ARCHIVING AND MANAGING RESEARCH DATA

DATA SERVICES TO THE DOMAINS OF THE HUMANITIES AND SOCIAL SCIENCES AND BEYOND: DANS IN THE NETHERLANDS

von Peter Doorn

INTRODUCTION

Data sharing has become a default requirement made by an increasing number of research funding and research performing organizations. Data should be findable, accessible, interoperable and reusable, in an as open as possible way, is the adagio of today. The idea is that the research system will be more efficient if data sharing will be part of the dominant research culture. This should lead both to a greater transparency of research, because FAIR data can be checked and will contribute to replicability of research. And for researchers it will make it possible to stand on the shoulders of predecessors, opening possibilities for comparative research or answering new questions on the basis of existing data.

This is the theory. But in how far does the above serve the needs of the users? In how far is data that is offered for sharing actually being reused? And in how far do "old" data contribute to new knowledge creation? Actually, not very much is known about the reuse of data, and even less about how this reuse leads to new scientific insights. Although recommendations for citing data abound, it even appears to be very hard to trace back reused data in the literature.

The core of this paper is on the use of a national data service, taking the EASY repository of Data Archiving and Networked Services (DANS) in the Netherlands as a case study, and presenting a quantitative overview for the period 2007-2019.

DATA ARCHIVING AND NETWORKED SERVICES

DANS is the Netherlands' institute for permanent access to digital research resources. It encourages researchers to make their digital research data and related outputs Findable, Accessible, Interoperable and Reusable (FAIR). Open if possible, protected where necessary. DANS provides expert advice and certified services. The institute was created in 2005, with the task to serve humanities scholars and social scientists. Over time, DANS evolved to serve additional audiences as well. The first predecessor of DANS was set up in 1964. DANS is an institute of the Royal Netherlands Academy of Arts and Sciences KNAW and the national funding organisation Netherlands Organisation for Scientific Research NWO.

The DANS core services include EASY (Electronic Archiving System) for long-term archiving, DataverseNL as a repository service for universities, research institutes and higher education, and NARCIS, the national portal for research information. By participating in (inter-)national projects, networks and research, DANS contributes to continued innovation of the global scientific data infrastructure. As the founder of the Data Seal of Approval (now: Core Trust Seal) DANS also provides expertise on trustworthy preservation of digital data and training for data professionals and researchers.

The majority of the datasets to which DANS provides access are stored in the EASY archive. This paper focuses on the use of this archive, both by data providers submitting data and by users downloading data. We will look into the growth of the

archive since 2007, when the EASY system went operational, and present metrics on subjects such as:

- the size distribution of the archived datasets: DANS typically serves the "long tail of science";
- the distribution of archived datasets according to type of access: the number of datasets that is publicly or openly accessible is rapidly increasing over time;
- the growth of downloads from the data archive (in total and according to discipline, in absolute and relative terms – taking variations in size of disciplines into consideration); and
- the popularity of datasets among users: frequencies of downloads of datasets, as well as the top 25 of datasets used.

GROWTH OF DATA ARCHIVE EASY

The number of datasets archived in the DANS Electronic Archiving System (EASY) increased from around 1,500 in 2007 to almost 120,000 by the end of 2019 (fig. 1). A dataset usually consists of the data belonging to a particular project and can contain one or more data files.

