowledge

The reuse of data that are produced by colleagues must largely be based on trust. Borgman notes that "[t]rust in other scholars' data is partly a function of whether those data can be interpreted" (p. 130). Reusability can be enhanced if researchers provide documentation and contextual information about the data, explaining how and why they were collected. Some studies (Kuula and Borg 2008; Borgman, 2007) suggest that the researchers who had produced the data are in the best position to provide the necessary context and explanations of their data, in order for subsequent users to be able to work with them and interpret them. Borgman (2007) suggests that, although data are "much less 'self-describing'" than final publications, researchers do not typically engage in the detailed description of the context and precise settings within which data were gathered because these are considered "day-to-day" or "tacit" knowledge within a certain discipline (p. 129). Indeed tacit knowledge, which according to Borgman is one of the primary ways to delineate the contours of a particular discipline, is a major barrier for the reuse of data, especially beyond the original field (Ibid, p. 165). Obtaining details about the origins and surrounding context of the data, as well as the methods used to collect it, is likely to increase trust in the respective dataset (Ibid, p. 131). Kuula and Borg (2008) note that "[w]ithout detailed documentation, data reuse may result in inaccurate, if not downright erroneous, interpretations." (p. 8). Some of the researchers interviewed were careful to document the structure and the contents of their databases so that they can become more accessible to others. Such documentation was sometimes added in the form of annotations which are included in the cells of a spreadsheet. In the case of the *Cinema Context* database, an extensive manual was written for the benefit of other users. A general recommendation for the description of data is given by Marilyn Deegan and Simon Tanner (2002) who suggest that scholars should consider the primary target audience for their data in order to make decisions about the most adequate description of their resources (p. 117). In addition, they recommend the use of controlled vocabularies and thesauri for description wherever possible (p. 143).

Format

DANS has formulated a *Data Seal of Approval* (DSA) which consists of a number of requirements for data and metadata to ensure that "research data can still be processed in the future in a high-quality and reliable manner, without this entailing new thresholds, regulations or high costs".[26] Amongst other thing, the DSA recommends data producers to deposit the digital objects "in a format that is recommended by the data repository". In the case of computer gaming research, the primary data often consists of specific scenarios in game play, and these can mostly be captured and viewed on specific devices only. This particular class of primary sources appears to be difficult to archive. Apart from these gaming data, this recommendation should not produce many difficulties, because respondents explained that they mostly use fairly basic computer applications,

26. www.datasealofapproval.org

such as MS Excel or MS Access. Research data may also be stored in the form of a PDF file, or JPG or TIFF images. All of these file formats are accepted by DANS.

Metadata

When researchers transfer the data from their own working environment to the DANS data archive, they are required to provide “the metadata requested by the data repository”. To the question if metadata were actually assigned during the data creation process, one researcher replied that this is not the case, because she worked with the assumption that she would be the only future user of the data, and she did not have a need for the metadata herself. Unless metadata are generated automatically by software, as is often the case with imaging software or text editors, scholars generally do not explicitly create metadata because there is no need for them. Nevertheless, if a data repository demanded such metadata, researchers would not mind creating these formal description as part of their research activities, provided that it is not too time-consuming.

In summary, the main difficulties in optimising the workflow for data curation are the following:

- Scholars often create unstructured and non-standardised research annotations. Such resources are mostly unfit for reuse.
- A temporal dimension should be taken into account. Researchers are usually not willing to share their data before these data have been discussed in a formal publication.
- Copyright or privacy laws may sometimes complicate open access to research data. Scholars who wish to share their data should make sure that publication does not produce any legal or ethical issues.
- Incidentally, scholars may produce resources which are strongly application-specific, rendering reuse near to impossible.
- Researchers generally did not know exactly how they could archive their data. Most respondents were familiar with DANS, but had never actually deposited their own datasets in their archive.
- Usually, researchers do not assign metadata to the resources they produce.
- There is a lack of a clear mandate. Universities, publishers or funding agents do not actively stimulate scholars to share their data.
- Researchers are often unaware of the existence of relevant datasets. There is no central registry of primary resources.

5 Workflow

5.1 Data curation lifecycle

Data curation focuses on “maintaining, preserving and adding value to digital research data throughout its lifecycle”.[27] Unfortunately, the research data which are created at present often face a highly uncertain future. Various agents offer some services in the overall workflow for data curation, but tasks and responsibilities have not been clearly assigned yet. To improve the present situation, tasks and responsibilities have to be identified, defined and assigned to specific parties and actors in the production chain of scholarly information. This section will propose a workflow for data curation which is largely based on the Curation Lifecycle Model developed by the Data Curation Centre in the United Kingdom.[28] It provides a "graphical high-level overview of the lifecycle stages required for successful curation" (Higgens, 2008, p. 135). Figure 3 is a slight adaptation of the original model. In this diagram, an attempt has also been made to indicate the agents responsible for the various activities which are mentioned.

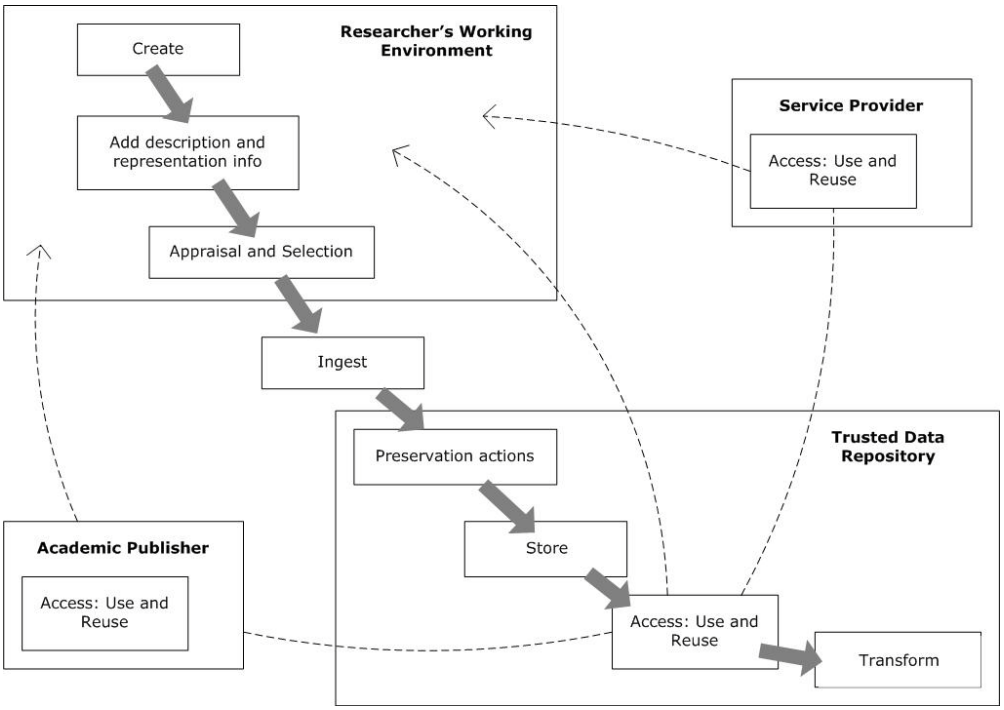

Figure 3. Curation Lifecycle.

