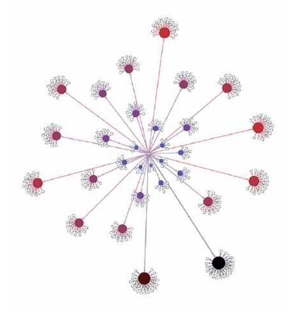
E-DATA & RESEARCH

Special issue 2014/2015

Newsletter on data and research in the social sciences and humanities.

E-data & Research is published three times a year. This special issue is sponsored by DANS.

CONTENTS



Use your smartphone to scan this QR code and visit our website **edata.nl**

The Dutch Data Prize is all about shared data

And the winner is...

Not everyone is aware of the importance of sharing research data. The Dutch Data Prize gives us the opportunity to express our appreciation for researchers who share the data they have collected, processed, curated and made available for others to use.

Drawing Auke Herrema

Both the quantity and the quality of the nominations have grown since the Data Prize was first awarded in 2010. All kinds of researchers and research groups want to demonstrate the painstaking work they have carried out to create wonderful data for innovative research. Of course there can only be one winner, although in a sense all data-sharing researchers are winners. They deserve more credit for making data available to the community, since this is arguably more important than the umpteenth publication in a formally peer-reviewed journal.

Great to see

The Dutch Data Prize is an initiative taken by Research Data Netherlands (RDNL), the national coalition of organizations with a mission to promote long-term archiving and reuse of research data. Although RDNL is a young and growing consortium, it builds on 50 years of data archiving and data sharing in the Netherlands. We have come a long way and it is great to see so many researchers sharing their data and encouraging others to do so too.

International data awards

Jury members had the difficult task of evaluating 30 nominations in the humanities and social sciences, and 16 nominations in the technical, natural and life sciences, all from the Netherlands. Would it not be great to take this Dutch initiative to an international level? The Digital Preservation Coalition in the UK has Digital Preservation awards; perhaps the Research Data Alliance could honour the champions of open research data by establishing international data awards.

This E-data & Research special edition is an initiative of DANS, but above all it is a co-production of people around the globe who express their vision of our joint challenges. Let's consider these challenges. Let's discuss the issues together. Today, tomorrow, this week, and in the months and years to come.

peter.doorn@dans.knaw.nl

Learn more about Horizon 2020 and Science 2.0

'Big data is definitely here to stay; the question is how we deal with it'

Learn more about Horizon 2020 and Science 2.0 from Robert-Jan Smits, Director-General of the European Commission's DG Research & Innovation.

How important is European research and innovation for researchers and their data?

"Affordable and easy access to the results of the research we fund is important for the scientific community and for innovative businesses. What is at stake is the speed of scientific progress and the return on R&D investment which has enormous potential for boosting productivity, competitiveness and growth. Therefore, all projects receiving Horizon 2020 funding – our new €80 billion research and innovation programme – will have the obligation to make sure peer reviewed publications (primarily journal articles) are openly accessible, free of charge. As for open access to research data, we have launched a limited pilot action in selected areas of Horizon 2020. Other projects outside these core areas may join the pilot on a voluntarily case-by-case basis. However, we also recognise that there are good reasons for not making data openly available and all projects therefore also have the opportunity to opt out of the pilot at any stage."

What are the obstacles for the ideal European research and innovation environment?

"With only 7% of the world population, Europe is still producing one third of the world's knowledge. This is very impressive for such a small continent. However, to remain competitive with our science base, we need to tackle the existing barriers that hamper our science and innovation system. For this reason, the EU has launched the Innovation Union, one of the flagship initiatives of the EU's growth strategy for the decade, Europe 2020, which is aimed at putting in place the right framework conditions such as the unitary patent, faster standard setting and a European Passport for venture capital funds."

What is the advantage of Horizon 2020 for researchers? How can they get involved?

"Horizon 2020 brings together all of the Union's funding for research and innovation. It offers researchers more opportunities for collabora-tion and promises more breakthroughs, discoveries and world-firsts by taking great ideas from the lab to the market. It couples research and innovation in many different areas to find solutions to the grand societal challenges for citizens. It is open to all researchers of the world, has a simple structure and red tape has been reduced dramatically."

Is the data explosion of the last few years a hype or are the data here to stay?

CONTINUED ON PAGE 3

Sharing data: good for science, good for you >>

1. Neelie Kroes, EU Commissioner for the Digital Agenda: "My vision is clear and straightforward. Open access to scientific information. And this will allow all of us to get the most benefit from science. For education, for innovation and for whatever creative reuse people can think of."

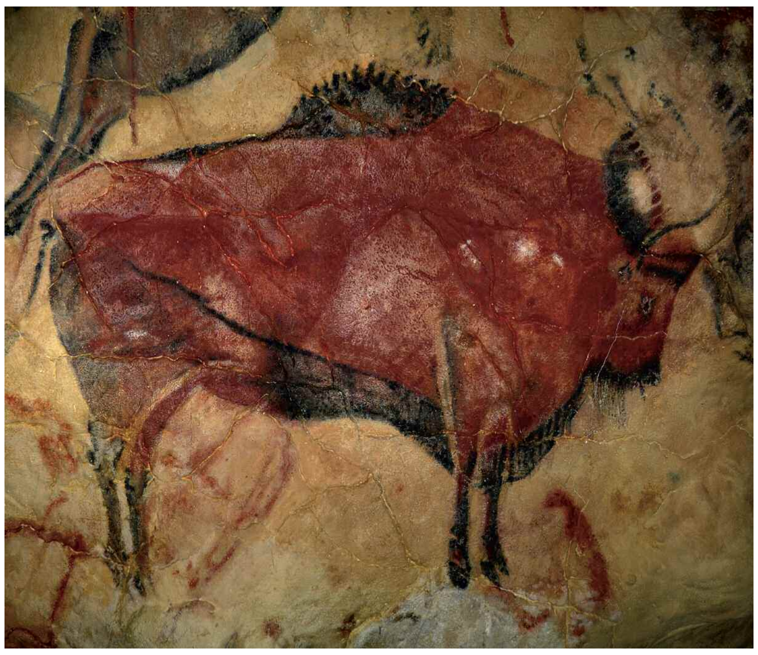
2. All scientists put a lot of effort into collecting and processing data. And quite rightly so. But what happens next, after publication? How do they store their research data? Are they simply stashed away somewhere, casually? Or stored permanently, accessibly?

From punch cards to online access to thousands of datasets

Fifty years of Dutch data archiving

Fifty years ago the first data archive in the Netherlands within the humanities and social sciences was established. Since then, a lot has changed, but the memories of the predecessors of DANS are still there.

HISTORY



An Altamira Bison; this cave painting from 30,000 years ago is still accessible and readable today. Will this also be true for the digital data recorded by us? It's the core question of digital data archives.
credit http://en.wikipedia.org/wiki/Cave_painting#mediaviewer/

1964 - Steinmetz Foundation

How to prevent valuable research data from being lost? Fifty years ago, the Steinmetz Foundation was set up to solve this issue. A few years later the Foundation was taken up by the Royal Dutch Academy of Sciences (KNAW). Marion Wittenberg, who works at the Steinmetz, remembers a story of a former colleague: "Punch files were taken to the University of Amsterdam's computer centre, SARA, on the back of a bike. There they were fed into a machine to be read, which sometimes went wrong. The cards would spray like a fountain from the card reader, warped or cracked, no longer reproducible. In the metadata, deviation rates were estimated. The punch cards were replaced with magnetic tapes, which were at first stored on SARA's mainframe computers, and later in the Steinmetz vault on Herengracht, with a back-up copy at SARA. In the early nineties, all data were copied to Optical Disks. The back-up CD-ROM was kept in Friesland, under the bed of one of the employees. If the bomb were to be dropped on Amsterdam, the data would not be lost. Ultimately, all of the data ended up in the EASY system at DANS."

She continues: "One of the aims of the Steinmetz Archive was that the collection could be reused. Director C.P. Middendorp was active in that respect. He replicated some 200 survey questions from 15 studies in his research on cultural changes in the Netherlands. This study from 1975 served as the basis for the biennial Cultural Changes survey by the Netherlands Institute for Social Research (SCP), which is still being held."

