On 2 December 2009, the symposium ‘Driven by data’ took place in The Hague to celebrate the anniversaries of a number of archiving institutions that are now all part of DANS (Data Archiving and Networked Services). It was a lively meeting that paid attention to various surprising and creative possibilities associated with the reuse of electronic data. DANS decided to publish this jubilee volume to give a larger public a taste of this versatility and creativity. The contributions cover topics as diverse as dendrochronology, politics, shipping traffic history, bird migration, philosophy of science and statistics.

DANS (Data Archiving and Networked Services) is the national organization in the Netherlands for storing and providing permanent access to research from the humanities and the social sciences. DANS manages existing data archives and works on further developments of the data infrastructure in new fields as well.

www.dans.knaw.nl
DRIVEN BY DATA
Driven by data
Driven by data
exploring the research horizon

Editors
Martijn de Groot
Marion Wittenberg

Pallas Publications
Contents

Using a deluge to our advantage 7
Peter Doorn

Data archiving in the Netherlands: A look back and a look ahead 9
Peter Doorn

Create your own search engine 15
Ewoud Sanders

Migratory birds pictured at 1.5 MB per second 21
Martijn de Groot

New insights into Dutch shipping and trade in the Atlantic in early modern history 25
Henk den Heijer

Preserving tree ring data: A repository for the Low Countries 29
Esther Jansma

The Social Statistical Database of Statistics Netherlands: Invaluable source for socio-economic research 35
Johan van Rooijen

Calculating the women-friendliness of parliament 41
Maarten Marx

Scientific research is changing 47
Martijn de Groot

About the authors 50
Driven by data
Information has gone from scarce to superabundant. That is what it said under the heading ‘Data, data everywhere’ in *The Economist* of 25 February 2010. Commerce, government and science, they are all increasingly driven by data, as revealed in a special on *The data deluge*. In recent years, this deluge has been reason for several authors to predict the end of scientific theory. Editor-in-chief Chris Anderson of *Wired Magazine* for example: ‘Out with every theory of human behavior, from linguistics to sociology. [...] Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.’

Authors like Jim Gray and Tony Hey of Microsoft talk about a paradigm shift in their book *The Fourth Paradigm: Data-Intensive Scientific Discovery*. After a scientific approach with models and simulations, data exploration is starting to take center stage: ‘Data-intensive science consists of three basic activities: capture, curation, and analysis.’

One can ask whether the ongoing revolution will really change the core of scientific practice. Researchers, after all, always want to answer questions and will always have a use for data to that end. That there really is a data revolution, however, is becoming harder and harder to ignore. In recent years, we’ve seen the emergence of e-humanities and e-social science, which are indeed highly data-driven. Humanities are now also experiencing the development of uncharacteristic data collections, which are increasingly interconnected. Everybody agrees that the practice of science is changing markedly under the influence of digitization. If the 1980’s and 90’s were the years of *me and my database*, the emphasis has now shifted toward connecting comparable data bases in various areas since the turn of the century. This trend can be observed everywhere. More and more DANS projects are also moving in this direction. And it doesn’t stop there: digital data sets are linked to electronic publications, and to information on research projects and researchers.

Data Archiving and Networked Services (DANS), wants to play a central role in this force field in the next few years and saw a good opportunity for scanning the horizon in the symposium held on 2 December 2009, the celebration of several anniversaries. That was done by way of a number of recent research projects from different disciplines, from dendrochronology to political science, from geo-ecology to statistical economics and from history to knowledge representation. Not only was there a wide range of disciplines, the types of electronic data and the ways they were used also showed a great deal of diversity.

The width of that spectrum demonstrated the versatility of what is becoming possible now that we’ve fully moved into the data-intensive era. We want to showcase that broad horizon in this publication. Variation in content goes hand in hand with variation in form. Some contributions were written by the speakers supplementing their contribution to the symposium, others portray their own contribution pretty accurately, whereas yet others add a more journalistic view on the messages the speakers delivered at the symposium, derived from a special supplement to the quarterly magazine *e-data&research.*
Driven by data
Data archiving in the Netherlands:
A look back and a look ahead

Peter Doorn

It just so happened that in 2009, there were four anniversaries of archives and institutions involved in making research data available – all part of DANS these days. We celebrated that in December with the symposium that led to this booklet. By far the oldest of those archives is the Steinmetz archive, of which the establishment in 1964 can certainly be seen as the starting point of data archiving in the humanities and social sciences in the Netherlands.

1964: Steinmetz Foundation
On 27 November 1964, the Steinmetz Stichting voor het Opslaan en Toegankelijk maken van Bestaand Materiaal van Sociaal Onderzoek1 was established from within what was called the seventh faculty of the University of Amsterdam (UvA). Initiators were Professor H.M. Jolles of the Social Sciences Council and M. Brouwer, M.A, of the Seminarium voor Massapsychologie, Propaganda en Openbare Mening2 affiliated with the UvA. Data archives were started in many European countries in the course of the 1960s. Today, their successors cooperate within the Council of European Social Science Data Archives (CESSDA). In those days, examples were scarce. The Roper Center for Public Opinion Research was established in United States shortly after the Second World War and the Inter-University Consortium for Political and Social Research (ICPSR) in Ann Arbor, currently the largest data archive for the social sciences, in 1962. One of the few European predecessors of the Steinmetz was the Zentralarchiv für Empirische Sozialforschung, launched in Cologne in 1960, now a component of GESIS – Leibniz Institute for the Social Sciences.

At the start of the Steinmetz Foundation, computers did not play an important role yet. Mainly counting and sorting machines were used, to count punch cards. In those early years, the Foundation predominantly focused on acquiring codebooks, interview protocols, and punch card collections such as those of the 1960 Census of Statistics Netherlands (CBS). Those files are now neatly digitally archived at DANS. The oldest data sets of the Steinmetz collection dates back to the 1950s. Data set number P00001 of the Steinmetz Archive is a coded survey among listeners and watchers of the marathon radio and TV campaign Open het Dorp of 1962, presented live by Mies Bouwman a famous Dutch television presenter.
Throughout the years, the oldest data sets were all converted into modern media and stored in such a way that statistics packages like SPSS can easily process them. DANS keeps
Royal Netherlands Academy of Arts and Sciences (KNAW) in 1971. At the time, the collection contained some two hundred data sets. Five years later, in 1976, that number had grown to 650. The Steinmetz Archive first was part of the Sociaal Wetenschappelijk Informatie- en Documentatie Centrum (SWIDOC) of KNAW. Later, in 1997, it became a component of the Nederlands Instituut voor Wetenschappelijke Informatiediensten (NIWI).

1989: Netherlands Historic Data Archive (NHDA)
A first proposal for a historic data archive was formulated within the Working Group for History and Informatics (WGI, the later Association for History and Computing VGI) in August of 1987. At the end of 1988, the working group and the Steinmetz Archive submitted an application based on that proposal to the Netherlands’ Ministry of Education, Culture and Science. The latter awarded a subsidy from within the onderuitputting van de begroting (under-utilization of the budget), which would remain an important funding source for the Netherlands Historic Data Archive (NHDA), operating on a project basis. In 1989, the Netherlands Organization for Scientific Research (NWO) acknowledged the archive as a center of expertise, but without financing attached.

The NHDA, meanwhile converted into a foundation, became a formal component of the KNAW office in 1995, carrying the title of Institute. Two years later, it became part of the NIWI.

The archive’s first data sets originated from the History Department of the University of Leiden, where it was also housed. One file however, on Anglo-African trade, arrived on magnetic tape from the inheritance of British historian Marion Johnson,

records for each file on how often it is downloaded and by whom, and there is a demonstrable though modest need for such old files. The Steinmetz Foundation became the Steinmetz Archive and part of the
had ordered Statistics Netherlands to earn back some of its expenses by means of data supply. Several reports paid attention to this complicated situation and helped pave the road for the launch of the WSA. It did not manage its own archives, but entered into contracts with data producers and arranged the supply of files to researchers at reduced rates, sometimes still a few thousand guilders. The most important contract was that with Statistics Netherlands: For one million guilders per year, researchers could use part of its files at reduced rates.