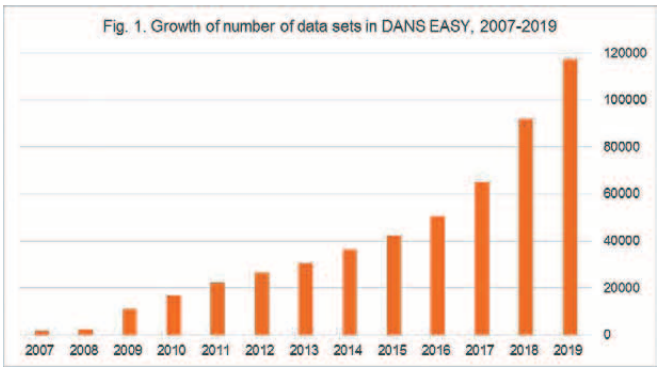

Fig. 1. Growth of number of datasets in DANS EASY, 2007-2019

In the early years, the annual growth rate was relatively high, due to the relatively small numbers before 2011 and due to a retro-digitization project of archaeological data and reports in collaboration with the State Archaeological Service (now: Cultural Heritage Agency of the Netherlands – RCE). Since 2012 the annual growth rate of the collection has fluctuated around 15 and 20 %, and over the last three years the growth rate has increased to 30-40 % per year. The recent acceleration of growth is caused by an increase of bulk uploads of institutional collections or repositories. Bulk ingest into the archive was already considerable in the time of the retro-project mentioned, and it has become more important in recent years, because of new institutional arrangements in which DANS serves as a background archive for a growing number of universities and other research organisations. These have set up their own institutional repositories, which are automatically ingested their data holdings into the DANS archive for long-term preservation and access. Yet, individual researchers are still uploading their data to the archive as well.

Many datasets consist of multiple data files. The average number of files per dataset varies considerably, but on average it is over a hundred. The total number of files in EASY increased almost linearly over time, from a little under 15,000 in 2007 to over 4 million in mid 2017 and is expected to be over 10 million by the end of 2019.

The overwhelming majority of the datasets stored in the DANS archives comes from archaeology, the (other) humanities and the social sciences, where the average data sizes are modest (see also further down, fig. 4). Before 2009, the mean size per dataset was less than 100 Mb; from 2010-2013, the mean size was around 150-200 Mb; and since then the mean size fluctuates around 400 Mb per dataset. The total volume of data archived in EASY (expressed in storage at one location, in reality there is multiple storage at two or more different sites) grew from 0,1 Tb in 2007 via about 15 Tb in mid-2017 to over 30 Tb by the end of 2019.

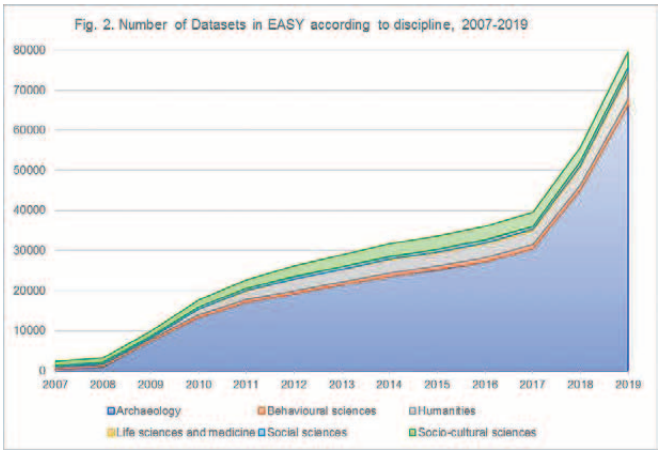

Fig. 2. Number of datasets in EASY according to discipline, 2007-2019

Figure 2 gives an impression of the size of the archive according to discipline over the past 12 years. The total number of datasets is inflated by about 3,000, because of datasets classified under more than one discipline. We restricted ourselves here to data from Dutch researchers and have not taken into account data from two international repositories for which DANS is the background archive: Mendeley Data and Dryad. Some 16,000 datasets from Mendeley Data from a broad array of disciplines have been ingested into the DANS archive, and the Dryad collection consists of about 28,000 datasets, mainly from the life sciences, biology and climate research.

It is clear from the graph that the relatively young archaeological data archive (E-Depot for Dutch Archaeology or EDNA, started in 2004) quickly became the largest section of the DANS archive, even though archaeology is a relatively small academic domain. The success of EDNA is largely explained by several factors: (1) the above-mentioned retro-digitization of archaeological data and reports contributed about 30,000 datasets; (2) the fact that RCE has made deposit at DANS obligatory for data belonging to every archaeological project carried out in the Netherlands; (3) recently, DANS has begun to store data from archaeological finds in private collections, which has resulted in an increase of about 23,000 datasets (Project Portable Antiquities of the Netherlands - PAN). We do not know which proportion of data from other areas is not archived at DANS, but compared to archaeology it is certain that in most research fields only a small percentage of the data that researchers produce is archived for long-term access. The greater attention to data policies by research funders, universities and other institutions has not yet markedly altered this situation.

