Archiving the Past while Keeping up with the Times

VALENTIJN GILISSEN and HELLA HOLLANDER
Data Archiving and Networked Services (DANS), the Netherlands

The e-depot for Dutch archaeology started as a project at Data Archiving and Networked Services (DANS) in 2004 and developed into a successful service, which has ever since been part of the national archaeological data workflow of the Netherlands. While continuously processing archaeological datasets and publications and developing expertise regarding data preservation, various developments are taking place in the data landscape and direct involvement is necessary to ensure that the needs of the designated community are best met. Standard protocols must be defined for the processing of data with the best guarantees for long-term preservation and accessibility. Monitoring the actual use of file formats and the use of their significant characteristics within specific scientific disciplines is needed to keep strategies up-to-date. National developments include the definition of a national metadata exchange protocol, its accommodation in the DANS EASY self-deposit archive and its role in the central channelling of information submission. In an international context, projects such as ARIADNE and PARTHENOS enable further developments regarding data preservation and dissemination. The opportunities provided by such international projects enriched the data by improving options for data reuse, including the implementation of a map-based search facility on DANS EASY. The projects also provide a platform for sharing of expertise via international collaboration. This paper details the positioning of the data archive in the research data cycle and presents examples of the data enrichment enabled by collaboration within international projects.

Key words:
Data Archiving; Preservation; Standards; Collaboration; Access; Portals.

SDH Reference:
Valentijn Gilissen and Hella Hollander. 2017. Archiving the past while keeping up with the times, SDH, 1, 2, 194-205.

DOI: 10.14434/sdh.v1i2.23238

1. DANS AND THE RESEARCH DATA CYCLE

Data Archiving and Networked Services (DANS) is the Dutch research data archive. Predecessors in data archiving in the Netherlands date back to 1964; DANS was established in 2005 following revisions of the existing initiatives by the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organization for Scientific Research (NWO) [DANS 2017].

DANS has a notable role within both of its founding organizations. The KNAW is a branch research organization of fifteen internationally renowned Dutch research institutes, including DANS. DANS promotes sustained access to digital research data and carries out research on this topic; its data

Author's address: Valentijn Gilissen and Hella Hollander, Data Archiving and Networked Services (DANS), Anna van Saksenlaan 51, 2593 HW Den Haag, The Netherlands; email: valentijn.gilissen@dans.knaw.nl; hella.hollander@dans.knaw.nl

archiving services additionally serve as essential support to all other research institutes. NWO is the organization which funds scientific research at public research institutions in the Netherlands. Research which is funded by NWO should ensure that the resulting data is archived in a sustained form, which DANS enables through the use of its online archive, the Electronic Archiving System: EASY [DANS EASY 2017].

The research data cycle can be summarized as follows: a researcher produces research data; another researcher needs to be able to find, access and re-use the data for new research; the initial research should be referenced; the new research results in its own dataset, and so on. The policies of DANS aim to ensure the sustainability of the research data cycle in the long term:

1) To ensure preservation of data, DANS states that data should be archived in a repository which complies with international standards and guidelines of trustworthiness: a certified "Trusted Digital Repository" (TDR).

2) To support the use of Trusted Digital Repositories, funding organizations should oblige researchers to deposit their research data in a TDR.

3) In order to make data accessible, DANS promotes Open Access, but understands that for reasons such as privacy sensitivity of certain data files it is not always possible to make data available without restrictions. The position of DANS is "Open if possible, protected if necessary."

4) To accommodate re-use of data as well as to attribute credit to researchers, Persistent Identifiers, unique hyperlinks following a specific manner of resolving in order to remain valid over the long time should be used for referencing data sources in scientific publications. A dataset should have the same scientific value as a research article; the Persistent Identifier could be likened to the ISBN code of a publication.

Archives use the Open Archival Information Systems (OAIS) reference model [CCSDS 2012] to identify all sorts of aspects which need to be taken into account when managing data between submission by a Data Producer and dissemination to a Data Consumer. While the OAIS is not a set of instructions on "how to build your own archive," the functions and concepts defined within the OAIS allow an internationally shared view of what an archive should encompass. The reference model supports the methods for certifying an archive as a Trusted Digital Repository according to international standards. There are three degrees of certification for TDRs: the basic/essential certification of the Data Seal of Approval (DSA); the extended certification of the Nestor seal (DIN 31644); the formal certification as an ISO standard (ISO16363). The DANS archiving system EASY was awarded with the Nestor seal in early 2016 and was the first digital archive in the world to obtain this certificate.

The OAIS reference model is presented as a schematic process of "Management" situated between the Producer and the Consumer. Merely having an archive that is OAIS-compliant would not fully meet the aims DANS has as a digital archiving organization for research data. If the archive takes its proper place within the research data cycle, the flow of data should continue with the Data Consumer becoming a new Data Producer, using Persistent Identifier citation for referencing the source data.