The agency later became part of the Netherlands Organization for Scientific Research. For data supply, it had set up Dataneth, a catalog and ordering system on the internet.

1994: Scientific Statistics Agency (WSA)

The Wetenschappelijk Statistisch Agentschap (WSA) was established in 1994 as an intermediary for obtaining large data files from organizations such as Statistics Netherlands (CBS), the Netherlands Institute for Social Research (SCP) and Topographical Services/Land Registry. In the 1980s and at the start of the 1990s, obtaining files was no simple matter and could be very costly. The micro file of the 1981 labor force sample survey at CBS for example cost 100,000 Dutch guilders (on the order of 30,000 to 40,000 US dollars in those days).

The Ministry of Economic Affairs who had spent a large part of her working life collecting and entering sources related to that topic. Around 1990, querying and downloading data sets was already possible on the mainframe computer of the computing center in Leiden. Soon after, the archive could be queried via Gopher, a predecessor of the World Wide Web. The first NHDA web site was already launched in 1991. The archive took inventories of, and kept a register of, files on the computers of historians all over the country. For the sake of survival, it also started experimenting in the area of scanning and optical character recognition (OCR).

2004: Electronic Depot of the Netherlands’ Archeology (EDNA)

Archeology is an area of the humanities in which the computer already gained a foothold in the 1960s. As of 1980, when the first micro computer was used in an excavation project, almost all its projects were documented in computer files. The archeological geographic information system (GIS) ARCHIS was started in 1992, later taken over by Rijksdienst Oudheidkundig Bodemonderzoek (ROB, now Cultural Heritage Agency RCE). It contained the core data of excavations, but not all the available digital data files.

In 2003, at a joint conference of the VGI and the Netherlands Foundation for Archeology (SNA) the idea for an Electronic Depot for the Netherlands’...
Archeology (EDNA) came up. The following year, the NIWI and the University of Leiden carried out a pilot project, in which the ROB was also heavily involved. This quickly led to the EDNA I project, followed by EDNA II. DANS continued both in 2005, after its establishment. Archeological data sets can be very large and varied. In the short time of its existence, EDNA has grown very rapidly, in number of data sets and in data volume. It now has become the largest thematic archive at DANS.

2005: DANS
After the NIWI’s discontinuation in 2005, all archives and activities of the Steinmetz archive, NHDA, WSA and EDNA were transferred to DANS. Data supply to researchers, previously arranged by WSA, became a free service within two years. The data contracts were reviewed as well, and the changing views on public access to publicly funded data led to the almost complete ending of payments for data. This year, DANS has been in existence for five years. Many areas in the social sciences and humanities are presented in the current offer of data sets, but that offer does show a considerable spread. Archaeology is relatively over-presented, sociology and history are slightly above average. However, disciplines like economics, law, psychology and the arts can still do with some fortification. That is not only true for areas in the social sciences and humanities. At the establishment of DANS in 2005, the assumption was made that the physical and life sciences would be able to manage on their own when it comes to data. Indeed, the scale of data storage in those disciplines is much greater than in the social sciences.
and humanities. But a recent report of the National Coalition for Digital Preservation (NCDD) as well as international studies reveal fragmentation in terms of organization, tasks and responsibilities for data supply in the physical and life sciences. DANS therefore plans to extend its work area to more, and perhaps all of the, sciences. The DANS archive already contains dozens of data collections in the area of public health and epidemiology, and DANS is also involved in the establishment of the data center of the Netherlands’ three universities of technology.

As an organization that is not discipline-specific, DANS can contribute toward a clear division of roles in the Netherlands with regard to digital durability. With the National library of the Netherlands focusing primarily on publications, the National Archive on government information, and the National Institute for Sound and Vision on audiovisual information, DANS will serve the sciences and humanities in the data realm. In addition, it will also empower, both nationally and internationally, the discipline-neutral activities that DANS is already developing, such as the Data Seal of Approval, metadata standards and persistent identifiers. Finally, it will also clear the way toward better integration of various forms of research information found among and side by side with (enriched) publications and data sets, such as e-journals, software and project information.

Notes
1 Steinmetz foundation for storage of and access to existing Social Science data material
2 Institute of Mass Psychology, Propaganda and Public Opinion
3 Social Science Information and Documentation Centre
4 Netherlands Institute voor Scientific Information Services
5 Scientific Statistical Agency

Figure 5. Steinmetz Archive was named after Sebald Rudolf Steinmetz (1862-1940) professor in the Sociography at the University of Amsterdam. Drawing by Jo Spier.
Driven by data
Create your own search engine

Ewoud Sanders

Scientists, journalists and researchers do make abundant use of digital sources, but they allow themselves to be hampered by the internet’s relatively primitive possibilities for searching. They could be searching in a much more advanced and systematic way, without a lot of effort. As a result of that, the quality of their research might increase immensely.

There are limits to finding information on the internet. Do we have to take that for granted? The short and sweet answer is: No, we don’t. Select the sources you use frequently and place them on your computer or laptop in a structured manner, let indexing software tackle them and you will suddenly be able to search and sort in a much more sophisticated manner. Whoever happens to be thinking ‘I’ve already been doing that for ages’ can stop reading now. All sorts of studies reveal that the digital revolution is radically changing our reading behavior. We for example want to know the conclusion right away. Well, this was the conclusion. But it doesn’t reveal how to do this and what you can do with it. That requires a longer response and it will have to start with a short introduction.

Thirty book cases
I am a historian and journalist and I specialize in writing about the history of words and expressions. For that reason, I have built a large library in the course of twenty five years. I ended up with more than thirty book cases, filled with language journals, dictionaries and encyclopedias – thousands of titles. It is a privilege to have that many books around you, but in the end, what started to annoy me more and more was that I was not putting them to the best use. For example, I had almost two hundred books with proverbs, but was mainly using only four of them. Why those four? Because experience had taught me that there was a large chance I would find the best information in these books. This selective use started irking me so much that I decided to scan most of my library. At the moment, I have over five million pages on my computer, in approximately twenty thousand documents, predominantly in PDF/a format, totaling more than three billion words. If you put that much information on your computer, it forces you to investigate how to unlock this information optimally. One of the above sentences contains the phrase ‘I’ve already been doing that for ages.’ In my experience, remarkably few researchers, journalists and information specialists circumvent the limitations of the internet by using indexing software. That is why I will explain here, step-by-step and somewhat like a schoolmaster, how I did that. To be clear, I am not recommending scanning on a large scale. It would cost a great deal of time and money because you would need all sorts of things such as a cutter (that is, if you...
Figure 1 and 2. Searching through books systematically is extremely time consuming. Digitalising them and then running indexing software will make it possible for you to search in a more rapid, systematic and advanced way. And suddenly you'll be able to hold tens of thousands of books in a single hand.
Driven by data

17

...want to cut up the books, which is the only way to scan relatively cheaply), a good scanner, software for optical character recognition (OCR) and good backup systems. No, if you want to be able to search a lot of information in a sophisticated manner on your own computer, you can actually make good use of the sources that are available on the internet at no cost. Of course, I don’t know what type of research you are into, but let’s take two examples: linguistics, and historic research.

Books at no charge
The linguist will want to search the largest possible number of sources. The historian will want to collect materials that match his or her research best.

Nowadays, you can download books from all sorts of web sites, free of charge. Nearly all Ph.D. theses are placed online (obligatory). The Digitale Bibliotheek voor de Nederlandse Letteren (www.dbnl.org) has a couple of thousand books and journals available for downloading as PDF. Google is placing 330,000 Dutch books online at high speed, together with the Gent University Library, limited to the years 1600 to 1870 to avoid copyright problems.

If you want, you can put thousands of books on your computer within a few days, and at no cost. In my experience, you need to do two things in order to manage such quantities well. It is useful to rename PDF files after downloading so that the name of the document always starts with a year. It’s also practical to arrange the sources in folders by type. For example, all biographies in one folder, and all journals, all bibliographies, all books with proverbs and so on. The only thing then still needed is a small investment in indexing software that places a file’s words in an index, which you can then search quickly and in a sophisticated manner.