1989 - NHDA

In 1989, the Netherlands Historical Data Archive (NHDA) was established. Heiko Tjalsma was involved from day one: "The NHDA was founded in Leiden. In a way, we followed in the footsteps of the existing Steinmetz Archive, but for the historical sciences. From the start there were differences, however. Until 1995, when it became part of the Royal Academy, the NHDA was not an established institute but rather a project with an uncertain future. Funding would vary from year to year. As a consequence, the NHDA with its small staff engaged in any work related to 'history and computing,' even if only remotely. We had to, as we were almost entirely dependent on external project income. In particular, many digitizing projects were carried out. We did little data archiving as there was no prospect of keeping data longer than say a year. Sometime later we also played a leading role in a postgraduate programme funded by the European Social Fund set up to train graduates in historical information science. All in all the NHDA could best be described as a very creative, slightly anarchistic but always stimulating organisation in which multitasking was the normal work attitude." The NHDA was integrated in NIWI in 1997.

1994 - WSA

The Scientific Statistical Agency (WSA) was established by the Board of NWO in 1994 to improve the accessibility and availability of data files. The Agency played an intermediary role between researchers and data providers, and worked closely with other institutions mediating the unlocking of data for scientific research. Ron Dekker was head of WSA from 1997-2002. He remembers: "In 2001 WSA, NHDA and the Steinmetz Archive organized the international IASSIST data conference in Amsterdam. In one week we received more than 200 participants from 20+ countries, including (for the first time) from Eastern Europe. That week, there was a trip – yes, to the windmills!, there were workshops, plenary presentations and parallel sessions. This was in the pre-wifi era and the internet connections were literally laid through Amsterdam: from the UvA auditorium to the Doelenzaal and Singelkerk across the canal. We opened with a keynote from Paul Schnabel who quoted from the Odyssey. When Odysseus blinded the Cyclops, the latter asked who he was and Odysseus replied, "Nobody." Here we see the first instance of anonymized data because when the other Cyclops asked who had done it, the first one answered "Nobody" and Odysseus came out unscathed. Nero, no not the emperor but a Steinmetz employee, had bought the original soundtrack of the film. Later Kubrick decided to use familiar music, but the soundtrack was already there – a nice bit of archiving! That collaboration of WSA, NHDA and Steinmetz anticipated the decisions made later. We were collaborating, but DANS did not come about until 2005. And the rest is history and, above all, a bright future."

Computer use by historians in 1989, photo taken from the book 'Toverwoord informatie'
credit Maili Blauw

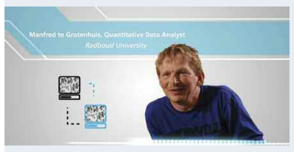
2004 - EDNA

The e-depot for Dutch archaeology, EDNA, aims at the sustainable archiving and disclosing of archaeological research data and the dissemination of related knowledge. EDNA is a collaboration between DANS and the Cultural Heritage Agency of the Netherlands. Milco Wansleeben was involved in the creation of EDNA: "From the very beginning, when we had a small website with four datasets and a modified version of NIWI's Itor software, Dutch archaeologists could experience for themselves what exactly a data archive was, how it worked and what they could do with it. This visibility has contributed to the awareness that digital research will not naturally remain available, and it has ultimately been the key to the inclusion of a national deposit requirement in the quality standard for Dutch archaeology. Demonstrators and pilots did really help, and they still do. We went from 4, to 50, to 22,000 archaeological datasets in an e-depot that is now completely up and running. More and more archaeological organizations have left their initial reluctance about accessibility behind and open access seems to become the new standard for archaeologists."

2005 - DANS

In 2005, NWO and KNAW together founded Data Archiving and Networked Services (DANS). DANS took over the activities of the Steinmetz Archive, the NHDA, WSA and EDNA. Despite all the mergers, the data have been preserved over the last fifty years. Research files that started out as boxes full of punch cards in the early days of the Steinmetz Archive, can still be accessed through the EASY system today. Some of them are being downloaded regularly. Peter Doorn, director of DANS: "DANS promotes sustained access to digital research data. To this end, we encourage researchers to employ sustainable data archiving and reuse, for example through our online archiving system, EASY. And with NARCIS, for example, we provide access to thousands of scientific datasets, e-publications and other research information in the Netherlands. We never know what the future brings, but today we want to celebrate the success of 50 years of data archiving in the Netherlands."

Sharing data: good for science, good for you >>



3. Dutch historian Martijn Kleppe (EUR): "Such data sets tend to be very rich, there are so many articles you could write. Now that I've finished my thesis, my interest in other research areas is growing while my orginal data set still contains useful information. Sure, I can store it in my drawer or on my computer, but that's not going to be around five years from now. Or, alternatively, I can store it at DANS, so that people might extract relevant information."
4. Data that's worth analyzing, is worth storing. For yourself and for your successors And it's not hard, if you plan it properly, up front. The past is a great source of future knowledge. Data archives can be real treasure troves.
5. Quantitative data analyst Manfred te Grotenhuis (RU): "It feels like walking around in a candystore, just taking what could be of use to us. The access speed is a huge advantage, direct access without mediation, so you can start immediately, that's what I find enjoyable."

COLOPHON Publisher: Stichting Uitgeverij E-data & Research Den Haag. **Editorial address**: Postbus 93067, 2509 AB Den Haag, +31 70 349 44 50, edata@dans.knaw.nl, edata.nl. **Editor in chief:** Heidi Berkhout. **Editorial secretariat**: Lucas Pasteuning. **Translator/proofreader**: Pieter Wijker. **Contributors**: Patrick J.C. Aerts, Almila Akdag Salah, Ashkan Ashkpour, Henk van den Berg, Cynthia Bogaert, Christine L. Borgman, Katy Börner, Olav ten Bosch, Leen Breure, Dan Brinkhuis (Science Media), Mercè Crosas, Ron Dekker, Ingrid Dillo, Peter Doorn, Elena Giglia, Manfred te Grotenhuis, Ashling Haslehurst, Auke Herrema, Denise Heiligers, Gary King, Martijn Kleppe, Albert Meroño Peñuela, Barend Mons, Randall Munroe, Jackie Middleton, Linda Reijnhoudt, Jeroen Rombouts, Andrea Scharnhorst, Robert-Jan Smits, Herbert van de Sompel, Heiko Tjalsma, Andrew Treloar, Milco Wansleeben, Marion Wittenberg, John Wood. **Layout** : Colette Sloots. **Production and press**: Quantes, Rijswijk. **Circulation**: 8,000. **ISSN**: 1872-0374. We have tried to approach all stakeholders regarding the images used. Those who believe they may assert rights can contact us.

CONTINUED FROM PAGE 1

Dealing with Big Data

"'Big data' is definitely here to stay; the question is how we deal with it: how will we store and curate the data? Who will be responsible for their management? How can we ensure interoperability? On these questions, the RDA can provide important input. Digital technologies are not only creating more data, they also provide us the tools to make sense of them. Text and Data Mining (TDM) can analyse and extract new knowledge and enables new research connections. A recent report indicates that prolific use of TDM would add tens of billions of euros in value to the EU's aggregate GDP."

What will the 'data world' look like in 2020?

Robert-Jan Smits, Director-General of the European Commission's DG Research & Innovation *credit EC DG Research & Innovation*

Science 2.0

Science 2.0 is a term used to describe the ongoing evolution in the modus operandi of doing research and organizing science. These changes in the dynamics of science and research are enabled by digital technologies and driven by the globalisation of the scientific community as well as the increasing societal demand to address the Grand Challenges of our times.

ec.europa.eu/research/consultations/science-2.0

"Systemic changes are taking place in the way research is performed and science is organized, which are sometimes referred to as Science 2.0. In 2020 the data world will be shaped by the main drivers of Science 2.0, which include amongst others the tremendous increase in the number of researchers, new emerging scientific powerhouses, the growing and increasingly pressing demand for solutions to grand challenges, or 'digital natives' becoming part of the researcher population."

http://www.iprhelpdesk.eu/Open_Access_in_H2020020.pdf.

COLUMN

"Let's always discuss what research must be like"

One day when my son was in his final undergraduate year at university he asked if I could tell him what it was like to do a PhD. Amazed that he was asking his father for advice I started to tell him what it was like for me and my students.

It was only afterwards that I realized I had not told him what I thought research would be like in the future. I was then prompted by a member of the European Commission to write up my thoughts on what research would look like in 20 years' time. Parts of this were subsequently published in Microsoft's Futures magazine and I thought no more about until I got a phone call from what was DG Infso to ask if I would chair a high level group on the future of scientific data. While I had led organizations that created and stored huge amounts of data from international experiments, I am not a computer or data scientist. Fortunately the members of the group were fantastic and we had tremendous support from the Commission in addition to contributions from a number of key witnesses. The result was 'Riding the Wave' or how to deal and cope with the tsunami of research data. Published at the end of 2010 it became a best seller since not only were the recommendations clear, but they resonated with other initiatives going on in other parts of the world. One of the key recommendations was to ensure there was a global forum to avoid different approaches by different regions of the world.