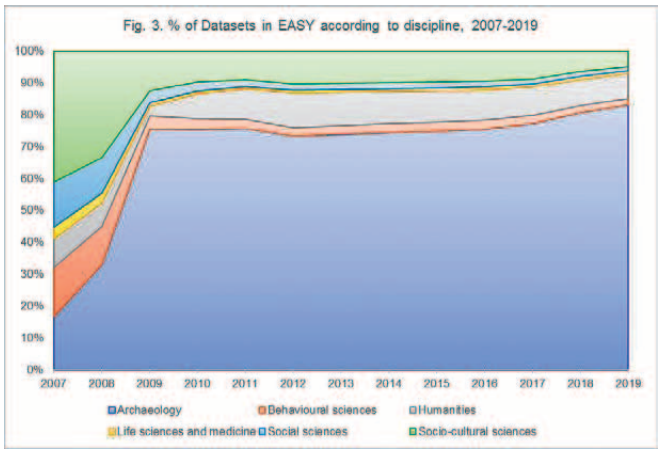

Fig. 3. % of Datasets in EASY according to discipline, 2007-2019

The large share of archaeological datasets in the archive is also apparent from figure 3. In the early years of DANS, the proportion of archaeological datasets grew from less than 20 % to about 75 % in 2017 and to 83 % recently (2019). Since 2009 the proportional distribution of archived datasets over the main disciplines has not changed much, except for the last two years due to the PAN data. The social and behavioural sciences are responsible for about 30 % of the downloads, and the humanities (without archaeology) for a bit less than 10 %. The life sciences are a growing, yet small category in DANS, taking about 3 % of the downloaded datasets in 2019.

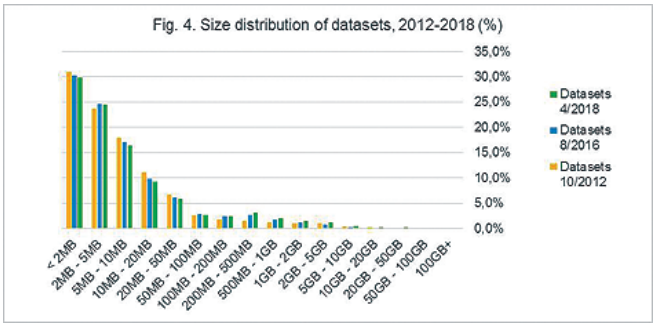

Fig. 4. Size distribution of datasets, 2012-2018 (%)

The size distribution of the archived datasets is reflected in figure 4. It is clear from the graphs, that the overwhelming majority of the stored datasets at DANS are fairly small, which is (still?) characteristic for the humanities and social sciences. DANS typically serves the "long tail of science". Although there is a tendency that datasets deposited in later years are somewhat bigger than those in earlier years, the effect on the size distribution is trivial. How time-related the term Big Data is, can be illustrated by the example of the population census of 1960, the first census in the Netherlands to be fully computerized. The 11 million punch cards of the original file take a little less than 0,5 Gb of storage. In the 1970s, one statistical run with SPSS on this dataset cost the full annual computing budget of the faculty of social sciences of the University of Amsterdam. The Big Data of the 1970 is just a modest-size dataset today.