From active working environment to trusted data repository
In practical terms, data curation usually involves transferring research data from the active working environment of researchers to an institution which can be acknowledged as a Trusted Data Repository (TDR). The storage and preservation of research data require considerable resources, normally beyond the capacities of individual researchers or research groups. These tasks are delegated to institutions who have the relevant expertise in this field and who hold the technical infrastructure for the preservation of digital resources and for technical and bibliographical metadata. The storage and preservation of research data directly connects to similar tasks these institutions have been performing in the past; it complies with their mission and remit. Kuula and Borg (2008) conclude that “the safest solution is to let experts take care of preservation” and that

27. http://www.dcc.ac.uk/digital-curation/what-digital-curation
28. www.dcc.ac.uk

"[c]entralised archiving is the best way to ensure that research data are documented according to national and international standards" (p. 26). An additional benefit of delegating the curation of data objects to centralised repositories is the common practice of such institutions to produce various materials describing or systematising data in the archive (such as bibliographies, databases, etc.), which may also facilitate reuse.

Create and add description and representation info

The lifecycle obviously starts when researchers create research data, as part of a specific research project. As was established above, this may include a wide range of data types and formats, such as images, databases, and audio or video files. Since high-quality metadata are indispensible for an effective curation of the data, it is highly recommended to ensure that metadata are assigned during or shortly after the creation of the files. This report recommends a workflow in which researchers provide the required metadata themselves. Borgman (2007) argues that "those who collected the data are in the best position to interpret them, as they understand the context for the questions asked and the decisions made in the collection and analysis processes" (p. 128). In the Social Sciences, the schema provided by the *Data Documentation Initiative*[29] is widely used to capture the details of research data. There appears to be no similar format for datasets within the arts disciplines. Nevertheless, it can be questioned if a similarly detailed format is required for the type of data that are produced by humanists. The main purpose of metadata is to enable research to find the objects, to understand them and to reuse them. On a minimal level, the following aspects appear to be essential:

- Intellectual ownership: the researchers who have created the resource
- The date of creation
- The date of the last modification
- The name of the research project in which the resource was created
- A brief description of the research project and the way in which the data were created

If necessary, references to codebooks or manuals may be provided as well. The compulsory metadata fields must be negotiated with the projected data repository. To reduce the risk that metadata assignment will ultimately be viewed as an unnecessary bureaucratic hurdle, a fairly simple metadata format such as Dublin Core[30] should be used. Recording metadata usually involves the creation of a simple text file or a spreadsheet is created in which the required aspects are captured.

Appraissal and Selection

In the workflow for data curation, it is important to ensure that relevant decisions are made right from the start of the data lifecycle. If a decision to preserve data is not taken until the end of the research project, it is likely that the researcher or the repository must invest much effort in cleaning data, reconstructing work processes and transforming data formats. Conversely, if researchers know from the onset of the project that data are ultimately going to be archived, they are likely to treat their research data and their documentation more carefully. The various objects which are created need to be evaluated, and a decision must be taken on which objects are worth preserving. Such decisions are sometimes bound by legal obligations, but if such regulations are missing, a selection must be made by the researchers themselves. Citing Kenney and Rieger (2000), Deegan and Tanner call for investing "wisely in the selection and creation of digital resources that are likely to be used and reused over time. Without this basic first step of good selection, the rest of the process will never appear valuable or affordable." (Deegan and Tanner, p. 98). Markku Leppänen (2006) provides a list of indicators that might be used as general guidelines in the process of selection for preservation: "usability and conditions of access, level of uniqueness, social, cultural and scientific value, and cost of preparation for archiving" (cited in Kuula and Borg, p. 8). Lastly, Christine Borgman also offers a rather broad selection criterion based on "the likelihood of reuse," stating that if the expected audience for the data is large enough, this should justify investment in the documentation and preservation of these data (Borgman, p. 141). Following recommendations made during the interviews with researchers, it will be argued that curation efforts must focus primarily on objects that have a certain 'factual' or 'encyclopaedic'

29. www.ddialliance.org
30. http://dublincore.org

quality or on objects that have a clear structure. Such objects could be images, transcriptions of interviews, databases or spreadsheets. Similarly, digital reproductions of primary sources will also be relevant to other scholars.

Ingest
Next, the selected data need to be ingested into the data archive. The infrastructure for the deposit of research data with a data centre should have a front office which is integrated as seamlessly as possible into the working environment of the researchers. Ideally, researchers should be provided with a button which enables them to “save into the data archive”. It is also important to ensure that, at the moment of transfer, researchers can specify the manner in which objects will be available. If there are certain legal or other conditions which warrant a degree of confidentially, there must be a possibility to enable a semi-open access or a clearly specified embargo period. DANS EASY is a lightweight deposit application which can serve as a good example of how data transfer to a data repository may take place.[31]

Preservation actions
After the ingest into the TDR, the maintenance of the dataset is no longer a concern of the researcher. This responsibility shifts to professional data archivists. Once the data have been received by the TDR, specific measures ought to be taken to ensure the integrity and the authenticity of the data in the long term. More concretely, this can involve activities such as "data cleaning, validation, assigning preservation metadata, assigning representation information and ensuring acceptable data structures or file formats" (Higgens, 2008, p. 138).

Access: Use and Reuse
Subsequently, the data need to be stored and can be made accessible to others. If, after a longer period, the technical format of the dataset becomes obsolete, it may need to be transformed into another format so that its contents can still be accessible.