Further discussions took place between Europe and the USA and these culminated in a pre-meeting at the ICRI2012 meeting in Copenhagen where colleagues from other interested countries, including Australia, met to discuss future actions. Towards the end of it Alan Blatecky from the National Science Foundation stood up and said: "Let's get on with it!" After a year of intense activity the Research Data Alliance was formed between the USA, Europe and Australia. Right from the start it was agreed that the RDA was not an academic debating chamber but was about tangible results that could be used by the community. The sheer energy and momentum of the RDA has surprised all of us.

Professor John Wood, RDA Council Co-Chair, Secretary-General of the Association of Commonwealth Universities and former High Level Expert Group on Scientific Data Information Chair & European Research Area Board Chair, member of the Research, Innovation and Science Expert Group Science 2.0. *credit Association of Commonwealth Universities*

Is it a false hope that research around tackling global challenges will lead to a sustainable future?

Only time will tell but the involvement of groups looking at agriculture, marine and social issues among many gives hope that the cultural change of sharing in an Open Science and Open Innovation environment will lead to a sustainable future for the world. Tall ambitions, but essential if the earth is to sustain 9 billion people in 2050.

The original Riding the Wave report looked forward to 2030. It is likely that many of the recommendations will be in place before 2020. So it is time to look ahead even further. My granddaughter is just over a year old, what will research be like when she starts her PhD? Discuss!

john.wood@acu.ac.uk

Combined data services of partners

Secure research data with one click of a button

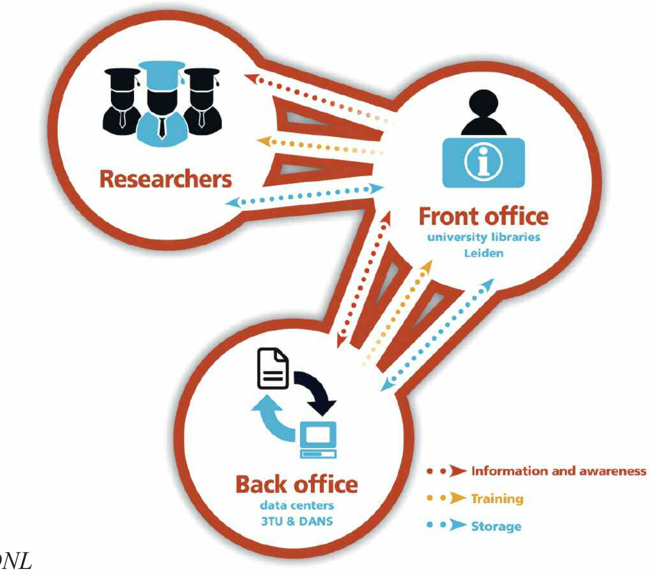
Research Data Netherlands (RDNL), founded in 2013, promotes sustainable archiving and reuse of research data. Jeroen Rombouts, working at one of the promoting organizations, tells us more about this 'open national coalition.'

For Dutch researchers it is getting increasingly easy to secure research data for future use. It may even require no more than one click of a button. That is, if it is up to the RDNL partners. Jeroen: "RDNL is a coalition of back office parties and currently made up of three partners: DANS, 3TU.Datacentrum and SURFsara. From RDNL, they offer their services to their customers as an integrated back office. If, for example, the customer is a university data library, the data library purchases data services from RDNL, which they then provide to their researchers. For staff members of research institutes this will lead to a more customer-friendly service offering and practice, particularly where cross-discipline and/or multi-supplier issues are at stake. Maintaining research project results could be entrusted to, for example, a single front office which records the data and then stores them in different repositories under similar conditions."

Ensured data value

He continues: "The partners benefit greatly from the coalition too. They complement each other through knowledge pooling and transfer. They also look for more efficient employment of their collective capacity. Moreover, cooperation improves the position of the individual parties at the national and international levels. Their main mission is to ensure the accessibility, usability and long-term availability of valuable data. To live up to their expecta-tions, RDNL will coordinate its roadmap with various stake-holders, including users, policy makers and research funders. The partners will then address these actions both together and with stakeholders from outside RDNL. 2015 will be an important year!"

J.P.Rombouts@tudelft.nl
researchdata.nl



credit RDNL

Sharing data: good for science, good for you >>

6. So durable and accessible data storage ought to be the norm really. It's a way to add longevity and value to your data. Still some scientists have their doubts - but why? "A lot of researchers are concerned that others will make off with their data. Take me, I have stored 4,000 catalogued photographs here, with a total of 20,000 parameter values. You might use it to write an interesting article, so why don't I write it myself?"
7. Marion Wittenberg, data manager at DANS: "After storing your data at DANS, you can decide how others may access your data."

Linked Open Data: taking a step forward

Researchers want to ensure that others can use their research data – now and in the future. Linked Open Data offers a solution.

Linked Open Data (LOD) is a way to connect data sources and publish them as one dataset online. The connections are provided by Application Programming Interfaces (APIs). Thanks to Uniform Resource Identifiers (URIs) new data sources can also connect to the existing datasets. Through LOD, data are thus always linked to other relevant data sources.

Using the same standards

International challenges, embodied in programs such as the EU's H2020 program, increasingly underline the need for research communities to work together. This also implies an increase in the number of data sources and data formats. Obviously, the internet has become the standard platform on which all parties exchange knowledge and expertise, also in the form of LOD. In order to be able to apply LOD in practice, it is crucial that those involved use the same standards. The international web standards body, the World Wide Web Consortium (W3C), leads the way in the development of these standards. Connected information sources in Linked Open Data format are available.

Reusability tools

Meanwhile web APIs are increasingly being used as 'data-reuse tools' (over 10k). Mendeley, Figshare and other platforms provide useful APIs for researchers. As an example, everyone can use the Mendeley API to retrieve publication data and combine them with data from Figshare.

A world of APIs

Want to know which APIs exist? ProgrammableWeb offers a list of web-based APIs. It's good to realise that most APIs have been designed to answer questions like "Which publications have 'data' in their title?" The LOD format enables a different type of API, allowing the retrieval of raw data. For example "Give me all the data on X where X is the URI of a resource". Some sources offer a mix of different types of APIs. Which API is most appropriate depends on the use case. However, using a web API is almost always better than retrieving a dataset file.

International public data

Together we must answer the question which standard(s) we need to create an API. Then we can go and warm data providers to the idea of connecting their data with other datasets. International collaboration is important. Global sharing and linking of public data should be a priority for all of us, with the aim of enabling everyone to do research using the same contextualized data.

christophe.gueret@dans.knaw.nl
linkeddata.org
programmableweb.com
dev.mendeley.com

KNOWeSCAPE: visualizing to know

Beautiful visualizations, often based on Big Data, can be found everywhere: in research, in science communication, museums and art galleries. To be able to make them is not longer the privilege of a small group of experts. However, if it comes to libraries and archives visualizing their collections and the artefacts in them, is far from being standard.

Visualizations can have many uses. For digital humanities researchers they are a useful method in their quest for the past of our culture. For the libraries and archives providing means for this quest, visualizations can be used to highlight special treasures in a collection, to support navigation through a collection and to monitor its use.

A European network of collaboration – KNOWeSCAPE – has been set up to foster the development and implementation of so-called knowledge maps for archives and libraries. The network has initiated all of the examples presented below. KNOWeSCAPE is a COST Action devoted to the analysis of the dynamics of knowledge spaces and map making for those spaces. It is unique in establishing a dialogue between data scientists, digital humanities scholars and information scientists and professionals.