Fig. 5. Distribution of archived datasets according to type of access, 2012-2019

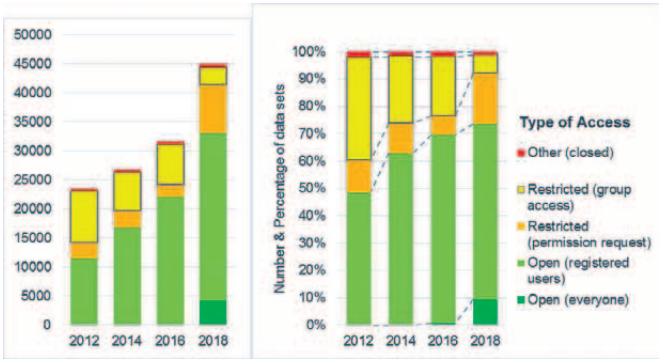

Fig. 5(1)
A. Excluding Mendeley Data and Dryad

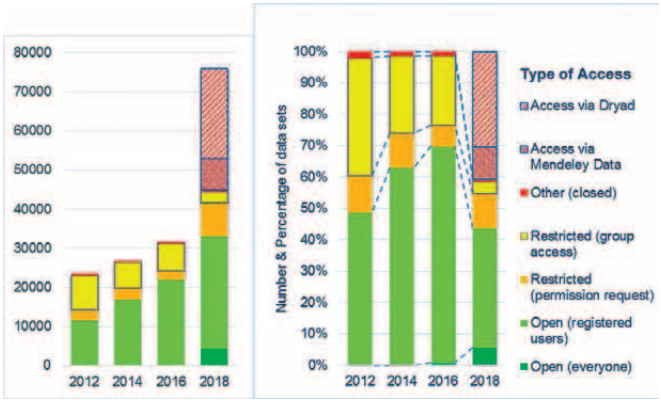

Fig. 5(2)
B. Including Mendeley Data and Dryad

Looking at the datasets according to type of access, it is remarkable that the number of datasets that is publicly or openly accessible is considerably increasing over time (fig. 5). This is both an indication of increased attention for open access in institutional data policies and for the increased awareness among researchers that sharing data is useful. It can be interpreted as a sign of acceptance of open science principles, at least among researchers depositing their data in the DANS archive. While the percentage of openly accessible data in EASY was less than 50 % in 2012, this figure had increased to 70 % in 2016. The datasets requiring explicit permission for access by the depositor decreased both in absolute and in relative terms. The category "group access" pertains to the archaeological sector, where access to detailed data is limited to professional archaeologists in order to prevent the disturbance of heritage sites. Originally DANS required users to register and log in if they wanted to deposit and download data. Since 2016 DANS also grants full open access without registration (also known as CC-0 license or public domain dedication) for data without any copyright (although data citation is still required according to the academic code of conduct).

After 2016 DANS began ingesting datasets from the international repositories Mendeley Data and later also from Dryad on a contract basis. Although this data is openly accessible through both repositories, the contract refers users back to the original repositories instead of granting access in the EASY archive. This

had a substantial effect on the total data accessibility at DANS (fig. 5b). However, disregarding the datasets from these two repositories, the percentage of openly accessible data further increased in 2018.

USE OF THE DATA ARCHIVE

Since 2007, the total number of unique visitors (counted as unique in any month of the year) increased from a little over 6000 in 2007 to about 100,000 in 2019 (fig. 6). If we count only registered users, who use EASY while being logged in, the numbers are much smaller: the logged-in visitor number grew from 500 in 2007 to 4500 in 2014 and since then fluctuated around 4000. This is probably related to the fact that DANS began to make data available for downloading without registration around that year. We no longer require logging in for a steadily growing number of fully open data.

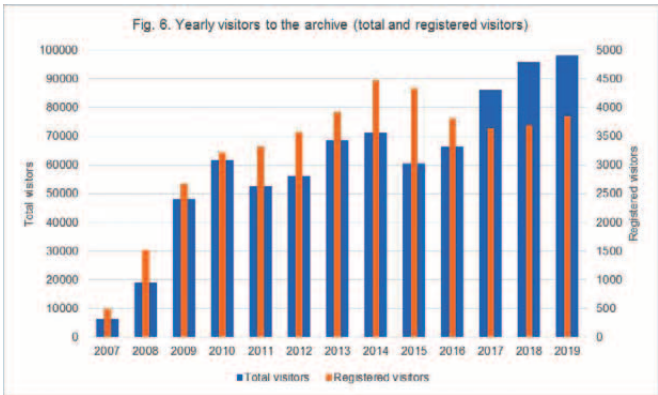

Fig. 6. Yearly visitors to the archive (total and registered visitors)

Between 2007 and August 2017, a total of 223,258 datasets, containing 2,361,588 files had been downloaded from EASY. By December 1st, 2019, these numbers had further grown to 312,472 datasets containing 3,017,817 files.