Transform
Higgens et al. (2008) assume that there may not always be a need to store research data indefinitely. Under certain circumstances, the data may also be deleted from the archive. Doek et al. (2009) propose that the decision to continue preservation efforts should be based on the usage and the appreciation of the research data. After a period of ten years, a timeframe which was chosen somewhat arbitrarily, this usage should be evaluated by considering the number of publications that are based on the dataset. This selection method evidently assumes that it is possible to count the number of citations, and this is currently still difficult. Metrics for the evaluation of the quality and the impact of academic publications are complicated for a broad number of reasons,[32] and the evaluation of the impact of datasets is likely to introduce a number of additional complications of its own. Usage statistics appear to be a more reliable method. When datasets have not been used for ten years, they may be moved or transferred to a difference storage facility which is less costly, such as storage on tape or DVD.

5.2 Role of institutions

The number of institutions that can actually function as trusted data repositories for research data produced in the Netherlands appear to be limited. It is not likely that the data centre that is implemented by the three technical universities[33] will accept datasets which are produced by humanities scholars. The *Netherlands Institute for Art History* aims to archive databases of art historical interest, but at the moment this institution does not have the capacity to accept all the digital research data that are generated in the Netherlands. It seems logical to assume that Dutch universities should take care to curate the data that are created by their own academic staff. Nevertheless, within most universities, the expertise and the resources necessary to set up such dedicated data centres are lacking. A large number of universities manage image repositories to provide access to digital reproductions of their special collections. When scholars have digitised

31. http://easy.dans.knaw.nl/dms
32. A number of these are discussed in Johan Bollen, “Usage Impact Factor: the effects of sample characteristics on usage-based impact metrics”. See also Adema and Rutten (2010)
33. http://datacentrum.3tu.nl/240/?L=1

primary or secondary sources which are held by a specific institution, these reproductions may be transferred to this institution's image repository. Nevertheless, since the images in such repositories are often subjected to high quality standards, the scans produced by researchers can not always be accepted.

National Library (KB)
In 2010, the Nationale Coalitie voor Digitale Duurzaamheid[34] has published the report *Toekomst voor ons digitaal geheugen: strategische agenda voor duurzame toegankelijkheid* (NCDD, 2010). Amongst other things, this report proposes a division of responsibilities. It is explained that the National Library (KB) in The Hague will focus primarily on sustainable access to textual publications, and that DANS will be the most evident destination for research data for which no other (subject-based) repositories are available and for those collections which have a strong national character. To a certain degree, this division is somewhat artificial, because the e-Depot of the KB[35] clearly does not only contain publications. Firstly, the e-Depot also contains many master images from the KB's own digitisation projects. In addition, there are also many datasets that come along (uncalled for) with publications. The current e-Depot was built in 2003, and the requirements for a new e-Depot are being drawn up at this moment. The new system should be operational in 2013. Whereas the bulk of the materials is flat, it is anticipated that, partly due to further technical developments, more datasets will need to be archived. The KB currently investigates what kind of e-Depot would be suitable for this purpose. It is clear that digital data archiving calls for a continuous process of technical innovation.

Central portal
Before researchers can reuse the data of others, they must evidently be aware of their existence first. Many respondents have signalled that it is often difficult to discover what other datasets colleagues have developed. Data archives such as DANS can improve this situation by providing web interfaces to their archives with good browsing facilities and effective search options, so that the scholarly community can find the relevant datasets. If different institutions will be responsible for the curation of data, it is also important to create a central portal through which researchers can discover relevant datasets. Service providers such as NARCIS or DRIVER[36] may ultimately have an important role to play in allowing researchers to discover data. Such a central provision may harvest descriptions from various archives. This will not solve the problem entirely, as there are currently many datasets which are not yet curated in an official data centre.

Academic publishers
It is clear that academic publishers can also have a function in the dissemination of datasets. In 2009, Amsterdam University Press (AUP) has set up the open access Journal for Archaeology in the Low Countries.[37] An interesting feature of this open access journal is that its articles are linked directly to the datasets, the images and the GIS data that have been used during the research. These supporting materials are derived from the *e-Depot voor de Nederlandse Archeologie*.[38] AUP's initiative to publish the final products of the scholarly process, the articles, in conjunction with resources that have been produced at earlier stages was largely the outcome of the Call for Tenders 2008 in which SURFshare encouraged institutions to experiment with such so-called enhanced publications.[39] Currently, AUP is also developing an enhanced Media Studies Journal based on the same format. When scholarly articles are enhanced, these publications can function as effective signposts to the datasets they are based on. This will clearly improve the visibility and the awareness of these datasets. The scholarly article may be seen as a documentation to the dataset as well, as the author may provide much contextual information. Instructions for academic authors to include citations to on-line databases if they have used them may also help to raise awareness of these resources.

Doek et al. (2009) also suggest that academic publishers may assist scholarly authors in the selection and the certification of data. Certification "establishes the validity of a registered scholarly

34. www.ncdd.nl
35. www.kb.nl/hrd/dd/index-en.html
36. http://search.driver.research-infrastructures.eu/
37. www.jalc.nl
38. www.edna.nl
39. www.surffoundation.nl/verrijktepublicaties

claim"[40] Similar to the way in which publishers currently organise peer review for publications, it has been suggested that they facilitate the academic evaluation of research data, especially if these data form the basis of the articles that they disseminate. The academic publishers that have been interviewed for this study were found to be sceptic about this potential responsibility. Publishers are mainly concerned with the academic results that offer a synthesis of the raw data. The data are primarily seen as an initial stage in a process that must ultimately lead to the publication of an article or a monograph. Peer review of data is not a task that publishers seem to be willing to take on, as this is considered to be the responsibility of the academic community. It is probable that the act of making data openly available also urges the creator of these data to certify that these data are free of errors. Authors know that if colleagues discover serious faults of even conscious falsification this will damage their reputation immensely. This effect is also visible very clearly in the ArXiv repository.[41] Although there is no formal peer review of the texts that are submitted to ArXiv, authors go to great lengths to secure the quality of their contributions because they are fully aware of the fact that these will be read by the most prominent researchers in their discipline. It can be expected that sharing data online will imply a degree of self-correction and that it will discipline researchers to ensure the academic quality of their resources.