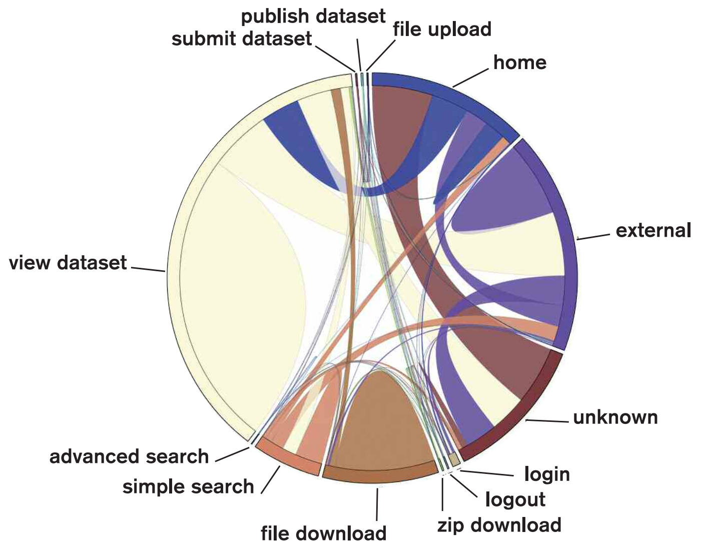
This page shows examples for three use cases of knowledge maps: research, navigation and curation.

The examples are taken from collections and services of DANS. Various visual elements are used: timelines, chord diagrams and complex network visualizations.

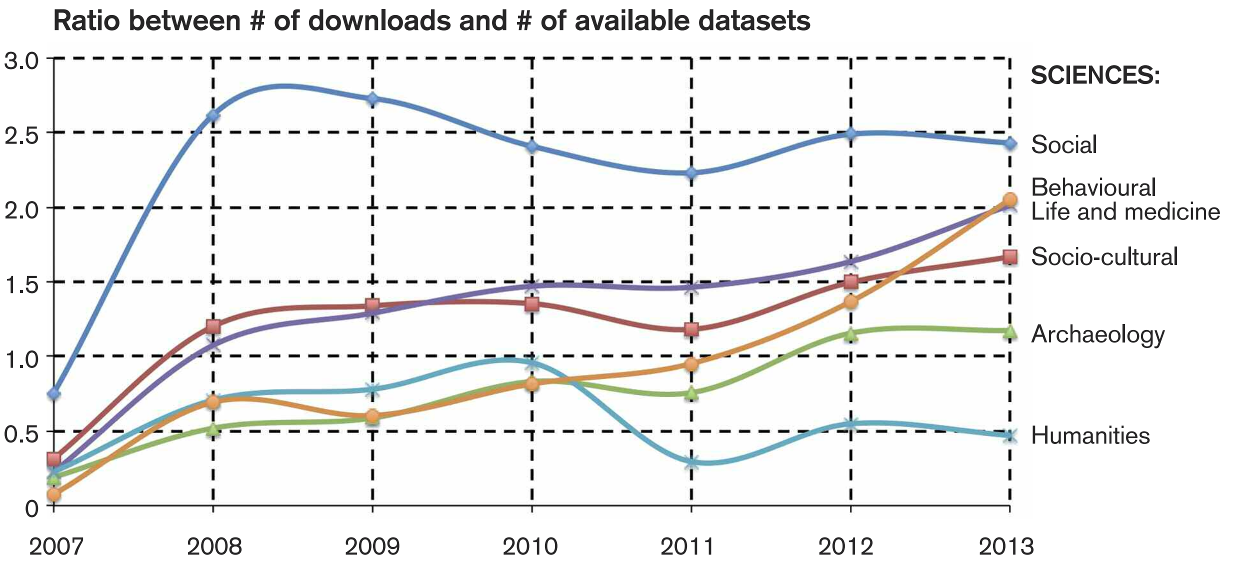
andrea.scharnhorst@dans.knaw.nl
www.knowescape.org

Curation

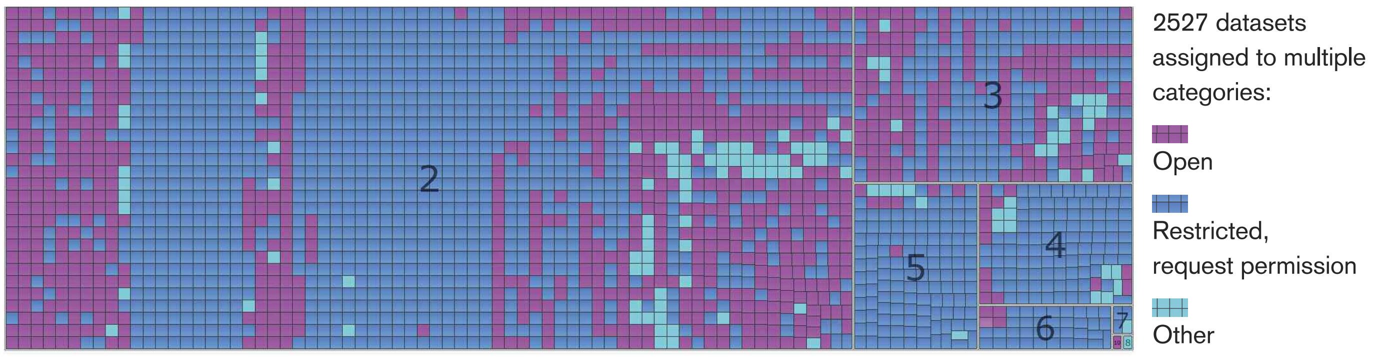
Each archive wishes to expand its collection for future use. At the same time, it also has to prove its present usefulness. DANS holds collections of digital research data in EASY (easy.dans.knaw.nl), and information about the Dutch research landscape in NARCIS (narcis.nl). Visualizations help to manage the services. Creating them is also a great opportunity to check the cleanness of metadata.



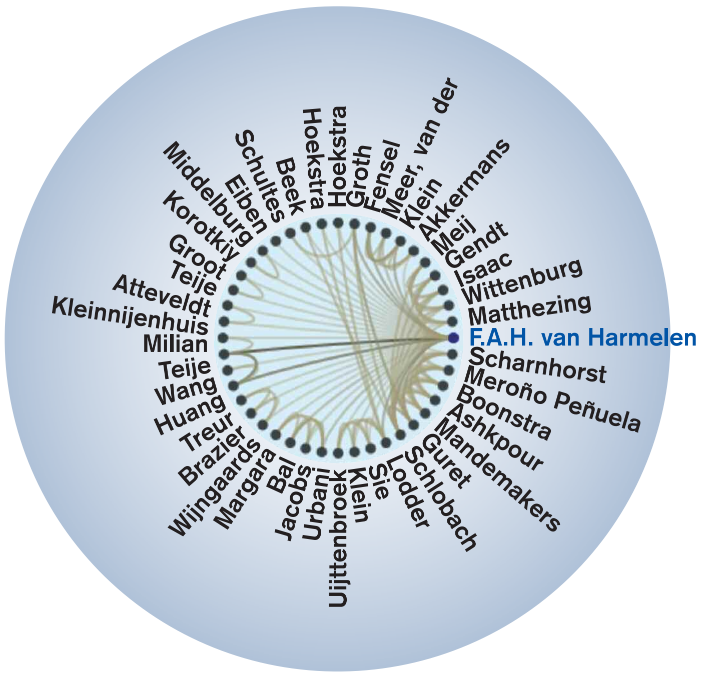
This chord diagram shows main paths of EASY users. Unsurprisingly, one sees a stream of transitions from external pages to views of datasets. What is remarkable is also that the advanced search option is hardly ever used. This kind of history of information seeking can be used to get insight into user behaviour as well as the user interface.



These timelines show that there is no direct correlation between demand and supply if it comes to archival information. Year after year more datasets have been deposited in EASY. They are tagged for different audiences, so we can arrange them into specific data set collections. In sheer numbers, archeology (EDNA, www.edna.nl) has the biggest collection. But if we divide the yearly download of datasets by the number of available datasets, we see that smaller collections - tagged as social sciences or behavioral sciences - are in greater demand. Collection building should be based not only on current demand from scientific communities, but also on the experience and imagination of the archivists who anticipate future needs.