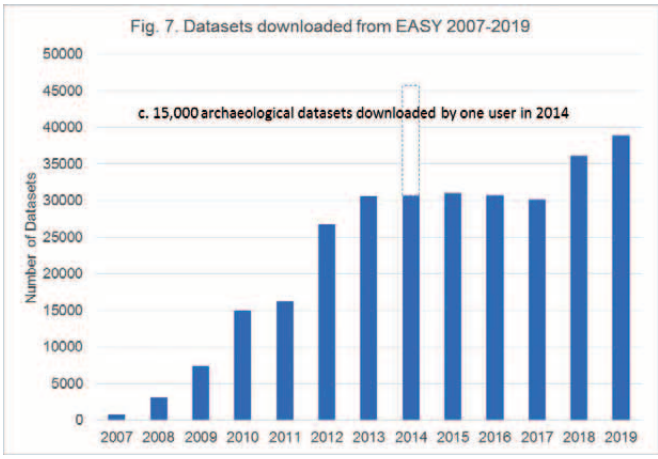

Fig. 7. Datasets downloaded from EASY 2007-2019

The reuse of datasets has increased in a fairly similar way as the growth of the archive until 2013. In 2007 only about 750 datasets were downloaded for reuse, a number which grew to around 30,000 in 2013 (fig. 7). From that year until 2017, the number of downloads per year more or less stabilized. 2014 contained an exceptional case, in which one user downloaded all openly accessible archaeological data, about 15,000 datasets in total. In 2018 and 2019, we saw a further growth in the number of downloads to almost 40,000 datasets. Downloads via Dryad and Mendeley Data are not counted, as they are not accessible directly from the DANS archive.

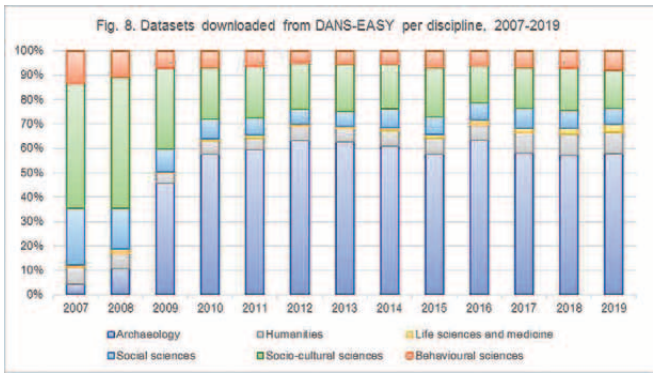

Fig. 8. Datasets downloaded from DANS-EASY per discipline, 2007-2018

Unsurprisingly, with a view on the size of the collection, archaeology is also the domain with the highest number of downloads in absolute terms: in recent years (since 2012), the number fluctuated between 20,000 and 24,000. Next are the socio-cultural sciences (between 5,500 and 7,800 downloads annually since 2012), the social sciences (between 2,000 and 3,100) and the behavioural sciences (1,600-2,700). The downloads in the humanities varied from 1,900 to 2,800 per year. Note that the absolute numbers of downloads underlying figure 8 may be slightly inflated because of some datasets being counted under more than one domain.

Also here we see the spectacular growth of the archaeological data archive: within five years' time, between 2007 and 2012, the proportion of downloads in this domain grew from less than 5 % of the total to about 60 %. From then on, it stayed more or less at that level, with a slight tendency to decrease in relative terms. This surge in archaeological data sharing went at the costs of the other disciplines (again: in relative terms), especially of the social sciences in the broad sense (including socio-cultural and behavioural sciences), of which the share of downloads went down from 90 % to around 30 % in the same five years.