At the time, I tested seven indexing programs including X1, Archivarius and Copernic. I felt that Isys Personale Edition stood out head and shoulders: Australian software that is used worldwide by more than fourteen thousand businesses and institutions, including many large legal organizations. The program costs one hundred and twenty euros per year, including maintenance and offers a free two-week trial (see www.isys-search.com). Isys enables you to search over two hundred file formats, including mail. You can store the data on your own computer or on a server, after which an unlimited number of users can search the indexes simultaneously (for this you need the more expensive version, Isys Workgroup).

Sorting
What can you do with Isys you can’t currently do online? Assuming that you have placed the downloaded sources in thematic folders and that the names of the documents start with a year, for starters you can:

- Always arrange all results in chronological order (from older to most recent);
- Always arrange all results in reversed chronological order (from most recent to older);
- Always arrange all results by type, based on folder name.

You can also, unlike on the internet, search in advance manner, not just by using Booleans (and, or, not) but also with wildcards. Suppose a Member of Parliament is crazy enough to call...
the Prime Minister a *goofball*. You will quickly want to know a few things, and not just vaguely (the internet would suffice) but as detailed as possible: how old is this word, which writers use it and in which forms? In every search, you can replace a random letter by a question mark and a random number of characters by an asterisk. That would enable you to carry out the following searches:

- **goofball** (gives you all words starting with *goofball*, such as *goofballiness*);
- *goofball* (gives you all words ending with *goofball*, such as *megagoofball*);
- *goofball* (gives you all words with *goofball* at the center, such as *megagoofballiness*).

The wildcards can also be combined. A search on ho*it* will give you results like *hobbits* and *hobits* – old English variants of *howitz* and *howitzer*. Can you find such words by using for example Google? Yes, but often provided you know of their existence. The power of searching with this type of indexing software, however, lies in the fact that you find all sorts of things you didn't know existed. In other words, you now have a tool that

Figure 4. An indexed collection of books. The documents can be sorted in chronological order, reverse chronological order, thematically or according to relevance.
Driven by data

A search like ‘kn*ks*kr*d’ in an index of three billion words will give you all sorts of variants of the word *knäckebröd* within a few seconds. That also enables you to search for variants of a name, which on occasion may save you a lot of time. You can indicate that two words have to be on the same page, in the same paragraph or not more than ten (or twenty, or thirty) characters removed.

In my opinion, a great advantage of this way of searching is that you can limit your collection of materials. The information offered on the internet is gigantic and will only keep increasing. Moreover, the web is nothing if not dynamic. This holds disadvantages for researchers. Statements like ‘We did not find this’ or ‘Such and such was its frequency of occurrence’ can become untrue tomorrow. If you create and manage your own corpus, you will be able to say ‘This research is based on one/two/three thousand novels in our corpus’, or ‘on the following list of sources.’

Unexpected finds

For about ten years, I have been an intensive to highly intensive user of digital sources, in the first place of the internet, a source that I will never skip. But my own digital collection is a blessing for the essence of my investigations. I am no longer faced with useful web pages that have suddenly disappeared, because when I run into a (moderately sized) web site that contains information that is important for me, I will grab it and add it to one of the 145 rubric folders in my digital library, and my indexing program will update its indexes automatically. The greatest step forward, however, I see in the very advanced search features. I am not exaggerating when I say that I now find unexpected results every week, and occasionally even every day: new, fun, original, important information in sources that I previously likely would not have consulted at all.

So if you really want to jump ahead in finding, sorting and quickly consulting information, concisely put my advice is the following: Put sources you often use on your laptop or computer and do it in a structured manner, and let some indexing software have a go at it, as you will be able to do so much more than you used to think was possible.

Spelling variants

Of course, a lot more is possible. An x can be used to replace any word, enabling you to search on expressions like ‘as strong as x x’ – meaning that all sorts of variants of this expression will pop up. *Popping up* is the appropriate phrase, as the software is very almost forces you to find the least expected or original results.

Figure 5. Spelling variants of *knäckebröd*.
20 Driven by data
Migratory birds pictured at 1.5 MB per second

Martijn de Groot

‘No models without observation, no observations without models.’ Willem Bouten, full professor of Computational Geo-Ecology at the Institute for Biodiversity and Ecosystem Dynamics of the University of Amsterdam, has faith in models for predicting the movements of animals on and above the earth’s surface. And he believes in observations to build these models on. Many observations.

Everything converges in the Virtual Lab for eScience (VLe), where Bouten collects observations in nature from a variety of sources in the Netherlands and beyond. ‘We make use of the most recent knowledge and the latest tools from computer science, and of the increasing capability of rapidly transporting enormous amounts of data via glass fiber cables.’ He is willing to give an example. ‘In my virtual laboratory, I can use bird migration data that are spewed out by all sorts of radar installations in Europe at one and a half megabyte per second. A few years ago, that was still impossible’. And this development will continue rapidly, the ecologist ensures us. ‘Our computers double their speed every year and our communication becomes five times more rapidly every year. As a result, the envelope of your research possibilities keeps expanding’.

Bouten and his modeling efforts venture into an area that is located between two former opposites in ecology, he explains. ‘We have theoretical ecologists, who work with mathematical equations representing the behavior of animals but never see those animals up close. And we have field ecologists, who make observations in the field, with the primary aim of protecting animals. I make models to predict animal movements, and I don’t need to leave my laboratory for them. But I do use all the field observations I can get. No models without observations.’

Well-founded assessments
All these data come together at the virtual laboratory, which requires that all barriers between the various systems and protocols are removed. Initially, that was only done for scientific research, but soon it became evident that the system was also very suitable for statements regarding the distribution of certain species for the construction sector.

Models can be used to make predictions. That is indeed what Bouten and his group are doing: It provides them with feedback on the quality of their models as well as with money for improving them. Bouten: ‘We for example look into whether and where compensating measures must be taken. We do this for Bouwend Nederland, the association of the construction and infrastructure sector. In the past, construction projects simply ground to a complete halt when

Driven by data
someone reported the presence of a rare species on the site. To support such decisions, the Gegevensautoriteit Natuur (GAN, www.gegevensautoriteitnatuur.nl) was established. We created an information system for GAN that lets us indicate species distributions. That enables GAN to formulate a well-founded assessment of the impact of a construction project for rare plants or animals. We obtain the data from a wide variety of sources. There is a web site with observations, www.waarnemingen.nl, but there are also twenty thousand registered volunteers in the Netherlands who report observations to volunteer organizations such as Vlinderstichting, Floron (flora protection) and Sovon (bird censuses). But also Natuurnmonumenten, Staatsbosbeheer, municipalities and provinces collect data for management and protection purposes.

Such a system is possible at a larger scale, too, says Bouten, and that is what the European Lifewatch project (www.lifewatch.eu) is trying to accomplish.

** Millions of birds fly across the screen **

Insight in bird migration makes it possible for Bouten and his team to make predictions of the migration’s intensity up to two days in advance, which is a great help to the Dutch Air Force when planning its aviation exercises. Last-minute cancellation of such an exercise can easily result in tons of wasted expenses. It takes two hundred people to prepare the flights of two pilots. Letting them go ahead while flocks of migrating birds are passing may lead to serious accidents, however. To make their predictions, the ecologists use two very different types of measurements that, particularly when combined, produce the desired results. Bouten: ‘From my lab, I can consult the military radar systems of the Dutch and Belgian armed forces. We also have meteorological radars at our disposal, which were developed to measure wind speeds. You can see up to millions of birds fly across the screen, but you don’t know which species they are. In addition, we have attached GPS transmitters to about fifty birds, of which we can follow the individual movements. I can adjust the settings of such a GPS system on a bird from the Virtual Lab, or even from my home. So I know which bird it is and I can see its movement. What we then do
is connect the observations of those small GPS systems with those of the radars. After combining them with other data, for example meteorological data, we can make bird migration predictions that are used daily by the air force and by bird watchers. To this end, so-called concept-driven models are being deployed, that can predict the routes and foraging locations of birds on the basis of theories and the weather. Bouten: ‘Say, a bird takes off from a particular site in Scandinavia with a certain amount of body fat. We can use aerodynamics to calculate how much fat it burns and where approximately it will have to land to forage’.