Academic Institutions

Although academic institutions are unlikely to realise facilities for the curation of data themselves, they do have a number of responsibilities with respect to research data of their academic staff. The ability to curate data crucially depends on the presence of high quality metadata and the use of the appropriate technical formats. Optimising metadata standards and file formats often requires specific expertise which is often lacking among researchers. A commitment to the goals which were laid down in the Berlin Declaration on Open Access,[42] which explicitly underscored the importance of open access to research data, also implies a need to ensure that the researchers who actually create the data are fully aware of the conditions which are needed to realise such open access. Research directors need to provide a sufficient amount of training to their academic staff. Library directors may also want to train library staff so that they can effectively assist researchers during the data curation process. In the SCARP report, it is argued that such data experts should have a remit "to reach out to and develop close ties with one or two disciplines or departments" and they need "domain-specific, task-specific and cross-cutting, general data curation training and support". Ideally, university libraries should train "data stewards". The traditional role of librarians as curators of authentic documents might be employed to serve the need for quality assessment of deposited datasets as well. One researcher strongly emphasised that the context of research databases and other sources should be assessed by a curator, whose bird's-eye view can easily identify links between different objects and reasons for inclusion or exclusion of specific datasets. As Deegan and Tanner (2009) point out, "curators are trained to know what they have and what its status is" thanks to the network of experts they have access to and the "bodies of meta-information (catalogues, bibliographies, etc.)" they can refer to (p. 185). Some of the ways in which data curators may help with quality appraisal is by guaranteeing the immutability of deposited objects and/or by keeping track of any changes or versioning they undergo (ibid, p. 186).

40. www.dlib.org/dlib/september04/vandesompel/09vandesompel.html
41. http://arxiv.org
42. http://oa.mpg.de/openaccess-berlin/berlindeclaration.html

6 Conclusions

6.1 Future implications

Attitudes towards data sharing

This report has explored requirements for the curation of the research data that are produced by researchers who focus on the arts and the media. The researchers that were consulted have mentioned various factors that affect their willingness to share their data. Many of the challenges that were explained are not unique to this particular discipline. Problems related to metadata formats, or to the discovery of datasets are probably equally pressing in other fields. The view that datasets cannot be shared before the results have been consolidated in a traditional publication is likely to be found among scholars across the entire academic spectrum. A feature which distinguishes arts and media research from other disciplines is that it is strongly focused on physical or born-digital primary sources held by cultural heritage institutions. There is a certain dependence on institutions such as museums, archives and libraries, who should ensure that they can offer their holdings in a form that is actually usable by researchers.

Reuse of external data is still rather limited at the moment. When researchers make use of artefacts produced by others, this usually involves digital reproductions produced by cultural heritage institutions. An additional data layer, consisting of primary resources created of other scholars is virtually non-existent. Nevertheless, the attitude towards data sharing was found to be positive. Respondents recognised that resources which are archived systematically can be reused by others, and that such resources can also be used to verify the conclusions of scholarly publications. A number of researchers also suggested that there is a tendency towards more openness in scientific practices in general. Previous generations of researchers often used to be very reluctant to make their data available to a wider public. They viewed the data as their own personal property, and kept their data to themselves. Newer generations of researchers increasingly realise that data can be relevant for others and take a more open attitude towards reuse. Today's researchers expect to find their information online, are used to combining different digital resources in the explorative and research phase of a scholarly project. On average, they are more inclined to share their data with colleagues. Since researchers generally recognise the benefits of data reuse, it is likely that, given time and adequate support, more and more datasets will ultimately become available online. Firstly, a critical mass needs to be achieved before shared datasets can actually prove their value. Well-defined discovery tools and more references from publications to digital data are also needed to draw attention to these resources and to ensure that sharing and re-using datasets can become more commonplace.

Analogue and digital research data

It must also be recognised that, at present, the majority of the primary resources that are created are still analogue in nature. Arts and media research mostly involves critical analyses or close readings of texts or audiovisual productions. The results of these activities are mostly recorded on paper only, and the general workflow is usually not supported by digital research instruments, such as, for instance, virtual research environments.[43] Nevertheless, the curation of data produced in research that takes place in a more conventional mode is as important for the field as digital humanities research. Making these resources fit for curation demands more time from the researchers involved and more effort from the curating institutions. Such resources need to be digitised, and, in some cases, also standardised. At the same time, this study concluded that rewards for these investments are absent at the moment. There may be budgetary reasons to focus data curation investments predominantly on the digitally born data. Nevertheless, from the general purpose serving the development of the humanities a sole focus on curating digital born data would be a mistake, as this would increase the gap between those projects that make use of digital research tools on the one hand, and the more traditional projects that don't on the other. It seems reasonable and worthwhile to digitise the analogue data as well. As a part of this endeavour, research can be done on how to further integrate new tools and instruments in more traditional research modes, making them more effective and, in the end, facilitating the curation process.

43. On the topic of VREs, see, for instance, Olson et al. (2008).

New possibilities

A growing willingness to share research data, combined with an increased use of digital instruments, will inevitably have strong repercussions for the nature of humanities research. When academic research is supported by sophisticated digital instruments, this will also lead to an emphasis on aspects that can be quantified, both instead of and in addition to qualitative research. It will also open up possibilities to pose and answer new research questions, departing from more traditional questions within the humanities. There will be a possibility to answer questions that could simply not be asked in an analogue research environment. In addition to more extensive and efficient analyses of primary data, it also becomes feasible to collect information on user behaviour in digital contexts. Registering human-computer interactions during, for instance, the consultation of digitised cultural heritage collections opens up possibilities to explore user behaviour on a scale and with an intensity that is new to humanities, demanding new skills in terms of observation and analysis. Proper data curation is a necessary step to realise the full potential of these innovations in humanities research workflow as well as, on a more general level, in the humanities as a scholarly discipline.

Standardisation

Research based on quantities of data and carried out with the aid of digital laboratories will lead to a degree of standardisation of technical formats. The vast majority of digital humanities projects produce XML documents, relational databases, images, audiovisual recordings, or combinations of these. The semantic contents of these files, however, may vary considerably. As was established before, research in the humanities is often centred on individual scholars, and, consequently, their products often reflect idiosyncratic methodologies and practices. It can be observed that the current scholarly domain is characterised by a great diversification of terminologies, tools and instruments. Before data can be reused within other projects, the scholarly community may need to take some measures to standardise terminologies, ontologies and practices. However, during the discussions with researchers, it became clear that such standardisation is not feasible nor desirable. Humanities research usually focus on the myriad of ways in which human beings have expressed themselves, and the interpretation of such highly diverse artistic utterances is not likely to be standardised, even if such interpretations and analyses are guided by digital research tools. Clearly, convincing scholars to archive and to share their primary resources is a first important step, but securing the semantic or intellectual interoperability of the various resources which are shared will clearly be a next major challenge.