2. INSIDE DANS EASY

Figure 1 shows the homepage of DANS EASY with its direct features highlighted in text boxes. A central search bar can be used to search through all metadata of all datasets. Information provided with "Search help" informs the user of search options such as the use of AND/OR Booleans and wildcard characters. Alternatively, the links to Advanced Search and Browse options can be followed for performing more specific searches.

A prominent button for submitting a new dataset deposit is presented in the center of the screen. For proceeding with a deposit or for accessing certain data files, users need to be logged in to the system; the options for registering or logging in remain present in the top of the screen until the user does so. It is not necessary to log in in order to search and browse through datasets, or to view their metadata - restricted access only pertains to downloading or opening data files.

The bottom of the screen holds a number of links to background information on the use of data, including full instructions for citing data. Additionally, the Trusted Digital Repository certificates which were obtained for EASY are presented here by their seals, with links to detailed information on the subject.

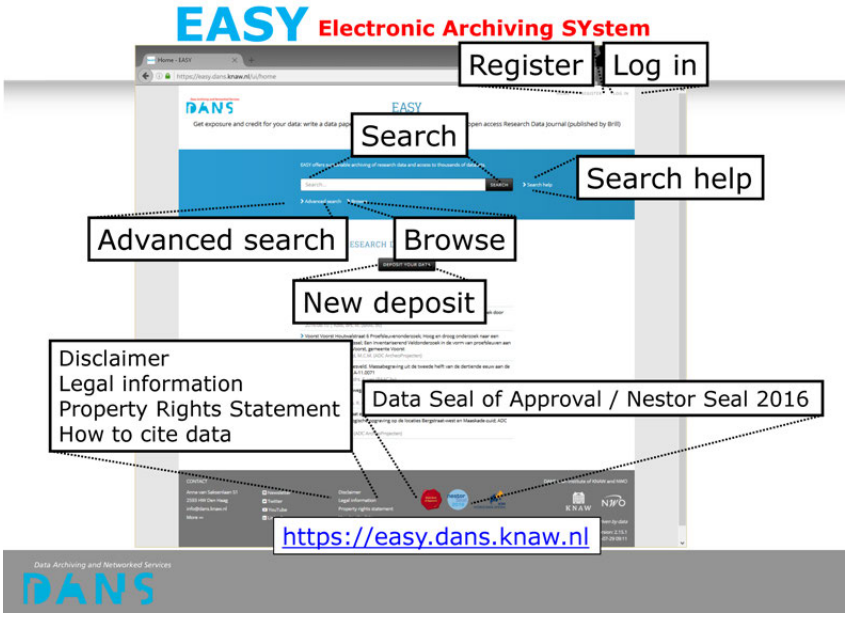


Figure 1. An overview of the homepage of EASY, DANS' Electronic Archiving System.

Datasets in EASY are described with metadata following the international standard (Qualified) Dublin Core. The Advanced Search options enable searches within specific Dublin Core metadata fields: Title; Creator; Description; Subject; Coverage; Identifier. A good example to show the benefit of using Advanced Search is a researcher who wants to find archaeological research projects conducted in the center of the Dutch city of Houten. "Houten" is also the Dutch word for "wooden." To avoid searching for Houten and additionally getting all the results of datasets that mention wooden objects

in their metadata, it is recommended that the user perform an Advanced Search on Houten within the metadata field "Coverage" (Temporal and Spatial coverage). Alternatively, if a researcher wants information on wooden artefacts but wants to avoid results which are only about the city of Houten, a search for "houten" should be done within the metadata field "Subject".

Browse options include means to refine search and browse results by audience, access category or by additional search queries. If, following on the above example, the researcher gets many results for a search for "Houten" in "Coverage," it could be that the Dutch word for "wooden" is part of a street name in a different city. A follow-up search could be another Advanced Search for "Coverage": "Utrecht", which is the name of the province wherein the city is located.

When a dataset is viewed, the user will see it presented in three tabs:

1) Overview: a highlighting of the dataset Title and (Abstract) Description from the metadata, which can be accompanied by pictures, illustrations, logos to represent the dataset. The overview also displays the correct way to cite the dataset in literature, by use of the "Digital Object Identifier" (DOI) Persistent Identifier.

2) Description: all of the (Qualified) Dublin Core metadata provided for the dataset.

3) Data files: the data files published with the dataset, with options for showing additional details if available, and for downloading if accessible. If not accessible, the page explains what conditions need to be met in order to download the data.

Datasets are described and deposited by researchers themselves. A deposit module takes researchers through the Dublin Core metadata fields as well as a page where they can upload the data files. Few fields are mandatory but a depositor is encouraged to fill in as much metadata as possible to make a dataset findable as well as understandable to anyone. If a research project was done in the city of Houten but Houten is only part of the Title and not specifically entered as "Spatial Coverage," the dataset would not be found in the use-case scenario given above.