Reducing dependency on personal judgement

Bird alerts were issued for the Air Force in the past as well, but those were based on radar observations and common sense, says the scientist. ‘Someone with a thorough knowledge of bird migration was often able to interpret the radar data reliably, but such alerts strongly depend on the person because transferring that type of knowledge is hard.’ His current, data-driven models reduce that dependence significantly, but that does not mean that the input of experts is now ignored. ‘The expert now makes his or her own predictions and takes the modeling results and measurements to see if there is room for improvement’.

This contribution is an edited version of a previous article by Martijn de Groot in the quarterly e-data&research.
Dutch Atlantic Connections

New insights into Dutch shipping and trade in the Atlantic in early modern history

Henk den Heijer

For a long time, Dutch historians dealing with overseas history focused their eyes eastward. The Dutch East India Company (VOC) had a large and certainly profitable trade network. That network was the basis for the later colony called the Dutch East Indies. The economic importance of the West, however, was not rated very high. The Dutch West India Company (WIC) was a financial failure, its trade supposedly did not amount to much, and from the 19th century onward, more money went into the colonies in the West than came out of them. But, was that really the case?

The views of the economic development of the Dutch Republic in the seventeenth and eighteenth century are fairly clear, but not always correct. One such view is that the Dutch mainly traded slaves in the Atlantic region and kept plantation colonies with slaves. The Dutch shipping and trade are said to have collapsed in the course of the eighteenth century, a development that proceeded more quickly in the Atlantic region than in the East. In the 1960s and 1970s, historians started placing question marks behind those opinions. Extensive quantitative research in historic sources was undertaken to verify whether the views on the past were actually correct. The University of Leiden, for example, carried out a large study into trade and shipping between Europe and Asia. The results were published in Dutch-Asiatic Shipping in the 17th and 18th centuries. This three-part publication of sources contained data for 4800 journeys from the Netherlands to Asia and 3500 journeys from Asia to the Netherlands. The resulting data file made it possible to formulate new questions and find answers regarding matters such as trade volume and fluctuations in Dutch shipping traffic to Asia. To the surprise of many, the
decline of the VOC turned out to have started much later than had long been assumed.

Unlocking all the data
At the end of the 1960s, studies of the Dutch transatlantic slave trade started as well. American historian Philip Curtin's publication *The Atlantic Slave Trade; A Census* served as catalyst for the Dutch investigations. Curtin estimated that approximately eleven million Africans were made slaves and taken to the New World from the sixteenth until the twentieth century. Prior to the publication of his study, the estimates ran from about 3.5 to 100 million. Was Curtin right, however? Historians in various countries started an extensive quantitative study in order to prove or negate his findings. Historian Johannes Postma, who is based in the States, took on the task of looking into the Dutch contribution to the transatlantic slave trade. Very little was known about it at the end of the 1960s. His research and that of others revealed that Dutch merchants removed approximately 550,000 slaves from Africa, on a total of about 12.5 million, as was shown later. That means that the Dutch share took up almost five per cent of the total. In 1991, an international group of researchers decided to gather all available information about the transatlantic slave trade in a large data file, and track down any lacking data by additional research. Their efforts resulted in the launch of the web site www.slavevoyages.com in 2006. It unlocks the data of some 35,000 transatlantic slave journeys, which represent 80 to 85 per cent of all slave journeys, so the editors estimate. Since the launch of the web site, many publications have appeared that offer new insights into the various aspects of the transatlantic slave trade. Now for the first time, we can determine with some certainty from which West-African areas the slaves originated, which percentage passed away during the voyage, and where exactly in the New World the survivors ended up.

Tedious puzzle-solving
At the beginning of the 1990s, research into the Dutch activities in the Atlantic region broadened. The historians’ attention diverged to include goods being traded with West Africa, parts of North and South America, and the Caribbean. Investigating this was not easy. The VOC, which held
the monopoly in trade and shipping traffic to Asia, had left behind an excellent and well arranged archive, with nearly all the data a researcher could possibly want. No wonder that many historians and institutes focused on it in the past. The WIC, by contrast, was an economic failure, and to make matters worse, most of its archive was demolished. But the trade with the West was of a very different nature than the trade with the East. The trade in the Atlantic region was mainly the realm of small private entrepreneurs. Hence, the researchers had to accomplish magic and acquire data from a large number of different sources. Next, they had to connect those data to gain insight into the Atlantic trade network of the Dutch Republic. This tedious puzzle-solving resulted in a number of interesting data files over the past twenty years. We now know much more about the extent of the trade and shipping traffic between the Netherlands and West Africa, Surinam, Curacao and parts of North America. The available data reveal that the economic activities of the Dutch in the Atlantic region were much more sizeable than had been assumed until recently. The Dutch colonies in the West were small and few, but the merchants operating there succeeded in rerouting part of the production of the surrounding French and British colonies. Part of the sugar, cocoa, coffee and tobacco produced in them were shipped to Dutch ports via Curacao and St. Eustatius, which were both Dutch. Furthermore, the investigation revealed that the Dutch trade with the West did not decline in the eighteenth century but by contrast, grew strongly.

Relational database available

It is not yet clear how the Dutch managed to secure a disproportionately large part of the transatlantic trade. Also not quite known is how large that Dutch part was. In 2008, a group of researchers of the University of Leiden and the Vrije Universiteit in Amsterdam started a large study aimed at clarifying those issues. The NWO-financed research project is titled ‘Dutch Atlantic Connections’ and will run until 2013. Existing data files of earlier research projects have been joined in a so-called relational database within the framework of this project. This also included the nearly 1600 transatlantic slave journeys that the Dutch carried out. Collecting them and making them uniform, made it possible to unlock these ‘old’ files digitally, and create a database of thousands of Dutch shipping movements in the Atlantic region. This relational database has been deposited at DANS and is available to other researchers now. More new data that the ‘Dutch Atlantic Connections’ project dug up from archives will soon be added. The database makes it possible to address new questions. That way, hard data can put a stop to existing assumptions and the view on the economic importance of the Dutch shipping and trade in the Atlantic in early modern times can be adjusted. It will engender new insights and a stream of new publications.
Preserving tree-ring data: A repository for the Low Countries

Esther Jansma

Dendrochronological research traditionally is directed at establishing the wood’s absolute age and origin. Encouraged by the current size of dendrochronological data collections, the focus is nowadays shifting towards larger-scale studies, in order to answer questions about the former (cultural) landscape, economy, wood-processing industry and wood technology. Throughout Europe, over a million ring patterns from the Holocene have been dated. These data need to be combined to facilitate the new research directions. Also, international standards are needed for data storage and metadata registration and to maximize exchangeability of data collections. The research program ‘Digital Collaboratory for Cultural Dendrochronology’ is aiming to achieve all this.

Wood from the Holocene is preserved in the soil, historical buildings, and objects such as paintings and furniture (Fig. 1). Its patterns of annually varying ring widths can be read as chronological records of biological, geological and climate processes throughout the last ten millennia (Fig. 2). They are also an important source of information about our former dealings with this material (Figs. 3 and 4). Dendrochronological research of wood from the cultural heritage traditionally is directed at establishing the wood’s absolute age and origin. This type of study focuses on single locations, objects and structures. But nowadays, the focus is shifting towards larger-scale studies, in order to answer questions about the former (cultural) landscape, economy, wood-processing industry and wood technology. This shift is made possible by the current size of dendrochronological data collections. Throughout Europe, over a million ring patterns from the Holocene have been dated, meaning that an absolute calendar year has been assigned to each of the growth rings in the patterns. These data need to be combined to facilitate the new research directions. Bringing dendrochronological data together is not as easy as it may sound. At present, dendrochronologists use a variety of locally developed data formats. The most common ones were developed in the 1980s and are not suited for storing metadata relevant to cultural heritage. Most dendrochronological software can only handle old legacy formats. As a consequence, many laboratories in Europe keep paper administrations of their descriptive and interpretative metadata.
Figure 1. For over 26 centuries this piece of oak was preserved in the soil of Uitgeest (NL). It belongs to an Iron-Age canoe that was excavated in 2003. Its tree rings have been dated to the year using dendrochronology.