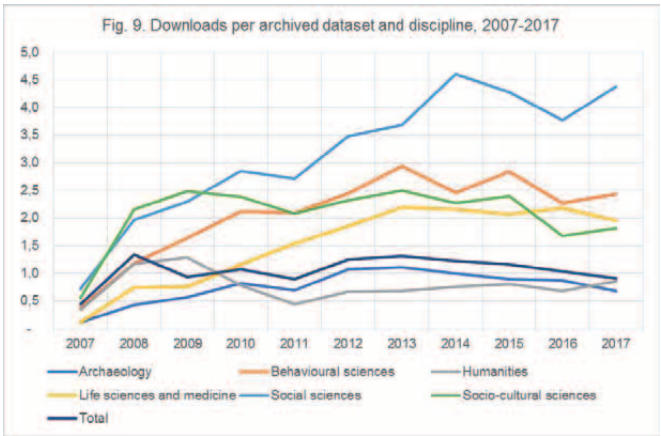

Fig. 9. Downloads per archived dataset and discipline, 2007-2017

It is instructive to look at the downloads relative to the size of the archive per domain (fig. 9). If we divide the number of downloads by the number of archived datasets per discipline, we get an indicator of relative data reuse. The mean for all datasets is about 1 throughout the investigated period, meaning that every dataset in the archive is downloaded about once every year. Despite the rise of archaeology at the cost of the social sciences just mentioned, the data reuse in the social sciences remained clearly on top since 2010, with a reuse ratio of between 3 and 4.5. The social scientists are followed by the groups of behavioural, socio-cultural and life sciences, with between 2 and 3 downloads per archived dataset per year. Archaeology and the humanities score slightly below average, with a reuse ratio of respectively around 1 and 0.75.
So far, we dealt with average downloads from the archive, in total and per discipline. But of course, the distribution of downloads is highly skewed: there are more and less popular datasets, and like there are books in a library that nobody ever borrows, there are also datasets that are not (yet) or only rarely downloaded. Until August 2017, of the 36,561 datasets then archived in EASY, 31,924 datasets had been downloaded at least once: a gross reuse percentage of 87 %, which represents the highest proportion of re-used data we recorded. Since then the rate of re-used has dropped. This clearly has to do with the very strong growth of the collection in the past two years. Obviously, datasets that have recently been added to the archive have a smaller chance to get downloaded than datasets that have been in the collection for several years. Where the content of the archive more or less doubled (not counting the datasets from Dryad and Mendeley Data), the re-use grew by less than 30 %.

Rank		Title of Dataset	Persistent Identifier	Dataset Downloads		Downloaded Files	
2019	2017			2019	2017	2019	2017
1	2	De steentijd van Nederland	urn:nbn:nl:ui:13-tg4-mof	1280	1108	1651	1452
2	1	Nationaal Kiezersonderzoek, NKO 2006	urn:nbn:nl:ui:13-4zd-x4e	1194	1125	5629	5398
3	5	Netherlands Longitudinal Lifecourse Study - NELLS First Wave - 2009 - versie 1.3	urn:nbn:nl:ui:13-54c-uie	1161	742	2728	1920
4	6	Geological-Geomorphological map of the Rhine-Meuse delta, The Netherlands	urn:nbn:nl:ui:13-nqjn-zl	1088	736	50519	34645
5	7	Nationaal Kiezersonderzoek 2012 - NKO 2012	urn:nbn:nl:ui:13-93iu-8p	1002	684	1881	1339
6	4	Nationaal Kiezersonderzoek, 2010 - NKO 2010	urn:nbn:nl:ui:13-9x4l-vy	979	829	4911	4279
7	12	Netherlands Longitudinal Lifecourse Study - NELLS Panel Wave 1 2009 and Wave 2 2013 - versie 1.2	urn:nbn:nl:ui:13-5wyt-c6	966	496	1917	1023
8	11	Nationaal Kiezersonderzoek, NKO 1971-2006 cumulatieve file	urn:nbn:nl:ui:13-e9w-iq9	912	553	2081	1306
9	3	Brabant cohort - derived student file	urn:nbn:nl:ui:13-zgkg-jv	887	884	2169	2161
10	14	NLGis shapefiles	urn:nbn:nl:ui:13-wsh-wv7	872	452	165019	149040
11	19	WoON2015: release 1.0 - WoonOnderzoek Nederland 2015	urn:nbn:nl:ui:13-pv3u-84	732	352	4241	2599
12	10	International Crime Victims Surveys - ICVS - 1989, 1992, 1996, 2000, 2005	urn:nbn:nl:ui:13-wx0-h0o	709	568	3614	3026
13	8	Nationaal Kiezersonderzoek, NKO 2002 2003	urn:nbn:nl:ui:13-hvz-17u	692	616	1330	1197
14	9	WoON2012: release 1.0 – Woon-Onderzoek Nederland 2012 (voor overheid, universiteiten en overige partijen)	urn:nbn:nl:ui:13-60fd-6i	628	603	5139	5043
15	13	Tijdsbestedingsonderzoek 2005 - TBO 2005	urn:nbn:nl:ui:13-v64-rd7	596	454	4115	3424
16	83	EURISLAM Survey-data & Codebook	urn:nbn:nl:ui:13-tk19-qq	550	137	1284	451
17	NEW	Anthropogenic land-use estimates for the Holocene; HYDE 3.2	urn:nbn:nl:ui:13-clts-tg	537	-	23850	-
18	15	Arbeidsaanbodpanel 1985 t/m 2010	urn:nbn:nl:ui:13-4js-jl3	468	430	6772	6577
19	18	Slachtoffers van oplichting en van poging tot oplichting (projectnummer 1742)	urn:nbn:nl:ui:13-h84-sjz	465	368	972	750
20	NEW	The SWELL Knowledge Work Dataset for Stress and User Modeling Research	urn:nbn:nl:ui:13-kwrv-3e	422	-	71974	-
21	16	Culturele Veranderingen in Nederland 2006 - CV'06	urn:nbn:nl:ui:13-73o-mbh	422	387	1547	1488