When a dataset is being deposited, a data manager of DANS checks the incoming dataset for completeness and understandability. The manager may make minor changes or additions to the metadata or may provide migrations of file formats if this benefits the long-term preservation and accessibility of the data. The dataset will only be published after the data manager has performed all of the relevant quality assessments and preservation actions.

At the beginning of 2017, EASY contains over 33,000 published datasets from various scientific disciplines; it continues to receive new datasets on a daily basis. DANS especially accommodates the scientific disciplines classified under the Humanities, Social Sciences and Behavioral Sciences; the Dutch Technical Universities manage Exact Mathematical Science data themselves, in co-operation with DANS, to ensure sustainable storage. The vast majority of datasets in EASY are datasets from the discipline of Archaeology, totalling over 27,000 datasets and continuously growing (Figure 2). The relatively large number of archaeological datasets can be attributed to the successes of the e-Depot for Dutch Archaeology (eDNA), which started as a project at DANS in 2004 and is now embedded as a service within EASY [eDNA 2017]. eDNA raised awareness of the necessity of data archiving within the archaeological scientific community, digitized many gray literature reports for publication in

EASY, and enabled all of the archaeological project bureaus, municipalities and universities to use EASY to store and disclose their larger datasets.

Figure 2. A collage for a poster, showcasing the archaeological content of EASY.

While a number of archaeological datasets in EASY contain only a single PDF of a publication, described with Dublin Core metadata, there are many larger datasets available, and DANS especially aims to archive and publish such sets, which generally contain data tables, photographs, digital drawings and specialist reports alongside the final publication (Figure 2). A dataset should contain the final data of a research project and the size of the dataset generally matches the extent of the fieldwork conducted. The dataset of a non-intrusive survey will contain only a single publication; the dataset of an intrusive survey may contain more files such as data tables. The dataset of a full excavation may contain a large set of photographs taken in the field, digital drawings and scans of drawings of pits and profiles, daily reports, a database with registration tables and specialist determination tables for the different material types of the finds.

3. DANS PARTICIPATION IN INTERNATIONAL PROJECTS

By participating in (inter)national projects and infrastructures, DANS contributes to sustainable access to research data. DANS is involved with a large number of projects; three projects will be highlighted here which are of special interest to archaeologists.

3.1 CARARE

From 2010 to 2013, 29 European organizations worked together in the European CARARE project [CARARE 2017] to make two million archaeological and architectural objects accessible via the Europeana website. This target has been exceeded by 5 million objects. DANS contributed the archaeological publications that are published in EASY (Figure 3). In the process, DANS gained

valuable experience with metadata mapping, harvesting of selected metadata records from EASY, having the resource displayed on a map through translation of the national Dutch coordinate system and having the resource link back to the content via the Persistent Identifier.

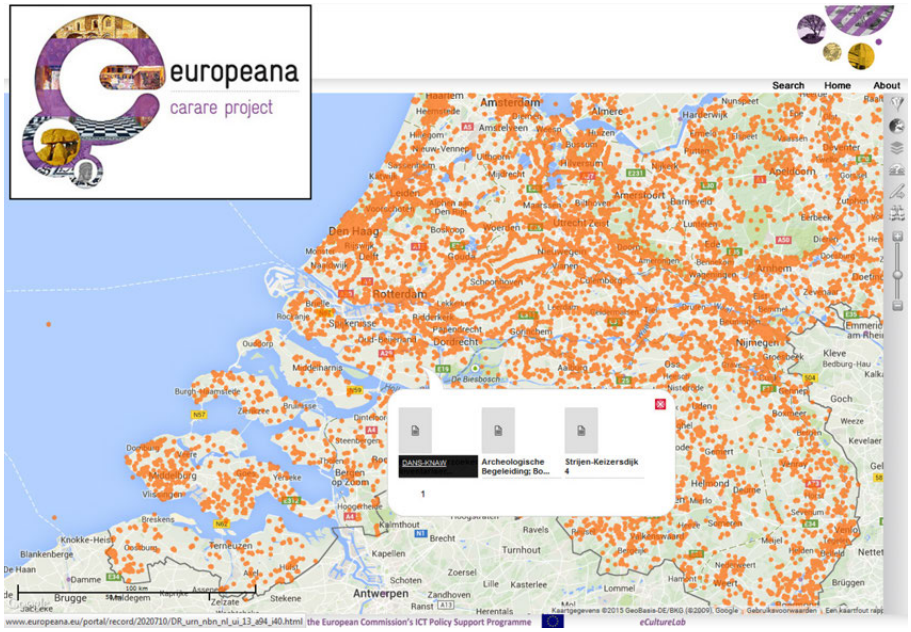


Figure 3. The first version of the CARARE map portal, which was available until December 2016. The portal is now incorporated in Europeana Collections [Europeana 2017]

3.2 ARIADNE

ARIADNE stands for Advanced Research Infrastructure for Archaeological Dataset Networking in Europe. 23 partners from 16 European countries collaborated in the ARIADNE project from February 2013 to January 2017 with the overall goal to establish a European research infrastructure for the integration of archaeological datasets. In addition, tools were developed that provide researchers access to this data [ARIADNE 2016]. DANS contributed data from the archaeological e-depot eDNA and the Digital Collaboratory for Cultural Dendrochronology (DCCD), making the data in EASY more visible internationally via the ARIADNE portal.