Dendrochronological measurement series of ring widths, as far as they are digital, are mostly stored in separate binary or ASCII files. Most of the data collected in the 1940s through the 1980s, if still available, exist only on paper (Fig. 5).

The Digital Collaboratory for Cultural Dendrochronology (DCCD)

The current challenge of cultural dendrochronology in Europe is to develop and maintain international standards for data storage and metadata registration and to maximize exchangeability of dendrochronological data collections. The Netherlands has taken the lead in this development, through the research program Digital Collaboratory for Cultural Dendrochronology (DCCD), managed by the Cultural Heritage Agency RCE. The program will be concluded in September 2010 with the launch of an online repository, developed by DANS and containing all dendrochronological measurement series and descriptive and interpretive metadata now managed in laboratories in

Figure 2. Example of a supra-regional environmental signal: the Minoan eruption of Thera (1628 BC) marks the start of a severe growth depression (red) in Dutch oaks (palaeo-ecological sites).
Driven by data

The Roman river barge ‘De Meern 1’ was excavated in 2003 in Leidsche Rijn (NL).

In a manner that it can be linked to other relevant digital archives in the Netherlands and abroad. Scientists in Austria, Denmark, Ireland, Latvia, Lithuania, Poland, Slovenia, Spain and the UK recently have selected the DCCD as their future vehicle for collaborative research.

The DCCD infrastructure consists of six building blocks: (a) an online ‘trusted’ digital repository of data and metadata, (b) a user interface for uploading, downloading, searching and controlling data access, (c) the XML data standard TRiDaS, which allows for storage of data and metadata pertaining to the historical and palaeoenvironmental sciences, (d) controlled multi-language vocabularies to ensure compatibility of metadata, (e) a stand-alone database for producing and ingesting TRiDaS XML and local data analyses, and (f) a digital platform for collaborative research, the exchange of DCCD-related products, and education.

Digital repository and user interface

The DCCD repository is developed by DANS and stores dendrochronological and related data: measurement series, descriptive and interpretative metadata, and related files such as articles, reports, photos and videos. When finished, the repository will receive the Data Seal of Approval developed by DANS, ensuring the quality and preservation of its content. The web application, also developed by DANS, enables non-members to produce general overviews of the repository’s content. Members can query the repository in more detail, and import or export selected data using TRiDaS XML. Although measurement series can be uploaded and downloaded in various old legacy formats, the repository only accepts metadata conform TRiDaS. The application validates the uploaded data against the community standards maintained in the DCCD’s digital collaboration platform. A key concept is that owners of uploaded datasets determine which DCCD member is allowed access, and to what information level. The levels of data access are equal to the information levels of TRiDaS.

The Tree-Ring Data Standard (TRiDaS)

During the startup phase of the DCCD program, the RCE developed the first outlines of a digital data standard for dendrochronology in collaboration with ICT specialists, dendrochronologists, cultural heritage specialists and geoscientists. The results were used for the development of an international standard, in close collaboration with Cornell University...
An international discussion forum was set up to ensure support and feedback from the global dendrochronological community. In addition, a web site was launched containing the technical description of the data model. This community-wide effort has resulted in the Tree Ring Data Standard TRiDaS (Jansma et al. 2010). The standard uses the eXtensible Markup Language (XML), which offers several advantages. XML is readable by computer software, but it is also clear to humans. In addition, it is open, meaning that anyone can use or read XML files. XML is multilingual and functions across platforms. Furthermore, XML schemas are easily validated. And finally, XML schemas are backwards compatible. The new data model is implemented by the DCCD community and Cornell University. USA-based organizations such as the International Tree Ring Data Bank are now looking for funding to do the same.

**Controlled vocabularies**

The DCCD-controlled vocabularies are designed to encourage standardization of terms used for objects and methods of dendrochronological research. They contain terminology for, among others, artifact types, timber types and wood species. By introducing multilingual lists of terms in the collaboratory, the community insures standardized terminology and the DCCD can be programmed in such a way that it becomes possible to search across languages. Currently, the controlled vocabulary of the DCCD contains English, Dutch, German and French terminology.

**Stand-alone TRiDaS database**

An Access database is being developed as a local tool for interaction with the DCCD repository and for data analysis. This stand-alone database will be publicly available. It is designed to administrate research projects locally according to the TRiDaS data model. These projects can then be exported to TRiDaS XML, making them suitable for uploading to the DCCD repository and other TRiDaS-based environments. The database can import TRiDaS XML that is downloaded from the DCCD repository and other TRiDaS-XML-producing environments, and can be

---

*Figure 4. Dendrochronology has shown that the ‘De Meern 1’ was constructed ca. AD 148 from oaks that grew in the Central Netherlands, north of the Roman frontier. Later additions to the ship, such as the cabin’s floor boards and bed planks, mostly were made of trees that grew further to the South and as far as southern Germany (Jansma 2007).*
Digital collaboration platform
A repository and online user interface, digital data model, controlled vocabularies, and a tool for local analysis, uploading and downloading are not enough to insure that this reorganization of dendrochronology will be effective. Durability can only be insured if participants continue to exchange ideas and products, and through research collaborations based upon the new infrastructure. This is why Utrecht University has set up a digital research and education platform for dendrochronology.8 This Virtual Knowledge Centre includes wiki, blog and forum functionality and is used for (inter)national research collaboration, education, information, and discussion. It offers protected environments for research and other activities and is linked with the DCCD web application. The DCCD’s stand-alone database, controlled vocabularies and newly developed software (including source code) are made available through this platform.

Program outlook
The DCCD program was developed to improve dendrochronological data handling, communication and research collaboration. One of the goals is to increase the research potential of cultural dendrochronology and to stimulate interaction with other scientific disciplines. The DCCD is already functioning as outreach to scientists in many other European countries. Most, if not all, European countries have no formal guidelines for the treatment of dendrochronological data from the cultural and natural heritage. As a result, large amounts of data have been lost and are still being lost, among others because researchers retire and because of deterioration of digital data files. Research using the DCCD will illustrate the benefits of proper data handling in dendrochronology. The Data Seal of Approval and key characteristics of the data model may well serve as a model in future quality agreements regarding dendrochronological data and metadata management.

Reference

Notes
1 Funded by Dutch research organization NWO
2 After this article was written, the end date of the program was shifted to June 2011. It was decided that the DCCD will be released twice: the core application will be released in September 2010 and a second version expanded with, among others, GIS functionality will be released in June 2011.
3 http://dendro.cornell.edu/.
4 The Dendro Data Standard Forum is managed by Dr. Peter Brewer (Cornell University) and open to all new participants. It can be joined by contacting p.brewer@cornell.edu.
5 www.tridas.org.
6 http://www.ncdc.noaa.gov/paleo/treering.html
7 The database, commissioned by RCE, is developed by Kit Sturgeon and Steve Mohlke (the Epison Group, Ithaca, New York) under supervision of Jansma and in close collaboration with Peter Brewer of Cornell University.
8 www.uu.nl/vkc/dendrochronology
Driven by data
The Social Statistical Database of Statistics Netherlands: Invaluable source for socio-economic research

Johan van Rooijen

The Social Statistical Database (SSD) of Statistics Netherlands (CBS) comprises data on persons, households, jobs, profits, social security benefits, addresses and dwellings. It is based primarily on registers and was developed to reduce the burden imposed by surveys on citizens and companies as well as to meet increasingly intricate requirements raised by policy makers and scientists. By linking the data, a wide range of socio-economic questions can be addressed statistically. This has already been borne out by countless studies, many of which by institutions other than CBS. This contribution shows results from some of these studies to illustrate the research possibilities offered by SSD data, thereby confining itself to descriptive statistics for the sake of readability.