Basic obstacles

In this project, three basic obstacles towards data archiving were found. Firstly, it is clear that there is currently no consistent or co-ordinated approach to curating the research data that are produced by scholars who focus on the arts. A number of researchers have come up with solutions to archive the data produced within their own projects, and in some cases, institutions have made ‘ad hoc’ arrangements for specific projects. A central vision is clearly lacking. All-in-all researchers omit taking measures to archive their data, because it is not demanded by their faculties nor by funding bodies. Other incentives or rewards for curating data and making them available for reuse are lacking. A second difficulty is the fact that there appears to be some confusion about the practical aspects of data archiving. Researchers generally do not know which institutions to contact, which data formats to create, and which metadata to provide. There also appears to be a degree of confusion among data centres themselves. It is not always clear which objects need to be archived by which institution. The NCDD is currently addressing this issue, by producing a strategic vision for data curation in the Netherlands, but much work is yet to be done on clearly defining roles and responsibilities. Thirdly, researchers currently have very limited means to discover other datasets. Before data can be reused, researchers must have an opportunity to learn about their existence. NARCIS, the national research portal is not widely used for this purpose at the moment. It is likely, however, that an improved accessibility of datasets will stimulate the reuse of these materials and that this will also encourage the community’s penchant to share their own data.

Acceptance of digital research data

The lack of a sense of urgency and the absence of formal obligations to archive data is largely caused by the fact that scholarly certification primarily takes place on the basis of publications. No direct rewards are given for resources that are produced at earlier stages of the research process,

such as databases or data visualisations. Even within the scholarly community, the preliminary results of research projects that make use of digital data and research tools are currently not regarded as highly as traditional publications. As research on the potential role of digital publications vis-à-vis printed books shows, there is a big difference in valuation of new modes of publications, being largely accepted in subfields as for instance media studies and linguistics, whereas in other domains print still has more prestige and print authors enjoy far more rewards and recognition than those publishing eBooks. There are strong indications that the difference in appreciation correlates with generations, implying that the shift may be completed within a few years.[44] In this context, it should be born in mind that the aforementioned new forms of data gathering and analysis are performed in addition to the more traditional and qualitative interpretative and ethnographic types of research that have been going on for decades in the humanities and that will continue in the years to come. One challenging issue is how developments towards digital humanities will influence the dominant mode of research within the field. Will the traditional humanities change as a result of digitisation or do we witness a growing gap between scholars relying on digital sources, resources, tools and instruments and those working in more conventional modes? That the present development or shift will lead to broader debates and disputes within the field is certain; how it will develop is not.

Quality and impact

At the moment, quality and impact is primarily measured on the basis of data provided by ISI/Web of Science. It is not probable that the parties that are involved in the assessment of research will start to recognise the value of research data in the near future. Nevertheless, it can be expected that, when universities and science foundations such as NWO formulate policies of mandatory curation and open access publishing of datasets, the importance of research data in the overall assessment of scholarly impact is likely to expand. Archiving data in trusted data repositories, or incorporating datasets in enhanced publications, leads to a new approach in evaluating the quality of scholarship. Peer review processes and other quality assessment procedures will no longer be based solely on the written record of the research in article or a book, but also on the review of the data, providing transparency in the process and the management of the workflow. Quality thus becomes both a characteristic of both the product and the process. This may also introduce a certain risk. When research methods and knowledge production become fully transparent, this can result in criticism regarding the sholar's work, thereby creating disincentives for data sharing. The wider implications of this development remain to be explored. However, it is clear that this will result in a redefinition of the professional role of the researcher. It will have far-reaching repercussions on the distribution of reputation and reward within the system of science, in which the key role of quality is increasing. Specifically for the humanities, this poses a clear challenge, since in comparison to the field of Science, Technology and Medicine, performance criteria are relatively underdeveloped, whereas the funding structure of the field is more in need of them as ever.[45] In short, it must be recognised that data curation will also have considerable political implications for the valuation of humanities scholarship in the future. It is likely to result in a new conception of scholarly quality and its evaluation, both within the process of scholarly publishing and in the assessments of individual scholars, research teams and faculties.

6.2 Recommendations

This section provides recommendations for the various stakeholder groups that can take actions to improve the current situation.

- **Data curation centres** and **research communities** need to agree on the metadata schema that should be assigned to research data. It is advisable to work with a relatively simple set of fields, since the creation of metadata would otherwise place too many strains on the research process.

44. Janneke Adema en Paul Rutten (2010). Digital Monographs in the Humanities and Social Sciences: Report on User Needs.
45. This can be illustrated by the fact that KNAW has established a committee to advise on quality indicators in the humanities. The committee will report in the fall of 2010.

- At the moment, researchers do not always know where to deposit their data. The field currently appears to be rather fragmented. It would be very helpful if **data curation centres** would co-operate and would invest in the development of a single portal for depositing research data. The various institutions that can offer services in the field of data curation could then agree behind the scenes on who stores what.
- In addition to a single facility for *depositing* research data, a similar centralised facility for the discovery of datasets may need to be supplied. A central registry which is actively promoted among the scholarly community will stimulate the reuse of primary resources.
- **Funding bodies** should consistently stipulate that if they have supported the creation of research data these data need to be preserved and published in open access.
- In addition, **universities** should stimulate their academic staff to archive their datasets, especially in cases where researchers have directly based publications on these datasets. Universities should also ensure that technical guidance in the field of data curation can be sought within the institution, for instance, by training library staff.
- **University libraries** and organisations that support academic research, such as **SURFfoundation**, should encourage researchers to make use of existing ontologies and thesauri. Terms and names provided by RKDartists, WorldCat Identities, GeoNames, DBPedia, and GeoNames can be very helpful in humanities research, but very few researchers actually use these. Conversely, researchers who, as part of their research, standardise terminology or formulate classification lists should be encouraged to publish these vocabularies in such a way that these terms can be reused. Humanities scholars need to be shown how they can contribute to the open linked data network.
- Experiments of **academic publishers** with enhanced publications must be stimulated, since this will raise the awareness of the benefits of the curation of digital research data. Academic publishers should also encourage their authors to include references to data-sets, if applicable.
- **Cultural heritage institutions** should ensure that they use the same protocols and the same standards that are in use in the scientific repository community. The most important requirement is that it must be possible to reference digitised objects on the basis of persistent identifiers. This provision is needed to ensure that academic publishers or other agents can enhance publications about these objects with direct links to these objects.
- The issue of data curation should be coordinated centrally as much as possible, for instance by the ***Nationale Coalitie Digitale Duurzaamheid*** (NCDD).
- A strategy of data curation should involve both digitally born data as well as those resulting from more traditional modes of research to avoid a gap between the two practices and to explore the application of new tools and instruments in the conventional research practice, making it more efficient and fit for curation.
- There is a strong need for research on the implication of data-curation and reuse for the future valuation of humanities research. This study should explore how different notions of quality (process and product) will evolve and how these changes will affect the professional profile of scholarship in the humanities and the future distribution of research funds, using quality as a central criterion.