Datasets in EASY are tagged with possible audiences (subject categories) and grouped into collections. The tree map lists all datasets with more than one audience. It is an interactive map in which each rectangular represents one dataset and its color indicates for open or restricted access. (see http://drasticdata.nl/DDHome.php?m=514)



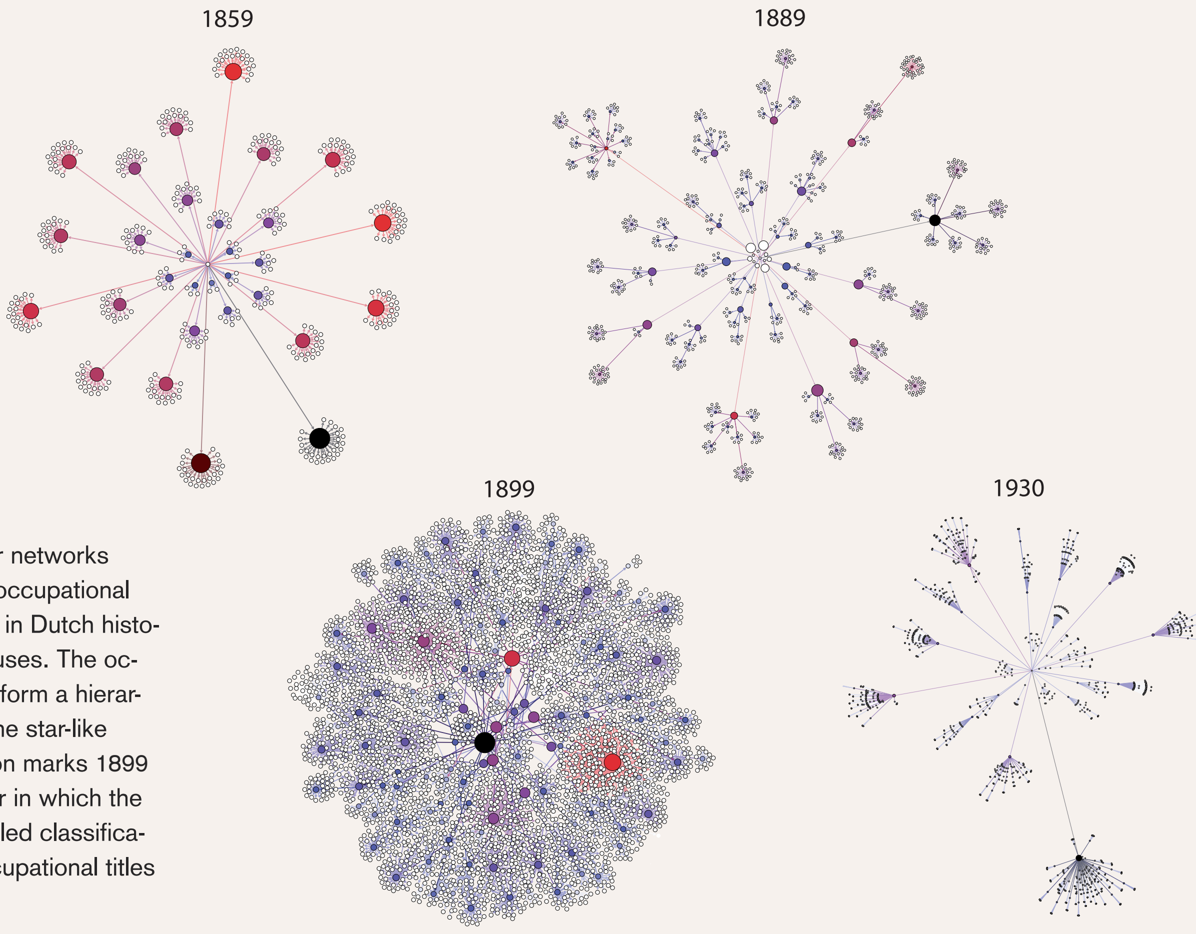
NARCIS holds information about researchers in the Netherlands, including all full professors (about 8000). The circular display of their collaborative network - based on publications and projects registered in NARCIS - allows this collaborative space to be browsed by clicking on collaborators.

Navigation

In the age of digital libraries, Online Public Access Catalogues (OPACs) determine how we see the collections of a library or an archive. Imagine a library user walking along open stacks or browsing with her fingers through cardstack drawers of subject catalogues, and compare this with looking at a web-interface and lists of retrieved documents. You immediately understand what is missing in digital access. Visualizing a collection as a whole is one way to close the gap between a physical encounter with a collection and browsing a collection online. But what would a user like to see: the growth of a collection over time or the interconnectedness of search terms and their possible targets? This can only be found out by experimenting with interactive visual interfaces.

Research

Digital humanities scholars and the general public are prominent users of collections in libraries and museums. Research visualizations are a way of getting a grip on the increasing amount of data and information that is digitally available. DANS supports the creation of such scientific visualizations, also to showcase some of its most interesting treasures.



These four networks show the occupational titles used in Dutch historical censuses. The occupations form a hierarchy, and the star-like visualization marks 1899 as the year in which the most detailed classification of occupational titles was used.

Automated sharing of research data

The Dataverse Network Project is all about control, visibility and academic credit. Mercè Crosas and Gary King tell us more about this solution for long-term data sharing and data preservation.

The archive unquestionably deserves to be thanked, but the author needs to get credit too or there will be little motivation for them to deposit the data in the first place. Alternatively, researchers can distribute the data themselves, and guarantee that the data only go to people who will appropriately acknowledge the author. However, this means that long-term preservation is much less likely.

Brand your own Dataverse

The Dataverse Network Project is an open-source web application designed to resolve this important tension in data sharing. The software allows researchers to brand their Dataverse as their own, retain control and gain credit for their data, while providing support for good data management practices through automated metadata extraction and reformatting, version control, standards and APIs. It allows journals to integrate data deposit with article submission and provides a persistent link between article and data.

Free worldwide service

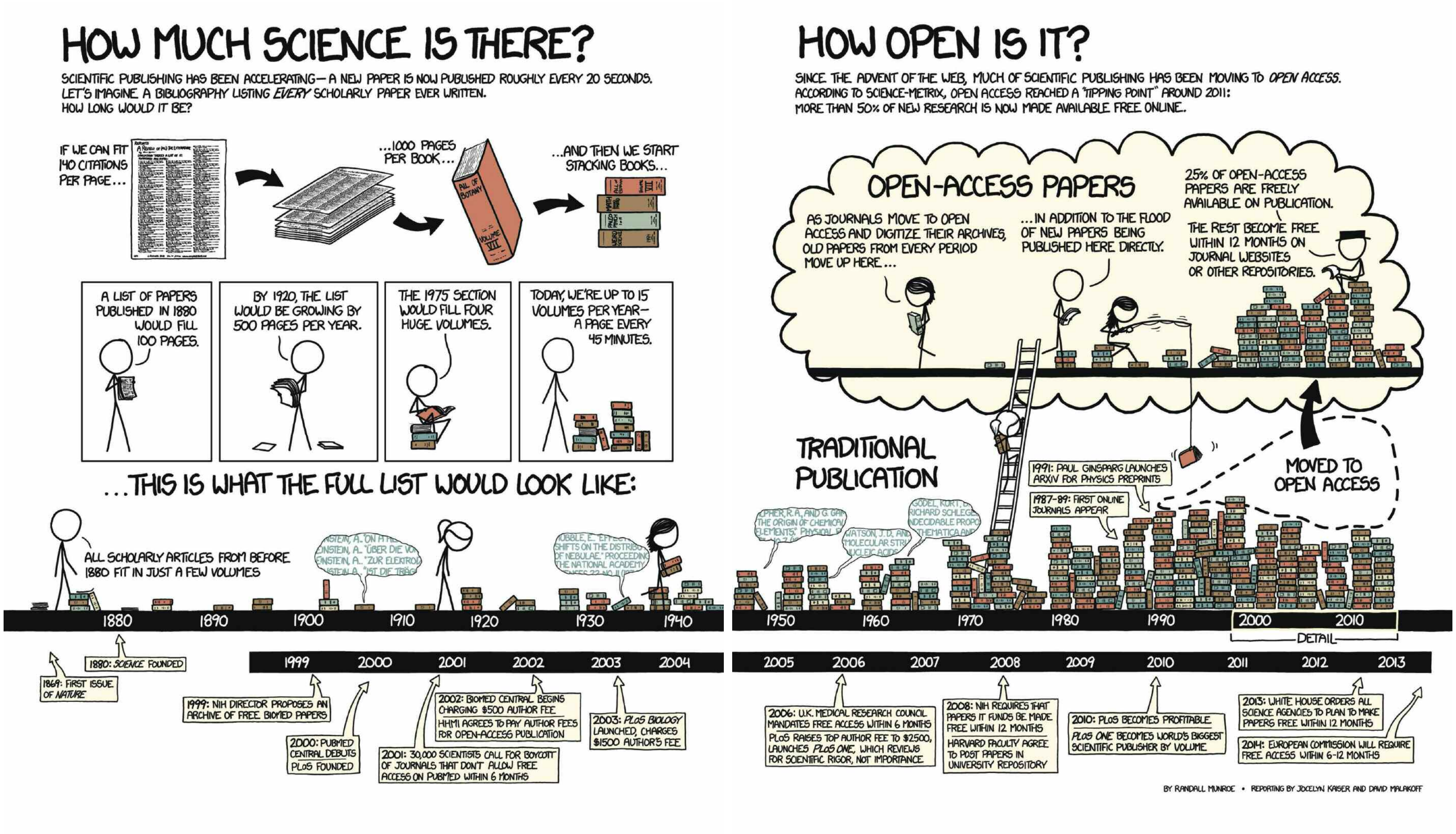
The Dataverse software is developed at Harvard's Institute for Quantitative Social Science (IQSS) and hosted by Dataverse Networks around the world (such as the Dutch Dataverse Network). The Harvard Dataverse offers a free service, open to researchers from all disciplines, worldwide. Dataverse 4.0, which will be released before the end of 2014, includes an improved user experience based on iterative feedback from the Dataverse community.