DANS collaborated closely with Leiden University in data mining and linked data activities, which allowed the mapping and translation of Dutch concepts from the national archaeological vocabulary to international vocabularies within the European infrastructure. Additionally, data mining was performed on the content of PDF publications from datasets that were lacking information in their metadata on coordinates. This enabled the addition of correct coordinates to the metadata of about 3500 datasets, a result of clear mutual benefit to DANS (elaboration of metadata) and the ARIADNE project (adding of content).

On the ARIADNE portal (Figure 4) thousands of resources can be found via a map, a timeline search and a keyword search, all with various options for filtering results. The mapping of keywords from national vocabularies allows comprehensive cross-searching the resources from all partners.

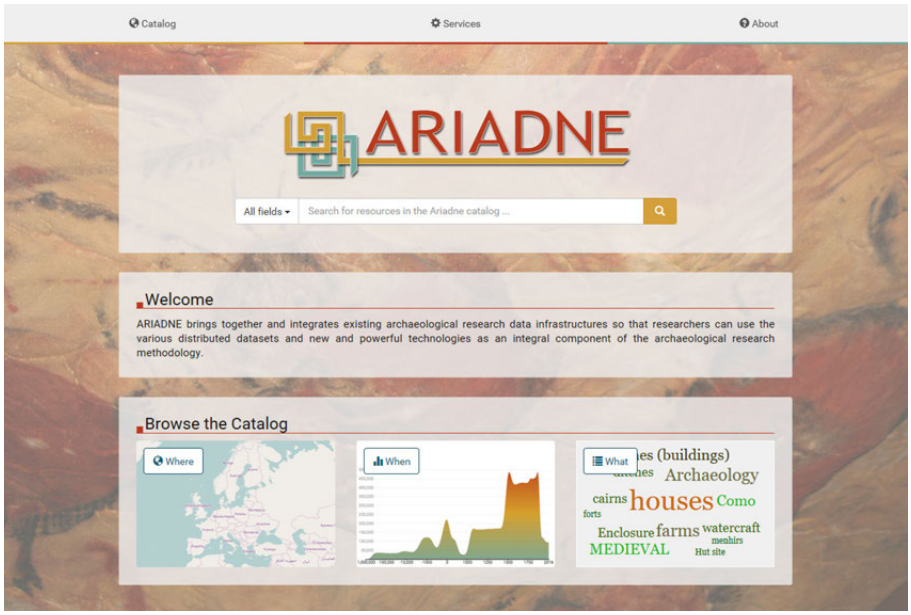


Figure 4. The ARIADNE portal [ARIADNE 2017].

Collaboration within ARIADNE led to the publication of new Guides to Good Practice, including a guide on Dendrochronology [Brewer and Jansma 2016] and a guide on 3D data [Trognitz et al. 2016].

The activities performed and the experience gained in order to have the EASY content displayed on the ARIADNE portal additionally enabled DANS to develop a map display feature in EASY, a feature which was implemented at the beginning of 2016. Search and browse results are initially shown in a list, as has always been the case, but the display can now be switched from "List" to "Map" (Figure 5a). All of the search/browse results which include coordinates are then displayed on OpenStreetMap in agglomerations of results. Zooming in will result in the agglomerations to spread out, to the point that single results are found.