Contents and background
The Social Statistical Database (SSD) of Statistics Netherlands (CBS) is a register-based system that comprises microdata on persons, households, jobs, profits, social security benefits, addresses and dwellings. The SSD is currently based on over forty registers and surveys and continues to expand (See table 1 on next page). The database was developed to meet two primary objectives. First, the burden traditionally imposed by surveys on citizens and companies had to be reduced by switching to administrative registers whenever possible (CBS Act, Section 5). Secondly, socio-economic statistics had to meet increasingly intricate requirements raised by policy makers and scientists as a corollary of an increasingly complex society (Al & Bakker, 2000). Examples of such requirements are:

- Information on specific groups in society, e.g. ethnic minorities;
- Longitudinal analyses, e.g. following labor market participation and income of discharged employees through time;
- Information on relationships between the various domains of the life course of individuals, e.g. the effect of having children on labor
Driven by data

market participation of women, controlling for educational attainment earlier in life;
- Information on fluxes instead of balances, e.g. information on whether an increase in labor market participation is the result of an increased influx, a decreased efflux or the combined effect of both.

**Statistical processes underlying the Social Statistical Database**

An extensive statistical process underlies the SSD. It concerns the following:
- Inspection and correction;
- Data security measures;
- Standardization;
- Micro-integration.

*Inspection and correction* are carried out at the level of the individual input register. This process entails checks of the record layout, population coverage, plausibility of information content at the variable level as well as corrections such as imputations of missing values. *Data security* measures include the substitution of identifying register data, such as social security number, name, date of birth by meaningless numbers. These numbers can also be used to link data from different sources in subsequent processes. The subsequent *standardization* is a comprehensive process. First, names of files and variables are transformed to comply with naming conventions. Secondly, files are converted to a standard format. Thirdly, meta-information is added in a standard format; descriptions regarding content of registers, codes of categorical variables, data models. In the final sub-process of *micro-integration*, different registers are confronted and conflicting information is corrected. For instance, if a job is extant according to one register whereas the corresponding employee has deceased according to another, the job is ended in the micro-integration process. In addition, longitudinal consistence is realized. Thus, micro-integration considerably increases quality and consistency of the data.

---

**Table 1. Subjects within the SSD (non-exhaustive) and their source registers**

<table>
<thead>
<tr>
<th><strong>Subject</strong></th>
<th><strong>Source register</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Jobs</td>
<td>Administration of Employee Insurance Schemes (AEIS)</td>
</tr>
<tr>
<td>Persons (demography)</td>
<td>Population Register (PR), administration Immigration and Naturalisation Service (INS)</td>
</tr>
<tr>
<td>Social security benefits</td>
<td>Social Assistance Benefits Administration (SABA, municipal authorities), Administration of Employee Insurance Schemes (AEIS), payroll data, administration Social Security Bank (SSB)</td>
</tr>
<tr>
<td>Persons (profits of self-employed persons)</td>
<td>Register Income Tax Declarations (tax authorities)</td>
</tr>
<tr>
<td>Persons (crime suspects)</td>
<td>Police Register</td>
</tr>
<tr>
<td>Persons (education)</td>
<td>Central Register for Enrolment in Higher Education (CREHE), Exam Results Register (ERR), Labour Force Survey (LFS), Education Number Registers (ENR), public employment service (PES), Student Finance Register (SFR)</td>
</tr>
<tr>
<td>Persons (lay-offs)</td>
<td>public employment service (CWI), courts of law</td>
</tr>
<tr>
<td>Dwellings</td>
<td>housing register (HR), valuation of real estate registration system (VRERS)</td>
</tr>
<tr>
<td>Addresses</td>
<td>geographic base file (GBF), register of addresses of institutional households (AIH)</td>
</tr>
</tbody>
</table>
Data security and privacy

Within the SSD framework, data security and privacy are evidently prominent issues. A strict data security policy is applied in order to comply with prevailing legislation as well as protect the reputation of Statistics Netherlands. This policy includes the following:

- Replacing identifying register data by meaningless numbers;
- Restricting access to the SSD data to personnel for whom access is imperative for the execution of their task;
- Restricting access to SSD data to only those data that are of immediate relevance for the execution of a task;
- Technically uncoupling access to SSD data and facilities for e-mailing attachments;
- Requiring personnel using SSD data to sign a confidentiality statement;
- Securing publications in accordance to strict guidelines.

Use of data

SSD data are disclosed in the form of so-called components, which are filed in a central library, accompanied by meta-information. Components are invariably provided with a key. A tool has been developed to link components on the basis of the key, resulting in data sets which form the basis of analyses and subsequent publications (Fig. 1).

By linking the data, a wide range of socio-economic questions can be addressed statistically. This has already been borne out by countless studies, many of which by external institutions. Examples are studies on labor market participation of discharged employees, socio-economic integration of immigrants and the relation between number of children and educational level of women. This paper presents some results from these studies to illustrate the research possibilities offered by SSD data, thereby confining ourselves to descriptive statistics for the sake of readability.

Illustrations of research based on SSD data

In the first study (Van Gaalen & Van Rooijen, 2009), discharged employees were followed for two years following lay-off in order to study the course of re-absorption by the labor market. Figure 2 depicts the participation of discharged employees. Most notable is the fast initial recovery followed by a surprisingly slow further recovery. Apparently, part of the employees had anticipated the dismissal and either already acquired new jobs or found new jobs soon after dismissal. Another part experienced

![Figure 1. Linking the components to form data sets for analysis](image1)

![Figure 2. Percentage of discharged employees having a job in the 24 months following dismissal in 2003](image2)
difficulties in re-entering the labor market. More detailed investigations demonstrated that the latter group is predominantly composed of typical weak employee categories such as older employees, women with children, and ethnic minorities (Scheele et al., 2008; Corpeleijn, 2009).

The study of return migration of immigrants (Van Gaalen & Bijwaard, 2008) concerns the relation between labor market participation and return migration of labor migrants. Fig. 3 shows the six-year course of return migration of a cohort of labor migrants that settled in the Netherlands in 1999. Striking is the fact that more than half of the labor migrants has left the country within six years after entering. An intriguing additional result is that labor migrants who were successful on the Dutch labor market had, counter-intuitively, a stronger inclination to leave the country again than those who were less successful. Figure 4 shows the relation between educational level and number of children for women of several birth cohorts (Van Agtmaal-Wobma & Van Huis, 2008). Clearly, a higher educational level is associated with a lower number of children irrespective of cohort. This difference was found to be attributable principally to a positive relation between educational level and the chance of staying childless.

**Future developments**

The Social Statistical Database will continue to expand. Periodically, new data will be added to existing data sets. This will enable longitudinal analyses over longer time spans as well as more detailed analyses of intergenerational transmission. Furthermore, new registers will be added so as to broaden the scope of SSD-based socio-economic research. For instance, an administrative register on primary education (Primary Education Number Register) is expected to be added this year. Finally, the combination of periodical additions and additions of new registers will in some cases lead to an increasing coverage of the Dutch population. For instance, every year, data on the educational level of approximately 300,000 people are added to those of the seven million already included in the SSD, which increases reliability and accuracy of statistical results.
References
Driven by data
Calculating the women-friendliness of parliament

Maarten Marx

More and more data are becoming available to social scientists in the form of raw text as material for text analysis. These data are usually easy to find on the internet. However, it is often still a big step to turn all sorts of formats into a nice input file for the popular analytical package SPSS. But suitable tools are available.

A recent article in *Science*, entitled ‘Computational Social Science’, contains a plea for curricula in which students in the social sciences and humanities learn to use the tools for handling vast amounts of text. We can demonstrate what these tools do by using an example, and showing that using these tools is relatively straightforward. Computers have become powerful and simple enough to enable even information-challenged persons to carry out quantitative research based on huge amounts of text. The required knowledge can be taught to any student in a moderate-size course on text analysis for the social sciences.

A simple research question...

We use an example to illustrate this. Let’s address the following research question. In the years 2006-2010 the Dutch parliament’s so-called *Tweede Kamer* contained a record-high percentage of women, namely over forty per cent. Are these women merely there for the record or do they get speaking time, proportionally? This question yields various subsequent questions. For example, does this vary by subject? And are there any differences in women(un)friendliness among the parties? How were things in the past?