The next challenge

Reusing and validating scientific results derived from sensitive and big data are key next challenges in data-driven science. 2015 will bring sensitive data support through secure storage and integration with DataTags, an application developed at IQSS in collaboration with a multidisciplinary NSF-funded project (privacytools.seas.harvard.edu/) that enables sharing of such data. DataTags helps depositors assess risk based on legal requirements, then assigns a tag that informs the repository how data must be managed and accessed. The future also includes support for large datasets (TB-PB scale).

mcrosas@iq.harvard.edu
king@harvard.edu
http://thedata.harvard.edu

Mercè Crosas is Director of Data Science, Institute for Quantitative Social Science, Harvard University.
Gary King is Albert J Weatherhead III University Professor; and Director, Institute for Quantitative Social Science.


HOW MUCH SCIENCE IS THERE?
SCIENTIFIC PUBLISHING HAS BEEN ACCELERATING—A NEW PAPER IS NOW PUBLISHED ROUGHLY EVERY 20 SECONDS. LET'S IMAGINE A BIBLIOGRAPHY LISTING EVERY SCHOLARLY PAPER EVER WRITTEN. HOW LONG WOULD IT BE?
IF WE CAN FIT 140 CITATIONS PER PAGE…
…1000 PAGES PER BOOK…
…AND THEN WE START STACKING BOOKS…
ALL OF BOTANY VOLUME VII
A LIST OF PAPERS PUBLISHED IN 1880 WOULD FILL 100 PAGES.
BY 1920, THE LIST WOULD BE GROWING BY 500 PAGES PER YEAR.
THE 1975 SECTION WOULD FILL FOUR HUGE VOLUMES.
TODAY, WE'RE UP TO 15 VOLUMES PER YEAR—A PAGE EVERY 45 MINUTES.
…THIS IS WHAT THE FULL LIST WOULD LOOK LIKE:
ALL SCHOLARLY ARTICLES FROM BEFORE 1880 FIT IN JUST A FEW VOLUMES
EINSTEIN, A. "ÜBER DIE VO… EINSTEIN, A.. "ZUR ELEKTROD…
HUBBLE, E. "EFFECT… SHIFTS ON THE DISTRIB… OF NEBULAE." PROCEEDING… THE NATIONAL ACADEMY…
1880 1890 1900 1910 1920 1930 1940
1869: FIRST ISSUE OF NATURE
1880: SCIENCE FOUNDED
1999 2000 2001 2002 2003 2004
1999: NIH DIRECTOR PROPOSES AN ARCHIVE OF FREE BIOMED PAPERS
2000: PUBMED CENTRAL DEBUTS PLOS FOUNDED
2001: 30,000 SCIENTISTS CALL FOR BOYCOTT OF JOURNALS THAT DON'T ALLOW FREE ACCESS ON PUBMED WITHIN 6 MONTHS
2002: BIOMED CENTRAL BEGINS CHARGING $500 AUTHOR FEE
HHMI AGREES TO PAY AUTHOR FEES FOR OPEN-ACCESS PUBLICATION
2003: PLOS BIOLOGY LAUNCHED, CHARGES $1500 AUTHOR'S FEE
HOW OPEN IS IT?
SINCE THE ADVENT OF THE WEB, MUCH OF SCIENTIFIC PUBLISHING HAS BEEN MOVING TO OPEN ACCESS. ACCORDING TO SCIENCE-METRIX, OPEN ACCESS REACHED A "TIPPING POINT" AROUND 2011: MORE THAN 50% OF NEW RESEARCH IS NOW MADE AVAILABLE FREE ONLINE.
OPEN-ACCESS PAPERS
AS JOURNALS MOVE TO OPEN ACCESS AND DIGITIZE THEIR ARCHIVES, OLD PAPERS FROM EVERY PERIOD MOVE UP HERE…
…IN ADDITION TO THE FLOOD OF NEW PAPERS BEING PUBLISHED HERE DIRECTLY.
25% OF OPEN-ACCESS PAPERS ARE FREELY AVAILABLE ON PUBLICATION.
THE REST BECOME FREE WITHIN 12 MONTHS ON JOURNAL WEBSITES OR OTHER REPOSITORIES.
TRADITIONAL PUBLICATION
ALPHER, R.A., AND G. GAM… THE ORIGIN OF CHEMICAL ELEMENTS." PHYSICAL R…
WATSON, J.D. AND… MOLECULAR STRU… NUCLEIC ACIDS
GÖDEL, KURT… RICHARD SCHLEGE… UNDECIDABLE PROPO… MATHEMATICA AND…
1991: PAUL GINSPARG LAUNCHES ARXIV FOR PHYSICS PREPRINTS
1987-89: FIRST ONLINE JOURNALS APPEAR
MOVED TO OPEN ACCESS
1950 1960 1970 1980 1990 2000 2010
DETAIL
2005 2006 2007 2008 2009 2010 2011 2012 2013
2006: U.K. MEDICAL RESEARCH COUNCIL MANDATES FREE ACCESS WITHIN 6 MONTHS
PLOS RAISES TOP AUTHOR FEE TO $2500, LAUNCHES PLOS ONE, WHICH REVIEWS FOR SCIENTIFIC RIGOR, NOT IMPORTANCE
2008: NIH REQUIRES THAT PAPERS IT FUNDS BE MADE FREE WITHIN 12 MONTHS
HARVARD FACULTY AGREE TO POST PAPERS IN UNIVERSITY REPOSITORY
2010: PLOS BECOMES PROFITABLE
PLOS ONE BECOMES WORLD'S BIGGEST SCIENTIFIC PUBLISHER BY VOLUME
2013: WHITE HOUSE ORDERS ALL SCIENCE AGENCIES TO PLAN TO MAKE PAPERS FREE WITHIN 12 MONTHS
2014: EUROPEAN COMMISSION WILL REQUIRE FREE ACCESS WITHIN 6-12 MONTHS
BY RANDALL MUNROE • REPORTING BY JOCELYN KAISER AND DAVID MALAKOFF

Drawing xkcd.com

The data archive of the Sudan census of 1973

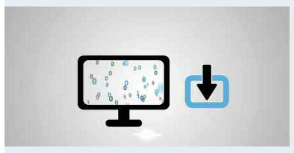
photo McCaa, Research Professor, Ambassador, IPUMS

Sharing data: good for science, good for you >>

8. All your data, preserved securely and sustainably. It adds extra value to your research, to your significance as a scientist. You even increase your chances of citations.

9. Martijn Kleppe: "Anyone who uses my data, should refer to me. That's another enjoyable reason to share my data."

10. Marion Wittenberg: "Also for verification. Researchers should make their data available so that others can check their modus operandi."
11. The repository had to meet requirements for certification and one of them is DSA, Data Seal of Approval.



12. Manfred te Grotenhuis: "You always want more. Ideally, all collected data should be available for further analysis. So, I would really like everybody to make their data available."

Enhanced data journal: next generation science

DANS is about to launch its own digital data journal. Leen Breure explains the initiative.

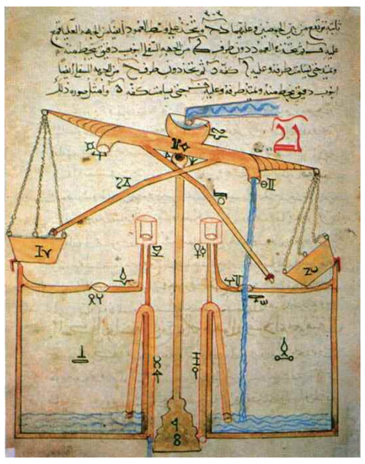
A digital data journal will contain data papers and reviews of datasets. A data paper is a scientific article describing data and their production against the background of the research project in which they were created. Some form of peer reviewing is usually part of the publication process. One essential property of such a paper is that it is formally cited and provides credits to the data creators, thus stimulating researchers to make their data more easily available to others.