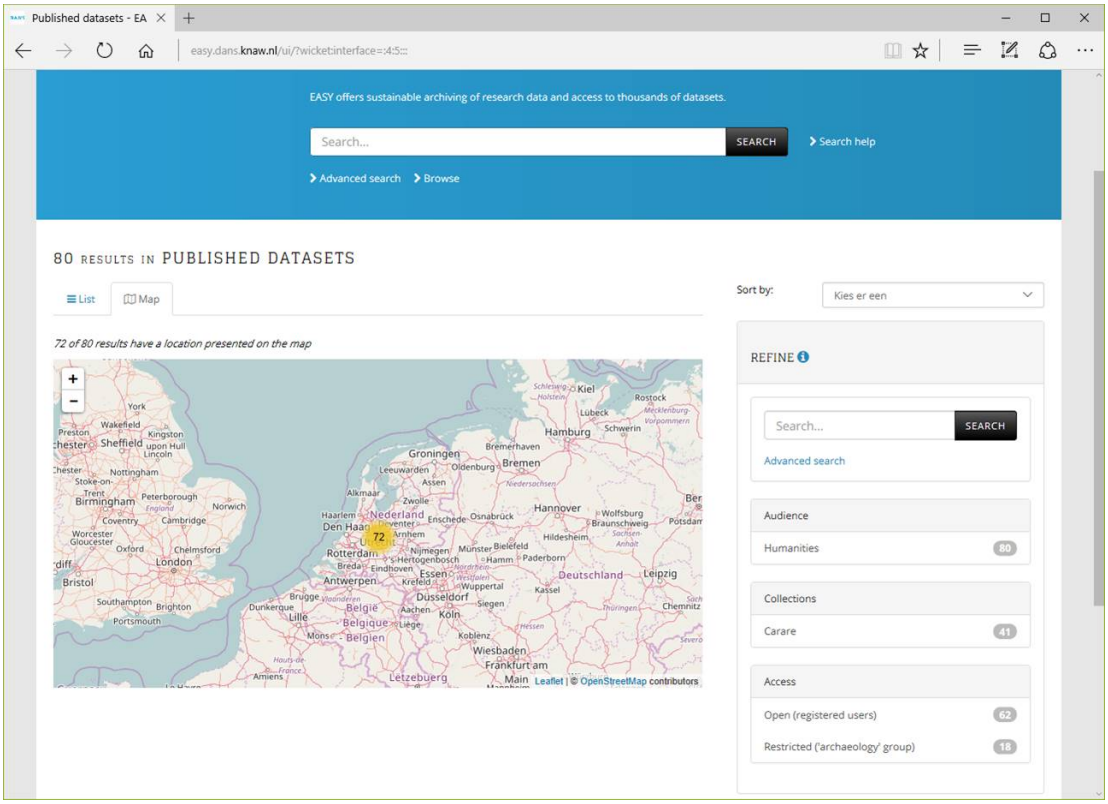


Figure 5a. The map display feature for browse results in EASY.

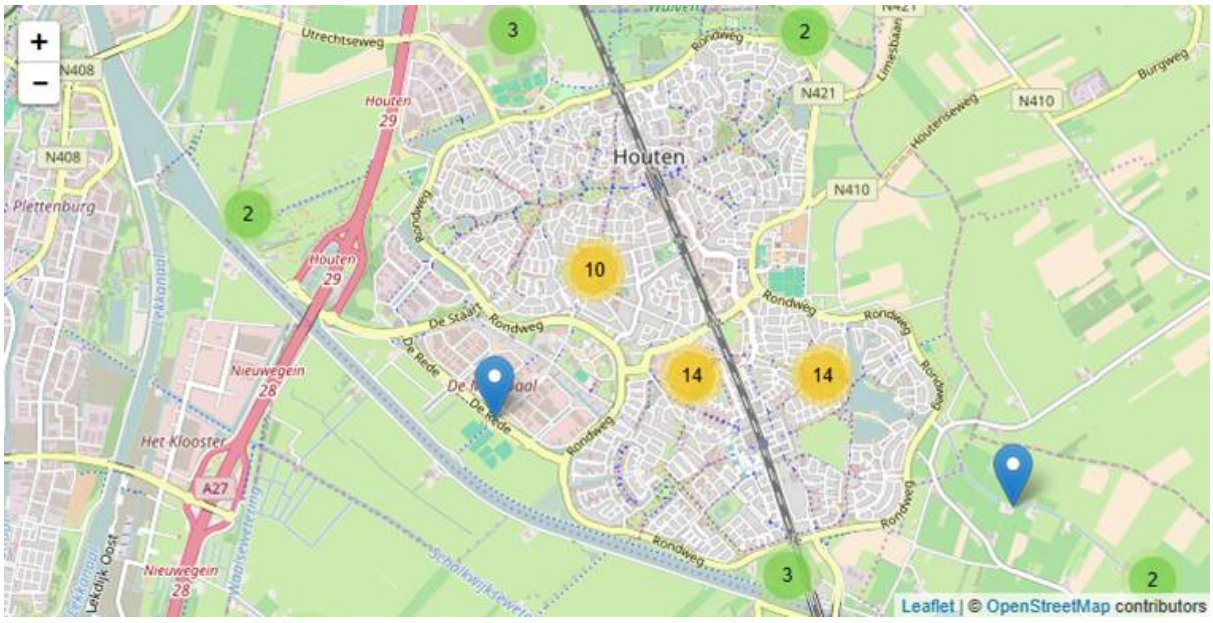


Figure 5b. A zoom on search results for Houten in the map display feature.

Figure 5b shows a zoom on an advanced search for "Houten" in "Coverage," following on from the example presented in section 2. With several dozen hits, showing the search results in List display would not be very helpful for a researcher who is only interested in archaeological research carried out in the center of the city. Switching to Map display allows zooming to the location and selecting single results in the target area.