All the data needed to answer this research question are present. *The Handelingen der Staten Generaal* (the Acts) are public; they are available on the internet for the years since 1917, as PDF files. They contain the exact transcriptions of what everyone has said in the *Tweede Kamer*. In addition, the website parlement.com contains extensive biographies of anyone who has ever been part of this Chamber.

We could therefore make a tally for each member of the *Tweede Kamer* how many times he or she contributed to a debate, how often a member interrupted, how long a member spoke, and how much speaking time that member used up. Although times are not listed in the Acts, we can approximate them by counting the number of spoken words.

...that cannot be easily answered

So far, nothing about this is difficult or special. And yet, it is not an easy thing to do because the data are not available in the right format. There are three problems. In the first place, it concerns a large amount of data:
Driven by data

3560 biographies and more than one hundred million words, spoken in the Tweede Kamer since 1995. Secondly, coupling the two data sets is difficult as the members of parliament are not consistently referred to by the same name. This problem is worse for data from before 1995, which were scanned and still contain errors as a result of optical character recognition (OCR). And finally, the Acts consist of text files in PDF format, with scant metadata. Of every word in the Acts, we do know on which day it was spoken and on which page it is, but not by whom it was said and in what capacity, as Member of Parliament or member of the government, as argument, interruption or response to an interruption.

The technique: Conversion to xml

Once the Acts are rendered machine-readable, techniques for text analysis can solve the problem of the recognizability and professional capacity of the speakers. With the aid of named entity recognition and reconciliation, we can recognize speakers and normalize their names and thus eliminate the thresholds for combining the two data sets. That is also the moment from which we can deploy computers to tackle the large data quantity. We no longer need to work with expensive coders and samples, and can carry out the analysis on the entire data set.

Next, we can take a look at how this conversion works in practice. These days, the Acts are already placed online as an html page the following day (www.tweedekamer.nl). Our example is based on that format. It works similarly for PDF input, but technically, it is a bit more complicated.

Here is an example from the debate of 13 January 2010 concerning the Davids Report (about the decision-making process surrounding the Dutch decision to back the war in Iraq):

Minister <strong>Balkenende</strong>:
<p>Well, what am I doing here?</p>
<p>The <strong>chair</strong>:</p>
<p>Yes, the Prime Minister is still standing too.</p>
Minister <strong>Balkenende</strong>:
<p>And it is already a quarter to three.</p>
Mrs. <strong>Kant</strong> (SP):
<p>A few hours ago, I would not have expected you to still be standing there.</p>

We briefly explain four differences:
1. The structure now clarifies who says what. All sections with text spoken by a person are nested in a <speaker> element.
2. Only what was actually said is in <p> elements now. This allows us to distinguish between the “Kant” who is speaking and the “Kant” who is mentioned.
3. The names of the speakers are rec-
ognized and linked to their unique number at the Parliamentary Documentation Center. For example, the biography of Balkenende (number 02207) can be seen at http://www.parlement.com/9291000/biof/02207. Also pay attention to the “Chair”. In the first version, it is not clear who this is at that time. Although changes of Chair are indicated in the Acts, this may have occurred pages ago.

4. The professional capacity in which speakers contribute is listed in the “type” attribute. The party is only listed for Members of Parliament.

The above consists of digital text and therefore does not contain any OCR errors. However, connecting a name with a unique person is not always simple. At the moment, there are two Members of Parliament with the name J. van Dijk, for example. Also, persons are not consistently indicated with the same name. See figure 1 for an example of a list we extracted from the Acts since 1995. Is that really only one person?

Figure 1. The Acts contain multiple references to the same person.

Figure 2 contains a number of party names from the scanned Acts between 1985 and 1995. Does this always refer to the Partij van de Arbeid (PvdA)?

Figure 2. OCR inaccuracies led to errors in the abbreviation PvdA.
One application: Summaries of speakers
In this new format, it is a piece of cake to use the Acts and produce all sections with text spoken by Balkenende in a debate. The following XPath query does that for the debate of 13 January 2010 which included the report of the Davids Commission:

```
doc('plenary_meeting_13_january_2010.xml')//
    speaker[@name="Balkenende"]//
    p
```

XPath is a very intuitive language, in which queries are set up as paths through the xml hierarchy. The above XPath query states: ‘give the sections (p’s) within the speaker element with the name Balkenende and extract them from said file’. We now have everything Balkenende said that day, on all topics covered that day. If we only want the text of what Balkenende said concerning the report of the Davids Commission in that meeting, the query becomes:

```
doc('plenary_meeting_13_january_2010.xml')//
    subject[@subject='Report Commission Davids']//
    speaker[@name="Balkenende"]//
    p
```

It is not very interesting to read everything Balkenende said without having its context. But computers can also summarize texts, for example, as a word cloud as given in figure 3.

Another application: a political iPhone app
Students at the University of Amsterdam created a fun application: an iPhone app in which curious sentences spoken in Parliament fly across the iPhone screen. Having the debates in the new handy format, the developers of the app were able to solve the really difficult challenge: How do you pull interesting and curious sentences out of texts in an automated manner? Figure 4 shows a screenshot of their app. A photo of the speaker is displayed together with the quote. When you click on the quote, you are taken to the moment in the debate when this sentence was spoken.

Merely filling the quota?
Currently, the University of Amsterdam is cooperating with the National library of the Netherlands on making the Acts available in xml format. This
Figure 4. iPhone app showing quotes from parliament made by Felix Cornelissen and Thomas Kuipers. www.poliquote.nl

would facilitate finding the answer to the research question posed at the beginning.
Meanwhile, we have carried out some quick tallies to measure women-friendliness, as a test. If we exclude the Chair, women used 33% of the speaking time in the Tweede Kamer in the period 3 February 2009 up to and including 8 October 2009. That is almost 20% less than you would expect on the basis of the number of female members. Women occupy the speaker’s position only 30% of the time. Of course, this can mean all sorts of things. Women may simply be less long-winded than men and more to the point when responding to interruptions.

Conclusion
Enormous investments have been made in language technology tools for the Dutch language. However, for ‘simple’ social scientists and the like, these tools still are often difficult to apply, certainly when they have to be used in combination with other tools. This is a shame as a plethora of raw data is freely available. The aforementioned Science article describes the real danger that the industry (Google, Amazon, etc.) will snatch the discipline of Computational Social Science out of the hands of science forever. Let’s make sure that this won’t happen.

Soon, everyone will be able to carry out this investigation on their own as we are placing all Acts in the EASY archive in DANS in a uniform xml format. And we are adding to it every day.

Reference
Driven by data

Kennedy heeft dat geweigerd, iets dat streft, in Groot-Brittanie altijd een vermoeden van den weigerenden beklaagde wekt. Men heeft het alleen als een vermoeden ingesteld van Kennedy's eigen mededaderschap opgetreden en vermoeden ook ten voordele van Browne's positie door de uiteenrafeling van Kennedys misschien versterkt geworden ware. Dit is van belang, want hoewel de rechter den gezwoen hart bond, dat Kennedy's verklaring niet als Browne's zeggend beschouwd mocht worden, blijkt, dat geen sterveling zich van den indruk van verklaring zoó geheel kan losmaken, dat zij in opzicht toch een zekere verzwarung van Broplevert. Wat Kennedy betreft, merkte rechter zijn richting-gevende samenvatting op, dat zijn allerlei tegenstrijdigheden en onverklaarbaar.
Scientific research is changing

Martijn de Groot

The World Wide Web is functioning as a science accelerator, says full professor in Knowledge Representation and Reasoning Frank van Harmelen who is based in Amsterdam. That is changing things. But this change barely compares to what is yet to come.

‘Of course the web has changed our lives as scientists. As researchers, we now blog and hyperlink as if our lives depend on it,’ said the computer scientist in Amsterdam when he, quite appropriately, held the foundation day speech at the Free University (VU) on 20 October 2009 in the presence of honorary doctor and web inventor Tim Berners-Lee. He continued: ‘But that’s all same old same old. I argue that the web will have a much more profound impact on science in the near future than what we currently see around us.’