Enhanced publication

The DANS data journal is an enhanced publication in more than one respect. The text is enhanced with direct links to datasets stored in the archived repository. Additionally, the journal is enriched with features that contribute to greater usability of the content in terms of overview and navigation by adding background information and various forms of visualization. Where possible, data can be previewed and explored online rather than through time-consuming downloads and offline applications. In short, an enhanced data paper provides an integrated view of data in their research context.

Mixed blessing of print

This enhancement could easily be considered a novelty that is essentially a luxury. After all, what really matters is some form of access to the dataset itself. However, looking back at the history of scientific publishing we see that integration of scholarly discourse and data – or evidence – has a long tradition. Remarkably, it has been the reproduction method which has often disturbed this integration. Medieval scientific manuscripts used to have diagrams and pictures, surrounded by notes and directly linked to the textual content. Printing created a problem: wood cuts are coarse and Renaissance anatomists and microscopists opted for 'high resolution' in the form of etching and engraving. But this so-called intaglio printing could not be directly combined with the relief printing used for text. The illustrations were therefore printed on plates at the end (or in the middle) of the publication, separated from the body text. Something similar happened to data.



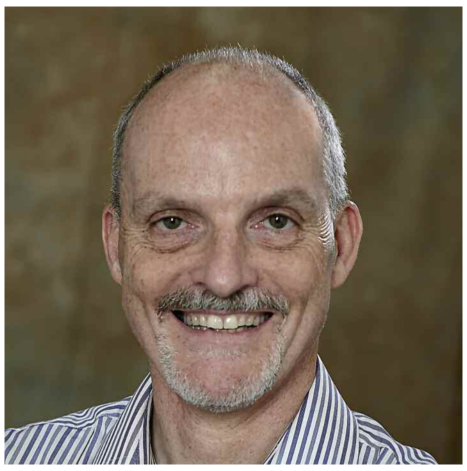
Picture of Al-jazari water device wtih Arabic handwriting
credit commons.wikimedia.org/wiki/File:Al-jazari_water_device.jpg#mediaviewer

In early reports of scientific experiments text is mixed with large tables and data lists (statistical graphics are a late 18th-century invention). At that time historians used to create disproportionate footnote sections with lengthy source citations. Massive computerization created not only much more data, but also posed the problem of relating printed and digital information. Finally, by fully embracing digital publishing we are now able to do justice to this natural coherence of text and data and take the next step in an old tradition.

Leen.breure@dans.knaw.nl

Christine L. Borgman *photo UCLA*

Katy Börner *photo edu.com*



Andrew Treloar *photo Treloar*

Herbert van de Sompel *photo Elena Giglia*

What does the world look like when sharing data is common practice?

The vision of visiting fellows

Every year DANS invites visiting fellows to contribute to the Dutch research data landscape by sharing their knowledge. In this article they answer three questions:

1. *Looking at your own daily scholarly practice, what is the aspect which frustrates your work the most, which you strongly dislike?*

2. *And what do you find most exciting?*

3. *If you were completely independent and able to create your ideal science/research lab – what would it look like?*

Christine L. Borgman:

“The ideal lab is one that brings the many stakeholders together”

1. I most dislike the administrative work, especially the lack of staff in public universities to handle some of the most basic duties. Then I remind myself that I am privileged to hold a professorial post in one of the world’s finest public educational institutions, the University of California.

2. Most exciting is the opportunity to influence practice and policy in the management of research data. My research and teaching interests have converged in the study of data and data practices, exploring how observations, models, artefacts and software become data; how these practices vary by individual and by discipline; and how these findings can be employed in the design of data collection, data management, data archiving and science policy. Factors contributing to the value of research data include the transition from print to electronic publishing, the ability to acquire and analyse large volumes of digital content in the sciences and humanities alike, and policies that promote openness and transparency.

3. On the surface, open access to data appears to offer vast benefits for research, education and innovation by leveraging public investments in these areas. Public policy documents suggest that releasing data is an easy task to be accomplished at the time of publishing articles or books, and that research data are yet another genre to be absorbed by libraries and archives. Underlying these simple claims is a morass of theoretical, social, policy and practical problems. This morass has proven to be fertile ground for research. The ideal lab is one that brings the many stakeholders together to explore, discuss and experiment with new models of data practices, management and technology. DANS may be that ideal lab.

@SciTechProf

Katy Börner:

“Let’s not waste expertise and time”

1. Many research teams that conduct science, technology and innovation studies use either Elsevier Scopus data or Thomson Reuters Web of Science data. The datasets are extremely comprehensive – covering much of humanities scholarly knowledge – but they are not perfectly clean. That is, all research teams apply their very own procedures to identify unique author names and institutions, to resolve errors in journal name and address data, etc. However, none of the teams is allowed to share cleaned data. That is, much of expertise and time is wasted cleaning data again and again. Even worse, teams cannot replicate each other’s results as they cannot access each other’s data.

2. More and more datasets are becoming available as part of the Linked Open Data effort (***lod-cloud.net***).

Institutions are making their own high quality data on faculty and their publications, funded research projects and courses taught available via research networking systems such as VIVO, Profiles NRN, SciVal Expert or CONVARIS. See the international network of emerging NRN systems at http://nrn.cns.iu.edu. Many of these systems are compliant with VIVO ontology, i.e. they can be cross-searched to find research collaborators, mentors and reviewers.

3. R&D today is interdisciplinary and global. Novel tools and team management approaches are needed to assemble winning teams dynamically, to make them work productively and to diffuse results widely. Systems and research results presented at the joint ***vivoconference.org*** are highly relevant for making this work.

@katycns

FUTURE

Andrew Treloar:

“I want to contribute to a better world”

1. The thing that I find most frustrating about my daily practice is all the things that prevent me from getting into, and staying in, a state of productive ‘flow’. I understand that constant interruptions from staff, phone calls and emails are part of the job, but on some days I cope better with this than on others.

2. The challenges that face the world are complex and scary, and what is needed to solve them is better knowledge and political will. I can’t help with the latter, but I can help with the former. Meeting the challenges of managing and working with data in support of our researchers is my way of contributing to a better world. And the fact that I find this intellectually stimulating is just a bonus!

3. Ah – an easy question. Cough. The ideal research environment for me would be one where the barriers to collaboration are reduced as far as possible, where the ability to build on previous work by anyone is maximised, and where researchers are able to spend their time focussed on making new discoveries rather than endlessly applying for research grants from a shrinking pool, or publishing just for the sake of building their CV. Describing this ideal is of course very different to making it happen. Because of the lack of political will alluded to earlier, I fear that we need the impact of humans on the planet to become undeniable before society will commit to resourcing researchers more appropriately. Let’s hope that by that time it isn’t too late.

@atreloar

Sharing data: good for science, good for you

13. Martijn Kleppe: “I estimate that 80% of all data are collecting dust in drawers or are dying on a hard disk. Along dies the information it contains, so a lot remains to be done.”