Apart from the invaluable contributions to the development on the Map display feature, DANS was also able to make use of the work done on mapping of metadata within the ARIADNE project to contribute to the national development of an XML standard for archaeological data tables. This XML standard was developed by the Dutch archaeological sector as a national exchange protocol [SIKB 2016]. The protocol serves to provide more and better metadata for archaeological projects and to have terminology standardized according to the national archaeological vocabulary. It allows for a complete export of archaeological database tables to standardized XML, which means that every archaeological company can use their own database system but still provide an export to make the data interoperable. DANS implemented the protocol in EASY to the effect that data depositors can now upload the XML with new deposits and have metadata from the XML extracted in the EASY Dublin Core metadata fields. This has proved to be a very efficient means to provide full and correct metadata with a data deposit, saving time and effort for depositors as well as for data managers.

3.3 PARTHENOS

PARTHENOS stands for Pooling Activities, Resources and Tools for E-Heritage Research Networking Optimization and Synergies. Whereas CARARE and ARIADNE focused on Archaeology, PARTHENOS empowers digital research across all fields of the (Digital) Humanities, including History, Language Studies, Cultural Heritage and related fields. The interdisciplinary four-year project, which began in May 2015, provides a thematic cluster of European Research Infrastructures (e-infrastructures and other world-class infrastructures), carries out integrating initiatives, and builds bridges between different but interrelated fields of research. [PARTHENOS 2017]

Central topics are the implementation of common AAA (Authentication, Authorization, Access) and data curation policies within the framework of the data lifecycle, including long term preservation, certification and Intellectual Property Rights (IPR).

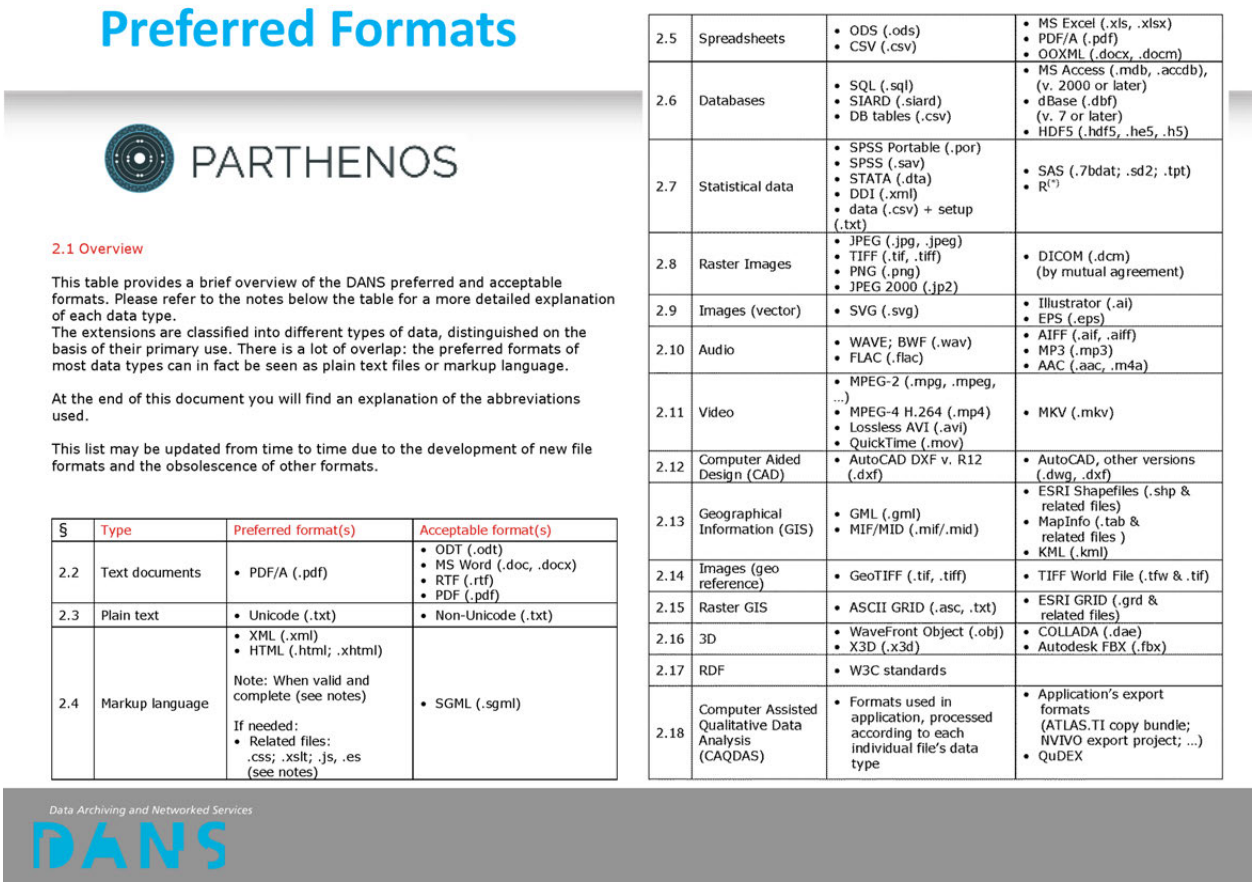
Expected key results of the project are:

1) Guidelines on data management: to produce a coherent, authoritative, well accepted set of policies/guidelines/tools concerning the management of data lifecycle and related issues such as IPR, quality and so on.
2) Standardization and semantics: to produce a wide set of standards and semantics, originated from community needs and tailored to the methodology and intended use by researchers.
3) Services and tools: to produce a coherent set of tools for carrying out research using and re-using data.

DANS is working on the harmonization of research data management within the various disciplines and the certification of their repositories. DANS contributes its guidelines and experiences on data

management to the PARTHENOS resources. In return, the international collaboration on data policies and protocols allows enhancing DANS' own guidelines.

One of the topics covered by such guidelines is the topic of Preferred Formats, the best choices for file formats over the long term. DANS published its Preferred Formats guide in September 2015 [DANS 2015], which details the best options for long-term preservation per file type (Figure 6). The guidelines aim to make data available in file formats which are, as far as possible: open formats; frequently used; independent of specific software, developers or vendors. A working group within DANS is responsible for maintaining the guidelines, which can be subject to revision based on issues such as new file formats occurring in dataset deposits, or new software developments. PARTHENOS functions as a platform for international discussions on the subject, which also contribute to updates of the Preferred Formats guidelines.



Preferred Formats

PARTHENOS

2.1 Overview

This table provides a brief overview of the DANS preferred and acceptable formats. Please refer to the notes below the table for a more detailed explanation of each data type.
The extensions are classified into different types of data, distinguished on the basis of their primary use. There is a lot of overlap: the preferred formats of most data types can in fact be seen as plain text files or markup language.

At the end of this document you will find an explanation of the abbreviations used.

This list may be updated from time to time due to the development of new file formats and the obsolescence of other formats.

§	Type	Preferred format(s)	Acceptable format(s)
2.2	Text documents	• PDF/A (.pdf)	• ODT (.odt) • MS Word (.doc, .docx) • RTF (.rtf) • PDF (.pdf)
2.3	Plain text	• Unicode (.txt)	• Non-Unicode (.txt)
2.4	Markup language	• XML (.xml) • HTML (.html; .xhtml) Note: When valid and complete (see notes) If needed: • Related files: .css; .xslt; .js, .es (see notes)	• SGML (.sgml)
2.5	Spreadsheets	• ODS (.ods) • CSV (.csv)	• MS Excel (.xls, .xlsx) • PDF/A (.pdf) • OOXML (.docx, .docm)
2.6	Databases	• SQL (.sql) • SIARD (.siard) • DB tables (.csv)	• MS Access (.mdb, .accdb), (v. 2000 or later) • dBase (.dbf) (v. 7 or later) • HDF5 (.hdf5, .he5, .h5)
2.7	Statistical data	• SPSS Portable (.por) • SPSS (.sav) • STATA (.dta) • DDI (.xml) • data (.csv) + setup (.txt)	• SAS (.7bdat; .sd2; .tpt) • R(*)
2.8	Raster Images	• JPEG (.jpg, .jpeg) • TIFF (.tif, .tiff) • PNG (.png) • JPEG 2000 (.jp2)	• DICOM (.dcm) (by mutual agreement)
2.9	Images (vector)	• SVG (.svg)	• Illustrator (.ai) • EPS (.eps)
2.10	Audio	• WAVE; BWF (.wav) • FLAC (.flac)	• AIFF (.aif, .aiff) • MP3 (.mp3) • AAC (.aac, .m4a)
2.11	Video	• MPEG-2 (.mpg, .mpeg, ...) • MPEG-4 H.264 (.mp4) • Lossless AVI (.avi) • QuickTime (.mov)	• MKV (.mkv)
2.12	Computer Aided Design (CAD)	• AutoCAD DXF v. R12 (.dxf)	• AutoCAD, other versions (.dwg, .dxf)
2.13	Geographical Information (GIS)	• GML (.gml) • MIF/MID (.mif/.mid)	• ESRI Shapefiles (.shp & related files) • MapInfo (.tab & related files) • KML (.kml)
2.14	Images (geo reference)	• GeoTIFF (.tif, .tiff)	• TIFF World File (.tfw & .tif)
2.15	Raster GIS	• ASCII GRID (.asc, .txt)	• ESRI GRID (.grd & related files)
2.16	3D	• WaveFront Object (.obj) • X3D (.x3d)	• COLLADA (.dae) • Autodesk FBX (.fbx)
2.17	RDF	• W3C standards	
2.18	Computer Assisted Qualitative Data Analysis (CAQDAS)	• Formats used in application, processed according to each individual file's data type	• Application's export formats (ATLAS.TI copy bundle; NVIVO export project; ...) • QuDEX

Data Archiving and Networked Services
DANS

Figure 6. The overview table of the DANS Preferred Formats guidelines.