And Van Harmelen put it as follows at the jubilee symposium of DANS: ‘Whereas Web 1.0 was a network of images and text and Web 2.0 one of communities – by groups of people for groups of people – Web 3.0 must be the semantic web, the network of our data: By computers for computers and hence useful to people.’

Particularly two consequences of the web’s existence will be important, Van Harmelen already explained in the accompanying issue of the quarterly e-data&research: In the first place, the manner of publishing and the possibility of exchanging data, and secondly, the possibility of using the web as a data source. ‘So far, publishing took place in scientific articles, which in fact is a sort of official farewell ceremony for scientific results. You create hypotheses, tables and plots, and no one else can use them afterward. No one can verify the data – because no one can access them. But neither can anyone use them for different, new purposes. That is all going to change. Now we can publish the data so that everyone can have them at their disposal. And right away, as opposed to a few years after the researcher has wrapped up his or her work.’

Like a wave

This leads to the emergence of a plethora of new possibilities, argues Van Harmelen, and not just for continuing someone else’s work but notably also for combining data and hence answering completely new research questions. This way of publishing is already quite common in some of the natural and physical sciences, such as astronomy, particle physics and the life sciences. But it isn’t yet in a few other disciplines, like chemistry and the social sciences. ‘This movement is traversing science like a wave, however.’ It does require more than the mere possibility for exchange, emphasizes the computer scientist. ‘You have to do it in such a way that the data are usable to others. The semantic web has delivered a nice standard to that end: The Resource Description Framework – or RDF – which describes ‘things’ and the ‘relationships between things’. Data sets consist of objects and relationships between objects, and of
variables and relationships between variables. RDF lets you describe the objects, variables, and the relationships among them for every data set. You end up with a labeled network of data. That makes it easier to reuse them and particularly for having that done by computers. In the life sciences, this is already fairly normal. There, they convert their data to RDF format as soon as they start sharing. That makes it easier to reuse them.’

**The web as observatory**

The social sciences already have examples of projects that use this approach to arrive at promising results, according to Van Harmelen. Communication scientists, for example, have their computers carry out content analyses of media items – instead of using a small army of students or stay-at-home moms – and store the data in RDF so that they can easily link them to data from many other sources. This turns the web into an observatory, on top of it being an efficient medium for data exchange. The VU’s Network Institute, an alliance that includes computer scientists, economists and social scientists, studies the effect of media on politics. It does that in conjunction with communication scientist Jan Klein Nijhuis. ‘In the past, they would hire small armies of stay-at-home moms and students, literally, to collect data for a content analysis. Nowadays, a computer reads those data and stores them in RDF format, thus linking them to all sorts of other studies and analyses.’

The Network Institute is also looking into the scientific process, together with Professor Peter van den Besselaar of the Rathenau Institute. Besides citation analyses – ‘that is not a very large data set and it lags five years behind,’ says Van Harmelen – it uses the entire web. ‘Scientists do a lot more than just quote and publish. They also attend conferences and write on blogs, and as it’s all visible on the web, you can measure it. In the long run, you no longer will have to hold researchers accountable on the basis of their publications alone, but there will be new methods for measuring the effect and the quality of their work.’

**Data want to be free!**

A similar approach – using the web as observatory – is applied in a study by organization scientists at the VU, into knowledge networks among
businesses. ‘Is it possible to predict startups as a function of the way of participation in knowledge networks? In the past, you would send questionnaires for such a study, and you would be only too pleased if your response rate was 30%. Increasingly, those networks can be traced on the web. Just look at LinkedIn.’ One condition is that as many data as possible are freely available. ‘Data want to be free!’

That leads to the question as to how reliable the sample quality is. Old-fashioned questionnaires did allow you to determine the exact composition of the sample, but if you grab data off the web, you have to wait and see who did and who did not place theirs online. Van Harmelen: ‘In the old days, you didn’t know who was going to reply so the problem was essentially the same. And of course you can still send questionnaires. The point is that an entirely new observatory has been added!’

**What to do**

The developments are rapid and the perspectives are very attractive, Van Harmelen feels. But they’re not only attractive. There is such a thing as going too fast, and he quotes part of the title of one of his publications: ‘What to do when success is becoming a problem?’

*This contribution is an edited version of a previous article by Martijn de Groot in the quarterly e-data&research*

---

**Figure 2.** Computer generated map showing relationships between weblogs. This map shows blogs as coloured spots and lines represent the links between them. Nodes within the centre are of a more technical nature; nodes at the outer part of the sphere are non-technical blogs. Colours represent the URL. This image was created by Matthew Hurst.
About the authors

Peter Doorn is director of Data Archiving and Networked Services (DANS). He obtained his Master’s degree and PhD in social geography from Utrecht University. From 1985 to 1998, he taught computer use in history at the University of Leiden. He was director of the Netherlands Historical Data Archive and department head at the Netherlands Institute for Scientific Information Services (NIWI). He acquired and managed a significant number of externally financed digitization projects and other projects at the interface of history and computer science. At the moment he is involved in DARIAH (Digital Research Infrastructure for the Arts and Humanities) and Life Courses in Context, a collaboratory for the study of 19th and 20th century population history.

Martijn de Groot is an independent professional writer and communications advisor. He studied communication at Wageningen University, where he graduated in 1981. Subsequently, he was science editor at the Landbouw-Economisch Instituut, head of communications at the Westelijke Land- en Tuinbouworganisatie, and editor-in-chief of national agricultural weekly OOGST. In 2005 and 2006, he was also editor-in-chief of the monthly magazine InformatieProfessional and still is editor-in-chief of the quarterly e-data&research, which he established for a number of institutions under guidance from DANS in 2006.

Henk den Heijer is professor in maritime history at the University of Leiden. He researches the Dutch presence in the Atlantic region between 1600 and 1800. In 1997, he defended his doctoral thesis Goud, ivoor en slaven, a study of Dutch West-African trade and shipping between 1674 and 1740. Den Heijer wrote several books and articles on maritime history, including De geschiedenis van de WIC.

Esther Jansma is dendrochronologist and project manager of the DCCD project on behalf of the Cultural Heritage Agency. She is senior researcher at this agency, professor in dendrochronology and paleo-ecology of the Quaternary at the Faculty of Geosciences (Department of Earth Sciences) of Utrecht University, and scientific director of the Foundation RING (Nederlands Centrum voor Dendrochronologie).
Maarten Marx graduated in political science (1990) and obtained a PhD in mathematical logic (1995), both at the University Amsterdam where he now works as an assistant professor at the Informatics Institute. His current interest is the integration of large quantities of semi-structured text-centric data. Part of it takes place within his PoliticalMashup project (see http://www.politicalmashup.nl).

Johan van Rooijen studied biology and has been working at Statistics Netherlands since 1998. He has been involved in the development of various register-based statistical processes as well as in the development of statistical output. He currently works on the production and further development of the Social Statistical Database. He is also associated with the Netherlands Centre for Biodiversity Naturalis where he carries out taxonomic research.

Ewoud Sanders is language historian and journalist. He has a weekly language column in NRC Handelsblad. In the past few years, he was intensively involved in digitizing and digitally unlocking (old) books and magazines. He now has a digital library with over 20,000 volumes, 5.5 million pages and 3 billion words.

Marion Wittenberg studied sociology and is project manager at DANS. Among other things, she deals with data projects in the area of the social sciences, the organization of symposiums, and student communication.
Driven by data
On 2 December 2009, the symposium ‘Driven by data’ took place in The Hague to celebrate the anniversaries of a number of archiving institutions that are now all part of DANS (Data Archiving and Networked Services). It was a lively meeting that paid attention to various surprising and creative possibilities associated with the reuse of electronic data. DANS decided to publish this jubilee volume to give a larger public a taste of this versatility and creativity. The contributions cover topics as diverse as dendrochronology, politics, shipping traffic history, bird migration, philosophy of science and statistics.

DANS (Data Archiving and Networked Services) is the national organization in the Netherlands for storing and providing permanent access to research from the humanities and the social sciences. DANS manages existing data archives and works on further developments of the data infrastructure in new fields as well.

www.dans.knaw.nl