Rank		Title of Dataset	Persistent Identifier	Dataset Downloads		Downloaded Files	
2019	2017			2019	2017	2019	2017
22	17	Enquête Beroepsbevolking - EBB - jaargangen 1987 t/m 2012	urn:nbn:nl:ui:13-sk6-fmg	421	381	3840	3521
23	22	Cohortonderzoek Onderwijs-loopbanen van 5-18 jaar - COOL 5-18 - Basisonderwijs 2007/08	urn:nbn:nl:ui:13-icz-r75	415	314	3414	2757
24	24	Cohortonderzoek Onderwijs-loopbanen van 5-18 jaar - COOL 5-18 - Basisonderwijs 2010/2011	urn:nbn:nl:ui:13-75t7-yy	395	283	3201	2461
25	48	Cohort Differences in Big Five Personality Factors Over a Period of 25 Years	urn:nbn:nl:ui:13-cf2d-zr	379	185	941	497

Table 1. Top 25 of downloaded datasets since 2007

Table 1 gives the top 25 reused datasets since 2007 in December 2019, compared to August 2017, both in terms of dataset downloads and in terms of downloaded files. Although the changes in popularity of datasets seem to be less great than those in pop music hit parades, the shifts are clear: the top 1 and 2 change places; there are 2 new entries, there are 12 climbers and 12 fallers. Most of the "greatest hits" in the top 25 are from the domain of the social sciences. If we took the number of individual files as the criterion for reuse instead of datasets, the top list would look rather different. Then the NLGis shapefiles would be the winner with over 165,000 files downloaded. This dataset contains the municipal boundaries of the Netherlands since the early 19th century till the present time. Geodata and archaeological data often consist of many different files, whereas most social science datasets in the top 25 usually consist just of a few files.

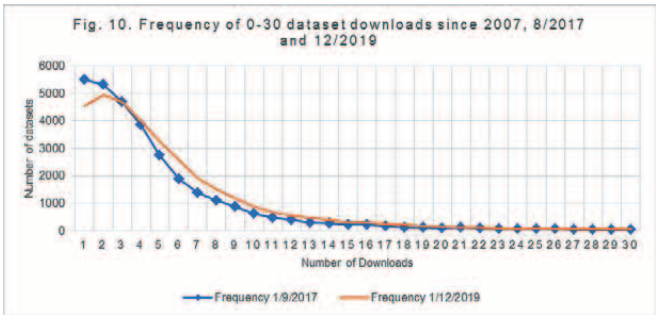

Fig. 10. Frequency of 0-30 dataset downloads since 2007, 8/2017 and 12/2019

When we look at the least frequently downloaded datasets (1-30 downloads since 2007), the numbers drop in an asymptotic way (fig. 10). It is noteworthy that over time, comparing 2019 with 2017, the curve tended to become a bit flatter, as the number of downloads per datasets tends to go up (especially for datasets with 5-10 downloads).