Bibliography

Adema, Janneke, “JALC User Needs: External Evaluation Report”, 2009. Available at: <https://www.surfgroepen.nl/sites/JALCproject/Project%20results/WP8%20External%20evaluation%20report.pdf>.

Adema, Janneke and Rutten, Paul, *Digital Monographs in the Humanities and Social Sciences: Report on User Needs*, OAPEN, January 2010. Available at: www.oapen.org/images/D315%20User%20Needs%20Report.pdf.

Angevaare, Inge, *A future for our digital memory permanent access to information in the Netherlands*, Netherlands Coalition for Digital Preservation: 2009. Availbale at: <http://www.ncdd.nl/en/documents/Englishsummary.pdf >

Borgman, Christine, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*, (The MIT Press: 2007).

Borgman, Christine, “What Can Studies of e-Learning Teach Us about Collaboration in e-Research? Some Findings from Digital Library Studies”, *Computer Supported Cooperative Work* 15.4 (2006), p. 359-383.

Brockman, William S. (et al.), “Scholarly Work in the Humanities and the Evolving Information Environment”, Report of the Digital Library Federation and the Council of Library and Information Resources, 2001. Available at: <http://www.clir.org/pubs/reports/pub104/contents.html>

Coles, Simon J. et al., “Enabling the reusability of scientific data: Experiences with designing an open access infrastructure for sharing datasets”, *Designing for Usability in e-Science*, Edinburgh, UK 26 - 27 Jan 2006. Southampton, UK, 5pp.

Deegan, Marilyn and Tanner, Simon, Digital futures: strategies for the information age, Library Association Publishing (London: 2002).

Doek, Afelonne et al., *IISH Guidelines for Preserving Research Data; A Framewok for Preserving Collaborative Data*. International Institute of Social History, Amsterdam, December 2009.

Graaf, Maurits van der, *The European Repository Landscape 2008 : Inventory of Digital Repositories for Research Output.* Amsterdam University Press: 2008.

Greenhalgh, Michael, “Art History” in *A Companion to Digital Humanities* (Susan Schreibman, ed.), Blackwell Publishing: 2004.

Harley, Diane (et al.), “Assessing the Future Landscape of Scholarly Communication: An Exploration of Faculty Values and Needs in Seven Disciplines”, Center for Studies in Higher Education, UC Berkeley, 2010. Available at: <http://escholarship.org/uc/cshe_fsc>

Harley, Diane, Henke, Jonathan, Lawrence, Shannon, Miller, Ian, Perciali, Irene, & Nasatir, David, “Use and Users of Digital Resources: A Focus on Undergraduate Education in the *Humanities and* Social Sciences”, Center for Studies in Higher Education, UC Berkeley, 2006. Available at: <http://www.escholarship.org/uc/item/8c43w24h>

Hey, Tony, et al., *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research: 2009.

Higgens, Sarah, “The DCC Curation Lifecycle Model”, *The International Journal of Digital Curation* Issue 1, Volume 3, 2008, pp. 134-140.

Kenny, A.R. and Rieger, O.Y. (eds), Moving theory into practice: digital imaging for libraries and archives, Research Libraries Group, 2000.

Key Perspectives. (2010), "Data dimensions: disciplinary differences in research data sharing, reuse and long term viability. SCARP Synthesis Study", Digital Curation Centre. Available at: <http://www.dcc.ac.uk/scarp>

Kuula, Arja and Borg, Sami, “Open Access to and Reuse of Research Data – The State of the Art in Finland”, Finnish Social Science Data Archive (FSD), University of Tampere, 2008. Available at: <http://www.fsd.uta.fi/julkaisut/julkaisusarja/FSDjs07_OECD_en.pdf>

Leppännen, Markku, “Miten tutkimusaineistojen säilytysarvo tulisi määritellä? Esitelmä Arkistoyhdistyksen syysseminaarissa” 3.11.2006 Tieteiden talo, Helsinki, 2006.

Netherlands Coalition for Digital Preservation (NCDD), *A future for our digital memory (1): permanent access to information in the Netherlands*, Netherlands Coalition for Digital Preservation: 2009. English-language summary available at: http://www.ncdd.nl/en/documents/Englishsummary.pdf ;

Netherlands Coalition for Digital Preservation (NCDD), *A future for our digital memory (2): strategic agenda for long-term access to digital resources*, 2010, English translation at http://www.ncdd.nl/en/documents/10-13strategicagendaNCDD_EN.pdf

Olson, Gary M., Nathan Bos and Ann s. Zimmerman, Ann S. (2008). “Introduction”, in: Gary M. Olson, Nathan Bos and Ann S. Zimmerman (eds), *Scientific Collaboration on the Internet*. Cambridge, Massachusetts: MIT Press, pp. 1-12.

Rockwell, Geoffrey and Mactavish, Andrew, “Multimedia” in *A Companion to Digital Humanities* (Susan Schreibman, ed.), Blackwell Publishing: 2004.

Tjalsma, Heiko et al., Selection of research data: A report by DANS and 3TU Data Centre. Commissioned by SURF Foundation - SURFShare program. April 2010.

Verhaar, Peter, “Enhanced Publications: Object Models and Functionalities”, DRIVER Digital Repository Infrastructure Vision for European Research II, Deliverable D4.2, 2008. Available at: <http://www.driver-repository.eu/component/option,com_jdownloads/Itemid,58/task,view.download/cid,54/>

Welshons, Marlo (ed.), “Our Cultural Commonwealth: The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences”, American Council of Learned Societies, 2006. Available at: <http://www.acls.org/cyberinfrastructure/ourculturalcommonwealth.pdf>

Appendix A. Summary of interviews

Connie Veugen, Faculty of Arts, VU University Amsterdam, 1 June
Connie Veugen investigates the ways in which computer adventure games differ from other media. In her research, she aims to build a theoretical framework for the description of video games and game play. Her databases contain many text fields which are used to classify specific findings. Veugen also records specific scenarios of game play. These data are currently application-specific. At the moment, she stores all the data herself, on DVDs, on the University network, or on her own computer at home.

