Lies, Damned Lies and Research Data: Can Data SharingPrevent Data Fraud?

Doorn, P.K.; Dillo, I.; van Horik, M.P.M.

published in
International Journal of Digital Curation
2013

DOI (link to publisher)
10.2218/ijdc.v8i1.256

document version
Publisher's PDF, also known as Version of record

document license
CC BY

The International Journal of Digital Curation
Volume 8, Issue 1 | 2013

Lies, Damned Lies and Research Data: Can Data Sharing Prevent Data Fraud?

Peter Doorn, Ingrid Dillo and René van Horik,
Data Archiving and Networked Services

Abstract

After a spectacular case of data fraud in the field of social psychology surfaced in The Netherlands in September 2011, the Dutch research community was confronted with a number of questions. Is this an isolated case or is scientific fraud with data more common? Is the scientific method robust enough to uncover the results of misconduct and to withstand the breach of trust that fraud causes? How responsible and reliable are researchers when they collect, process, analyse and report on data? How can we prevent data fraud? Do we need to adapt the codes of conduct for researchers or do we need stricter rules for data management and data sharing?

This paper discusses the conclusions and recommendations of two reports that were published recently in consequence of this data fraud. The reports are relevant for scientific integrity and trustworthy treatment of research data. Next, this paper reports on the outcomes of enquiries in data cultures in a number of scientific disciplines. The concluding section of this paper contains a number of examples that show that the approach towards data sharing is improving gradually. The data fraud case can be regarded as a wake-up call.

International Journal of Digital Curation (2013), *8*(1), 229–243. http://dx.doi.org/10.2218/ijdc.v8i1.256

Introduction

After a spectacular case of data fraud in the field of social psychology surfaced in The Netherlands in September 2011, the Dutch research community was confronted with a number of questions. Is this an isolated case or is scientific fraud with data more common? Is the scientific method robust enough to uncover the results of misconduct and to withstand the breach of trust that fraud causes? How responsible and reliable are researchers when they collect, process and analyse and report on data? How can we prevent data fraud? Do we need to adapt the codes of conduct for researchers or do we need stricter rules for data management and data sharing?

In the course of 2012, two more cases of inappropriate conduct in the treatment of data in The Netherlands were extensively discussed in the press and in online media.

The first and most spectacular case of data fraud that emerged was the Stapel Affair (Wicherts, 2011). The social psychologist Diederik Stapel was an influential professor until a group of young researchers, on the basis of outcomes which they considered "too good to be true", smelt a rat. In the final report, which was recently published by a committee that evaluated all of Stapel's work at the universities of Tilburg, Groningen and Amsterdam, it is now clear that 55 of his publications are based on fictitious data.[1]

Two other cases in which data were at least treated unscrupulously came to light in 2012, both at Erasmus University, Rotterdam. The seriousness of the scientific sin committed by Professor Poldermans at the Erasmus Medical Centre is not entirely clear: although he was suspected of committing scientific fraud by fabricating data, it could only be proven that he was guilty of negligence in the collection and treatment of research data. The final report on this affair accuses the researcher of serious scientific misconduct, including "failure to preserve essential source documentation," so that "further verification and analysis " was impossible. Moreover, he failed "to record actual medication use in a study focusing on an intervention using medicines"; and he did not obtain informed consent from his patients, or did so improperly (Erasmus MC Follow-Up Investigation Committee, 2012). In another case, the marketing researcher Dirk Smeesters resigned in June 2012 after an investigative panel found problems in his studies and concluded it had "no confidence in [his] scientific integrity." This researcher conceded to "massaging" his data in some papers to "strengthen" outcomes, while defending his actions as common in his field (Committee for Inquiry into Scientific Integrity, 2012).

In consequence of these data fraud cases, two reports were published that are relevant for scientific integrity and trustworthy treatment of data: the report on Scrupulousness and Integrity in dealing with Scientific Research Data by a committee chaired by Professor Kees Schuyt (Committee on Scientific Research Data, 2013); and the final report on the Stapel affair by three committees (abbreviated here as 'Flawed Science' or The Levelt Report after Professor Pim Levelt, chair of one of the committees).

[1] The full report of the committees chaired by Levelt, Noort and Drenth is published on the web under the title "Flawed science: The fraudulent research practices of social psychologist Diederik Stapel"; see: http://www.commissielevelt.nl

The Schuyt Report

The Committee on Research Data, chaired by Professor Schuyt and installed by the Royal Netherlands Academy of Arts and Sciences (KNAW), presented their advisory report in September 2012 recommending greater alertness and peer pressure, adherence to the code of conduct for researchers and better access to research data.

One of their key recommendations was for scientists to ensure that data are properly archived and thus available for review, replication and other use. Surveys and case studies carried out by us (and others) over the past few years have shown that data archiving practices and cultures of data sharing vary widely among individual researchers and scientific disciplines, but they can also change quite suddenly and rapidly.