DISCUSSION

Borgman et al. (2019)[1] carried out a qualitative research into the reuse of datasets archived at DANS. They conducted 28 interviews with different users of the DANS archive and found a remarkable variety of uses. The information from the interviews was enhanced with weblogs, ethnography, and document analyses, revealing that a few large contributors provide a steady flow of content, but most individual depositors are academic researchers who submit datasets infrequently and often restrict access to their files. Data consumers are a diverse group that overlaps only marginally with the depositors. The usage appears to be typically infrequent and diverse. The aims of locating and downloading data varies not only across disciplines and data types, but also depends on the characteristics of the user: there are students, researchers, museum curators, employees of private companies, government employees, etc.

Perhaps the ultimate justification of maintaining a data archive is in the degree to which reused data leads to new knowledge creation, which might be counted as scholarly publications based on the downloaded data. However, it appears that data citation practices have less permeated academic cultures than the guidelines for such citation might suggest. Finding back reused data in the academic literature appears to be very hard. References to persistent identifiers are still the exception rather than the rule, and title information of a dataset is not as fixed as the title of an article or monograph. Moreover, there is not yet much research on this subject. Piwowar and others (2011)[2] suggest that "data archiving gives a high return on investment", but the underlying evidence is thin. In the Netherlands, Tessa Pronk and colleagues reported on their attempts to trace back datasets from a number of repositories in the literature, and concluded (Pronk et al, 2017)[3]:

- sharing of data for reuse often takes place in an informal manner;
- shared sets are not centrally registered;
- reuse is not registered by default;
- data citation is not yet taking place on a large scale.

1 Ch. L. Borgman, A. Scharnhorst, M. S. Golshan: 'Digital data archives as knowledge infrastructures: Mediating data sharing and reuse'. Journal of the Association for Information Science and Technology (ASIS&T), 70(8) 2019: p. 888-904. First published: 24 January 2019. https://doi.org/10.1002/asi.24172.

2 H. Piwowar, T. Vision, M. Whitlock: 'Data archiving is a good investment'. Nature 473, 285 (2011) https://doi.org/10.1038/473285a.

3 T. Pronk,. et al. (2017) Rapport hergebruik onderzoeksdata van Nederlandse universiteiten in kaart. UKB werkgroep Research Data. Versie 31 januari 2017. https://wiki.surfnet.nl/download/attachments/10875068/Hergebruik.%20Seminar%20The%20Making%20of%20RDM%20Policy%20Wageningen%201Dec16.pdf?version=1&modificationDate=1488474114142&api=v2 (accessed 1/12/2019).

ARCHIVIERUNG UND VERWALTUNG VON FORSCHUNGSDATEN – DATENDIENSTE FÜR DIE GEISTES- UND SOZIALWISSENSCHAFTEN UND DARÜBER HINAUS: DANS IN DEN NIEDERLANDEN

Data Sharing ist zu einer Standardanforderung geworden, die von einer wachsenden Zahl von Forschungsförderern und forschenden Organisationen gestellt wird. Das Adagio von heute ist, dass Daten auffindbar, zugänglich, interoperabel und wiederverwendbar sein sollten, und zwar auf eine so offene Weise wie möglich. Das ist die Theorie. Aber inwieweit dient dies den Bedürfnissen der Nutzer? Inwieweit werden Daten, die zur Weitergabe angeboten werden, tatsächlich wiederverwendet? Und inwieweit tragen „alte" Daten zur Schaffung neuen Wissens bei? Tatsächlich ist nicht viel über die Wiederverwendung von Daten bekannt, und noch weniger darüber, wie diese Wiederverwendung zu neuen wissenschaftlichen Erkenntnissen führt. Der Kern dieses Papiers ist die Nutzung eines nationalen Datenservice, der das Repositorium des Data Archiving and Networked Services (DANS) in den Niederlanden als Fallstudie nimmt und einen quantitativen Überblick für das Jahrzehnt 2007-2019 bietet.

Dr. Peter K. Doorn
Direktor DANS
Data Archiving and Networked Services (DANS)
Anna van Saksenlaan 51, 2593 HW Den Haag
The Netherlands
E-Mail: peter.doorn@dans.knaw.nl