Veugen explains that the reuse of the data of other researchers is rare. Her experience is that the databases of other researchers tend to be designed for the purpose of a specific research question. Data is usually shared only among researchers who know each other very well. Veugen states that she personally has no objections to share her data, on the condition that this takes place after she has written a publication about these data. She is convinced that the reuse of data could save work. She also clearly sees the importance of metadata and would not object if making such formal descriptions would become mandatory for preservation purposes. Veugen's database contains a vocabulary that is used to classify the data. It would be useful if the discipline could agree on a fixed vocabulary. Veugen also tries to annotate her databases as much as possible by including comments.

In general, Art Historians and researchers in the Comparative Arts appear to be reluctant to make their data available to a wider public. People keep their data to themselves and are usually not concerned with digital preservation. Veugen also mentions a very practical problem. Audiovisual archives such as the Academia service of SURFnet generally limit themselves to national broadcasts only. Similarly, the programs which are offered by foreign equivalents of "uitzending gemist" are not accessible outside those particular countries.

Karel Dibbets, Media Studies, University of Amsterdam, 9 June
Prof. Dibbets is editor of the Cinema Context Project (www.cinemacontext.nl). This project investigates the reception of films in the Netherlands. One of the most important results of this project is the Cinema Context database. The encyclopaedia-style presentation encourages data analyses by third parties. The strength of the database lies in the consensus that it is the one and only source for information on this topic. Cinema Context explicitly does not duplicate information, but tries to link to other systems whenever possible. According to Dibbets, museums and archives in particular are more prone to aggregation of data into their own systems than to integration with external databases. If this is the case, the duplicated database will not be maintained. The maintenance of datasets, but also the archiving of audiovisual media can clearly be better regulated. In this context, Dibbets emphasised the need to establish a digital curator.[46]

The fact that NWO demanded that the datasets were made fully public was not problematic for Dibbets. Nevertheless, the attitude of Dibbets it is not representative. Many researchers prefer keeping their data to themselves. Dibbets further states that for an organization like SURF it is important to look beyond the abstract level when it comes to providing grants and building infrastructure. For research fields that remain under the radar, assistance is often not available immediately.

Henk Borgdorff, Amsterdam School of Arts, University of Amsterdam, University of Gothenurg, Royal Academy of Art, 14 June
Henk Borgdorff studies the field of artistic research, in the way that it is practiced by artists. From that perspective, art is not necessarily considered as an expressive practice, but also as a field of scholarly contemplation and research. To meet the academic criteria, a space for discourse with possibilities for scientific positioning towards non-scientific expressions should be created.

46. See: Dibbets, Karel. "Op zoek naar een digitale conservator", in: Bert Hogenkamp and Mieke Lauwers red., Jaarboek Stichting Archief Publicaties, deel 5: Audiovisueel (The Hague: SAP, 2006), p. 189-197.

Henk Borgdorff is engaged in a number of fronts relevant to SURF. Firstly, there is the SHARE-project. Secondly, there is plan to set up a the Journal of Artistic Research (JAR). This should become a DILPS-resembling[47] multimedial platform that serves the rights and needs of the artist. JAR's capacity to meet research requirements will be supported by a research catalogue.[48] The challenge of constructing such a heterogonous database lies in the creation of a certain depth, the ongoing added value of such a database should therefore be in its dynamic character.

In order to accomplish this challenge, the volume of the collection should decide over its functioning. Furthermore, the database should provide a clear and transparent field of research. This can diminish the gap between the artifact and its re-usability, a perpetual feature in the arts. As a whole, this concept of data-curation can be interpreted as a paradigm. All these ambitions assume a substantial user friendliness, as well as a quality standard of the data in de repository, with a selection solely on the basis of a system of clear academic peer review, although resulting in a hurdle of artistic arguments that are multi-interpretable. The threshold may therefore be low, but this large project might form the soil for the emergence of a large and diverse database.

William Uricchio, Professor of Media History at Universiteit Utrecht, Comparative Media Studies at the Massachusetts Institute for Technology, Boston University, 14 June
The interview with William Uricchio dealt with the consequences of the advent of the so called Digital Humanities for the future of data reuse. Uricchio compared the situation in the United States to that in Europe. Uricchio observed that in many fields within the humanities there is no tradition of producing data based on observation of so-called original works or objects. In that sense the concept of the curation of data-collections is not valid. In those cases were humanities scholars do reuse existing data, these data take different shapes and sizes demanding a differentiated approaches.

To illustrate his argument, he indicates that the legal context of research in Europe and USA are different, especially when it comes to the reuse of data. The US is a litigious country. The potential violations of both IP and privacy laws can have severe consequences for those accused of it. Therefore, scholars refrain from putting up their data using somebody else's data. In Europe the climate is still different, but this may change. Another obstacle in the United States is that, due to high compensation demands, using material in publications is increasingly out of the question for budgetary reasons. An interesting and developing field is the so-called Digital Humanities. Here data gathering and reuse of data is core. The development of tools and instruments for future research are the main goals of most of the projects in this field.

Julia Noordegraaf (Media Studies, University of Amsterdam), 22 June
Julia Noordegraaf currently works in a research project which investigates the implications of the transformation from analogue primary sources into digital sources. What sort of knowledge transformation takes place if objects are taken out of their original context, and if they are incorporated into an on-line archive? What kind of implications does the carrier have for the content?

Humanistic research usually consists of consulting resources and making annotations of these resources. Often, such annotations cannot be made public, because they have value only to the researcher who makes them. Nevertheless, there are a number of artifacts such as the transcription of an interview, which have a more 'factual' nature and which are clearly of value to other researchers. Noordegraaf also thinks that archiving collections of hyperlinks may be useful. The curation of research data may be relevant because it allows others to evaluate the validity of an argumentation. It makes the scholarly process more transparent. A great advantage could be that researchers do not have to take care of the materials themselves. They can simply transfer the materials to institutions that specialise in this.

In the disciplines that focus on the arts, there are usually no fixed methodologies. Noordegraaf thinks that standardisation of the scientific method is not feasible. Sources are highly diverse and heterogeneous. Methodologies inevitably differ. Reaching agreement on which classification terms to use may prove to be a challenge. At the moment, it is very difficult to find relevant sources.

47. DILPS (Distributed Image Library Processing System) is a joint development with the Institute of Art History of the University of Frankfurt, the University
48. www.jar-online.net

Ideally, there should be only two portals: one for retrieving research data, and one user-friendly portal where data can be deposited. An editorial board should organize them; the presentation of the materials should be discussed with potential users. Researchers get no rewards for making their data available, so this is not something that researchers can currently be expected to do.