14. Sharing data: it’s good for science and for you. Please watch this movie at www.youtube.com/user/DANSDataArchiving

Herbert van de Sompel:

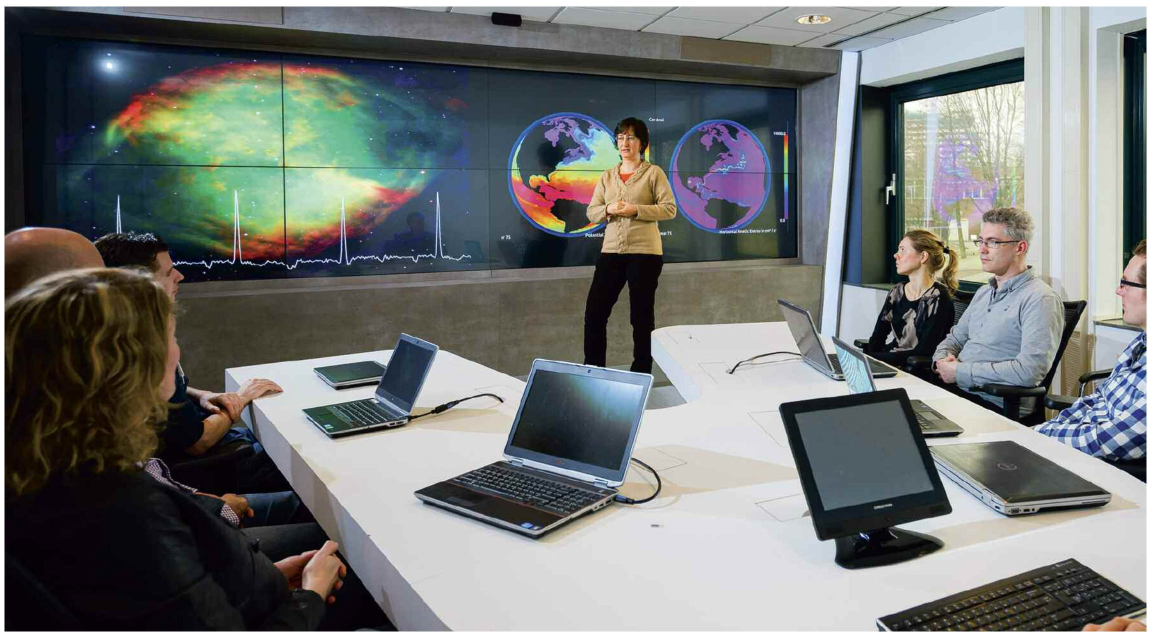
“I am excited when I see efforts that fully embrace the web”

1. I work at an institution that does open and classified research. While the policy constraints that rule classified work are very stringent for obvious reasons, unfortunately, also open research is subject to strict policies that sometimes make things that are taken for granted at other research institutions difficult / impossible, like certain cloud services.

2. I am excited when I see efforts related to research and research communication that fully embrace the web rather than devise a research-specific enclave that merely uses the web as a conduit. Recent examples include the W3C PROV work and the Open Annotation effort that recently became an official W3C work item. I think the ongoing Research Object work is moving in, what I would call, the right direction too.

3. I am extremely intrigued by the notion of using global sync&share technology such as DropBox as fundamental research infrastructure. Currently, those technologies are missing features that are crucial from a research communication perspective. But, researchers absolutely love DropBox for its sharing, mobility, and versioning features. Architecturally, sync&share technologies blur the boundary between local file systems and the web; every file potentially has a URI. The result is a technology that allows every asset to seamlessly move into the web flow when so desired. In that regard, in my opinion, DropBox scores better than some special-purpose infrastructures. Hence, to me a crucial question is: How could we overlay the features that are missing from a research communication perspective on sync&share technology?

@hvdsomp



The Collaboratorium *credit esciencecenter.nl/about-the-center/collaboratorium/*

NLeScience deals with scientific challenges

Access to data and the exchange of data across domains is a very important aspect of modern scientific practice. Patrick J.C. Aerts, Director of Strategic Alliances at the Netherlands eScience Center (NLeSC), explains about their focus on the actual deployment of data, software and tools to address complex scientific challenges of the present and the future.

Set up in 2011, NLeSC aims to reinforce and accelerate multi-disciplinary and data-intensive research by developing and applying eScience methods and approaches. eScience ('enhanced science') enables new scientific breakthroughs. NLeSC helps bridging the gap between advanced ICT and e-infrastructures and the scientific disciplines that seek to engage their scientific challenges. Scientists from all disciplines, from humanities to astronomy and from ecology to water management, can count on NLeSC. The particular focus at NLeSC is on excellence in eScience: it develops and reuses tools in expertise domains such as cross-type data integration, data-driven & multi-model simulations, visualization & analytics, high performance (exascale) computing and networking.

Network organization

As a network organization NLeSC collaborates with scientists at universities and maintains a network of eScience Integrators. They are professors in their application domain who act as ambassadors for eScience towards their domain while providing input to NLeSC from their disciplines. Also, NLeSC has established a platform of Dutch eScience/data research organizations (ePLAN) who share a common vision of the role of eScience and Big Data in science and research in both academic and commercial sectors. The importance of software sustainability can be compared to the exponential growth of the web. This growth was enabled by open web codes which could be easily copied and reused allowing every new generation to stand on the shoulders of its predecessors. Today there are libraries and containers with fully fledged solutions to almost everything. By sustaining scientific software a similar track can be followed by reusing software across domains rather than reinventing the wheel.

Community of practice

eScience has been called a community of practice growing towards a discipline per se. This reflects the fact that eScience in 2020 will be different from eScience today, but will develop organically with the ever increasing growth in ICT potential and e-infrastructure capabilities on the one hand and the ever more demanding challenges in public and private science and research on the other.

p.aerts@esciencecenter.nl
eScienceCenter.nl

The European framework for audit and certification

How can we bring trust to the system of digital preservation and sustainable access? Ingrid Dillo, Policy Director at DANS, has the answers.

Data sharing is far from common practice. Although there are a lot of advantages (transparency, ability to replicate and verify research, reuse of data, higher return on investment), researchers still argue that they cannot trust data produced elsewhere by somebody else. So, how can we build trust into the system of digital repositories?

Three levels of evaluation

In 2010, a European Framework for Audit and Certification of Digital Repositories was set up to provide an answer to the question of trust. The parties involved have come up with three levels of certification that offer increasing reliability:

1 Basic Certification is granted to repositories that obtain Data Seal of Approval (DSA) certification. DSA, initially developed by DANS, ensures that research data can be processed in a high-quality, reliable manner, provided the 16 guidelines for self-assessment are followed.

2 Extended Certification is granted to Basic Certification repositories that perform an additional structured, externally reviewed and publicly available self-audit based on ISO 16363 or DIN 31644. The DIN 31644 standard, an initiative by NESTOR, is a catalogue of 34 criteria that trusted digital repositories should meet. The ISO 16363 standard presents over 100 metrics for different aspects of a digital repository.

3 Formal Certification is granted to repositories which, in addition to Basic Certification, pass a full external audit and certification based on DIN 31644 or ISO 16363.

If you are a researcher you can deposit your data in a trusted digital repository, or you can ask the repository how they make sure data are treated well. If you are a funder of research proposals you can draw the attention of researchers to the importance of depositing their data in a trusted digital repository after completing their research. Or you can make it mandatory for researchers to deposit their data in a trusted digital repository. You can provide researchers with some funding to take care of their data management during research and to archive them afterwards. Alternatively, you can withhold a percentage of the initial funding if they don't. If you work at a research infrastructure, make sure that the resources you offer remain meaningful and usable over time by applying sustainable business models. If you work at a digital repository yourself, make sure it is trustworthy.

The time is now

We all endorse the critical importance of sharing and preserving reliable data produced during the scientific process. Trust is at the very heart of storing and sharing research data. The time is now; together we can make the digital world more reliable.

ingrid.dillo@dans.knaw.nl
trusteddigitalrepository.eu

DSA seal *credit DSA*

COLUMN

Let's take data to Broadway

Dear eScientist, I bet one of your first actions in the morning is switching on your computer(s). Science without computers is inconceivable today. Yet, we make life difficult for them.

From 'reading to form a hypothesis' we move towards 'reading to confirm a hypothesis' that was 'thrown in our faces' by data pattern recognition. For decades we have been integrating data, using Extract-Transform-Load (ETL) into yet another non-interoperable data warehouse for local use. ETL is no longer a viable option in eScience, but we continue to bury our results in narrative, clumsily linked 'supplementary data' and non-interoperable formats, using 'our own standards' which create nightmares for computers which we should be supporting as our main assistants by now.

Semantic technologies enable 'functional interlinking' of datasets for everyone to use and reuse, and they can be processed by machines 'as one', even though the underlying sources are disperse and in different formats. The 'narrative publishers' are also exploring business models that involve exposing the associations buried in their texts in machine-readable format.
Key associations reported in these data sources need to be carefully selected for their KD value by pattern recognition. What's more, those associations need to be made available in so-called machine-readable FAIR format: Findable, Accessible, Interoperable and Reusable. The underlying resources will increase in value, can be found and used more easily and will always remain crucial for 'confirmative reading'. As long as *de novo* data 'speak the same language' immediate patterns can be discerned that were escaping our attention without Functionally Interlinked Data.
All this means that our 'local shows' need to be 'taken to Broadway', which implies Professional Data Publishing with safeguards against all the mistakes made in narrative publishing. Listing the latter would make this column way too long, and we know them all too well.

Barend Mons
photo marietmons.nl