4. CONCLUSION

This paper explained the role of DANS as a Trusted Digital Repository within the Research Data Cycle and described the EASY Electronic Archiving System and its archaeological content. Furthermore, it gave examples of results from DANS' involvement within three international projects and the mutual benefits from DANS bringing content to those projects and enriching its own output in return.

The paper has different objectives, depending on the background of the readers. Researchers are encouraged to use the described services. The search and browse options of DANS EASY allow finding datasets from Dutch scientific research including a large number of archaeological records; the portals of Europeana/CARARE and ARIADNE assist in finding datasets of sources from all over Europe which can be useful for further research. When creating data, recommendations and guidelines coming from the PARTHENOS project should help in keeping the data findable, accessible, interoperable and re-usable.

When re-using data, researchers are strongly encouraged to keep the research data cycle alive – by citing the data source and by depositing their own datasets. Storing data in a Trusted Digital Repository will give the best guarantees for sustainability in the long term. Organizations with a TDR like DANS will continue to provide content to existing portals and to participate in projects to disclose data. Therefore, new datasets deposited in a TDR such as EASY will also be enriched by making them findable through innovations such as the ARIADNE portal.

Readers working in other parts of the research data cycle such as data management are encouraged to participate as much as possible in (inter)national projects and to work together to promote sustained access to research data. The examples given in this paper show that international collaboration can generate great mutual benefits. There is much to gain in connecting data across borders and disciplines, for anyone involved – and there is no reason not to work together; we should all be friends.

5. REFERENCES

ARIADNE - Advanced Research Infrastructure for Archaeological Dataset Networking in Europe. 2016. Building a research infrastructure for Digital Archaeology in Europe. ARIADNE Booklet, December 2016. Retrieved April 13, 2017 from http://www.ariadne-infrastructure.eu/About

ARIADNE - Advanced Research Infrastructure for Archaeological Dataset Networking in Europe. ARIADNE portal. Retrieved April 13, 2017 from http://portal.ariadne-infrastructure.eu

Brewer, Peter and Esther Jansma. 2016. Dendrochronological Data in Archaeology: A Guide to Good Practice. Archaeology Data Service / Digital Antiquity. Guides to Good Practice. Retrieved April 13, 2017 from http://guides.archaeologydataservice.ac.uk/g2gp/Dendro_Toc

CARARE - Connecting Archaeology and Architecture in Europeana. Retrieved April 13, 2017 from http://www.carare.eu

CCSDS - Consultative Committee for Space Data Systems. 2012. Reference Model for an Open Archival Information System (OAIS). Recommended Practice CCSDS 650.0-M-2. Magenta Book, Washington, DC: CCSDS Secretariat. Retrieved April 13, 2017 from https://public.ccsds.org/pubs/650x0m2.pdf

DANS - Data Archiving and Networked Services. Retrieved April 13, 2017 from https://dans.knaw.nl/en

DANS - Data Archiving and Networked Services. 2015. File formats, preferred formats and accepted formats. Preferred Formats. Version 3.0, September 2015. Retrieved April 13, 2017 from https://dans.knaw.nl/en/deposit/information-about-depositing-data

DANS EASY - Data Archiving and Networked Services Electronic Archiving System. Retrieved April 13, 2017 from https://easy.dans.knaw.nl/ui/home

eDNA - e-Depot for Dutch Archaeology. Retrieved April 13, 2017 from https://dans.knaw.nl/nl/over/diensten/data-archiveren-en-hergebruiken/easy/edna

Europeana. Europeana collections. Retrieved April 13, 2017 from http://www.europeana.eu/portal/en

PARTHENOS - Pooling Activities, Resources and Tools for E-Heritage Research Networking Optimization and Synergies. Retrieved April 13, 2017 from http://www.parthenos-project.eu

SIKB – Stichting Infrastructuur en Kwaliteitsborging en Bodembeheer. 2016. SIKB0102 Archeologie, XML exchange standard. Retrieved April 13, 2017 from http://www.sikb.nl/datastandaarden/richtlijnen/sikb0102

Trognitz, Martina, Kieron Niven, and Valentijn Gilissen. 2016. 3D Models in Archaeology: A Guide to Good Practice. Archaeology Data Service / Digital Antiquity. Guides to Good Practice. Retrieved April 13, 2017 from http://guides.archaeologydataservice.ac.uk/g2gp/3d_Toc

Received March 2017; revised July 2017; accepted August 2017.