Marcel Ras and Barbara Sierman, KB, 14 June

In a future division, the KB will particularly focus on sustainable access to publications.[49] This is a problematic issue. Especially the museum field is too fragmented. The introduction of a central institution for digital data of the museum world would be helpful. However, no natural organisation is qualified to play such a central role.

The KB wants to give access to everything made *by*, *about* and *for* the Netherlands.[50] Sustainability and long term access are the main goals. Regulations help reaching these goals. At the moment, the KB mainly archives scientific publications in its e-Depot. One of its research questions is what to do with datasets that come along with publications. The provisional solution is to keep publications (KB) and datasets (DANS) separate. However, the institutions agree that this separation is artificial and not sustainable. As part of the DRIVER-II project, the KB investigates in collaboration with (among others) JALC, what kind of new e-Depot is suitable for archiving enhanced publications.

The KB preserves many kinds of data that could be of interest to researchers in the field of Art or Media Studies. Especially preservation of websites is a challenge because websites are increasingly complex, just as their archives. Also the current archiving agreements should be changed. Concrete plans for archiving research data are not (yet) written in the policy plan. A general collection policy is very welcome. In addition, there should be one central desk to which all scientists could bring their data. This would lower the threshold. Which institution handles the files behind this desk is not a primary concern.

Reinier van 't Zelfde, Rieke van leeuwen, Netherlands Institute for Art History / Rijks Kunsthistorisch Documentatiecentrum, 16 June

The RKD is a documentation centre which preserves important sources for art historical research. The RKD collections database provides information on the holdings of the RKD, but this database is not entirely complete. Large sections of the collection have been digitised. They are made accessible through the online databases RKDimages and RKDportraits. Images are maintained in an adLib database. Most of these images are offerered in open access. The RKD has not taken specific measures to secure the long-term preservation of their digital collections. The RKD would be very interested in experiments with enhanced publications, but the current technical facilities probably do not allow for such forms of linking.

The RKD offers various sources of information which may be relevant to researchers. It must be stressed that the RKD is not only a documentation centre, but also a research institute. The staff investigates biographies of artists and studies ascriptions of works of arts. It is clear that RKD databases such as RKDArtists and IconClass can help art historians to do their research. Staff of the RKD often gives workshops at Art History departments in the Netherlands.

Art historians who have completed research can transfer their archive to the RKD. At the moment, there are no good facilities for curating digital files other than images. The RKD tries to incorporate the information that is collected by researchers into their own databases. Since art historians outside of the RKD often use their own formats and their own databases, matching the information to the databases of the RKD is often a very difficult task. Sometimes some agreements can be reached about the structure of data, but this is highly exceptional. In some cases, data proves useless to the RKD.

Jeroen Sondervan, Publisher Humanities at Amsterdam University Press, 11 June

The Journal for Archaeology in the Low Countries (JALC) is an on-line open access journal, in which the articles are connected to resources that are available at the EDNA (eDepot voor Nederlandse Archeologie) which is managed at DANS. Conditions were very favourable. In JALC, AUP could

49. The NCDD report'Toekomst voor ons digitaal geheugen (2): strategische agenda voor duurzame toegankelijkheid': www.ncdd.nl/documents/NCDDToekomst_2_Strategischeagenda.pdf
50. The KB policy: www.kb.nl/bst/beleid/bp/2010/index.html

basically combine a number of ingredients which were available already. A survey has been held among researchers to investigate what they think about enhanced publications. It is a relatively new phenomenon and people clearly needed to get used to the interface. Another aspect that will be investigated is the workflow between authors and AUP. AUP also plans similar projects within other disciplines. A challenge in some of these projects may be the fact that repositories in the cultural heritage sector do not always work with the same standards that are used in the repository world.

Sondervan claims that datasets themselves are not likely to become publishable and marketable products. Also, the current generation of researchers demands open access, and expects to find information on-line, and is more willing to give access to their data than previous generations. Raising awareness of the benefits of on-line communication and open access publishing should also become part of training of academic skills. Sondervan holds the view that quality assurance of datasets is not the task of publishers. This should ultimately remain a responsibility of the research community.

The use of on-line data is still a relatively new phenomenon and practices surrounding the use of such resources still need to evolve. Considerable investments are needed, and if there are no clear immediate financial benefits, publishers are usually unwilling to make these investments.

Frans Havekes, Brill, 28 June

A first class of sources which may be relevant for researchers who focus on the arts are the many primary sources which have been digitized or filmed on microfilm by IDC Publishers. IDC Publishers was acquired by Brill in 2005. Brill also provides online access to scholarly databases. An important example is the World Christianity Database.[51] Other sources that may be relevant to researchers are bibliographies.[52]

Brill does not have a facility for researchers to deposit their datasets. Brill can offer a degree of continuity by taking care of the availability of databases after a project has ended. Brill usually works with an editorial board which tries to ensure the academic quality of a specific resource. Brill does not automatically accept all databases. The resource must fit within the publisher's profile; the focus is primarily on the humanities. Secondly, it has to be a product which is interesting from a commercial point of view. Brill also tries to "enhance" its publications. Improved searchability may clearly add value. Brill has never worked with a "Data Availability Policy". The lack of on-line supporting materials has never been a reason to reject a publication. Nevertheless, if authors have used databases, they are encouraged to include links.

A difference between Brill and DANS is that Brill accepts databases only if there is an active user community and if its exploitation is interesting from a commercial point of view. Archiving databases for its own sake is not interesting for Brill. Brill may make an arrangement with an external provider such as DANS but this certainly involves certain risks.

Peter Doorn en Heiko Tjalsma, DANS, 19 May

DANS is not necessary the place where all Dutch research data needs to be stored. One of the main mission of DANS is to ensure that there are good guidelines to secure the long-term preservation of data. For this purpose, DANS has developed a Data Seal of Approval (DSA) with requirements for the creator of data, for the curation centre and for the user of the data. Contrary to similar guidelines such as DRAMBORA, the DSA was kept relatively lightweight and easy to use. For DANS, it is very interesting to learn more about the attitudes of researchers towards data curation. When are researchers willing to share data? Do they have certain expectations or reservations? A well-known issue is the fact that researchers usually do not earn credits for the creation or the sharing of databases. Confusion about intellectual property may also affect the willingness to archive data. NWO is preparing a policy which stipulates that all data that have been produced on the basis of taxpayers' money should be available in open access. It will be interesting to explore what researchers would actually think of such an obligation.

51. www.worldchristiandatabase.org/wcd/
52. See, for example, the Linguistic Bibliography Online, www.blonline.nl/public/