The committee interviewed scientists and scholars in a variety of disciplines about the research practices in their field, in particular with respect to data management. As one might expect, data management practices appear to vary a lot across disciplines, but also within fields. If one regularity exists, it is that in small-scale research the researcher has to do pretty much everything alone, with very little support and very few checks and balances; whereas in large-scale science, where research is carried out in groups, professional data management arrangements are in place. Moreover, if we regard the whole research cycle, it is especially the phase in between the formulation of a project proposal and the writing of a paper to be published, that supervision or intervention by peers can be lacking, especially in "small science" and individual projects. Many PhD projects are individual enterprises and PhD supervisors tend not to spend much time on data management issues. The Schuyt committee thinks that unsupervised, solitary work is the most risky area of a research project, where sloppiness in the collection and treatment of data may creep in, or where researchers may be tempted to fiddle with their data in unacceptable ways.

Two concepts are central in the report: scientific scrupulousness (in Dutch: *'zorgvuldigheid'*) and scientific integrity. Both concepts must also be distinguished. Whilst the first has to do with how meticulously, accurately, precisely, strictly or exactly a research project has been carried out, the second has to do with honesty.

There is also the aspect of responsibility or accountability: choices made in research and data management should be accounted for. Data is often transformed, processed, manipulated, enhanced or enriched before results can be found. Whether it is statistical, image or text analysis, a form of pre-processing of the data is often indispensable. Outliers need to be removed for certain statistical tests that require a normal distribution[2]; images need to be enhanced to see what remains otherwise invisible; disambiguation or other forms of linguistic pre-processing may be needed to do a content analysis on unstructured texts. This is all allowed, provided it is accounted for, reported and documented.

[2] The well-known phrase: "There are three kinds of lies: lies, damned lies, and statistics" was attributed to the 19th-century British Prime Minister Benjamin Disraeli (1804–1881) by Mark Twain. However, the phrase is not found in any of Disraeli's works and the earliest known appearances were years after his death.

One might say that scrupulousness is at stake when important decisions taken in the research process and in the data management are not documented and can not be accounted for. In fact, the research may become so sloppy that it becomes impossible to establish the integrity of the research and the researcher. This was quite clearly the case in the Poldermans' affair: whether data was deliberately falsified or whether it was merely a case of 'flawed science' cannot be established because both the original data and the documentation are lacking. Here we see that a lack of scrupulousness and a lack of accountability are destructive for trustworthiness and raise suspicions about a researcher's integrity to such an extent that his scientific reputation can be damaged.

Data Quality

Authenticity of Data

The authenticity of data receives a lot of attention from the data archiving community, for instance in the APARSEN project.[3] The debate about authenticity is usually connected to the provenance of the data and what data curation organisations can do to establish that provenance. How far can one go in establishing the provenance and hence the authenticity of data? In a way, the whole data cycle would need to be recorded and verified – or rather, be verifiable – to be sure that data is authentic and "true". However, in most cases this is not feasible in practice. Moreover, making sure research data is verifiable is not the primary responsibility of the curator of the data, but of the creator: the data curator can only require the depositor of data to document his or her work. Data fraud could be seen as the extreme instance of data *in*authenticity. To some extent, a depositor has to be trusted that the data they offer for archiving is reliable.

Trust is another big subject in the APARSEN project and in data curation in general. When we speak about trustworthy digital repositories, we usually refer to the procedures of the repositories, which need to be sound and well-documented. But what about the procedures of the data creator, the researcher who sampled, collected, digitized or measured the information?

Quality Checking by Data Curators and Data Archives

In discussions among data curators or archivists about data quality, the subject of fraudulent or falsified data rarely arises[4]. The quality of research is usually measured in terms of scientific output. The quality control of scientific output, especially as published in journals, has traditionally been assured through the system of peer review. For electronic publications, new ways of reviewing articles in digital repositories have been sought and implemented since the mid 1990s (Harnad, 1996). Applications for project funding are also usually peer reviewed.

However, the databases underlying scientific publications are seldom reviewed, although an increasing number of journals are requiring the submission of such datasets in publicly accessible data repositories. Such journals have been labelled as

[3] APARSEN project: http://www.aparsen.eu

[4] Parts of Chapter 3 of the e-IRG Report on Data Management by the Data Management Task Force (e-IRG, 2009) were edited and updated to reflect this.

"DAP-Journals' or journals with a Data Availability Policy (de Moor & van Zanden, 2008). Data archives, which have been established since the 1960s, have always used the potential for checking possible errors in data collections as a rationale for their continued existence. Such digital data archives are the main advocates of quality assurance for research data. Quality control by data archives is usually achieved by painstaking and labour intensive checks on the data, carried out by data archive staff. Quality checks include checking the format of the data files, whether a codebook is available for coded data, whether personal data has been anonymized, and whether the dataset is complete and consistent with the descriptive metadata. All these checks are, of course, rather formal and do not guarantee the validity of the content of the dataset.

Data archive staff are usually not in a position to assess the quality of the content of the data itself. The users of the data are best equipped to scrutinize and judge the value of the data. They are the most likely to come across omissions, errors or even downright forgeries. Of course, there is no guarantee that problems will be identified, but nevertheless, if peer specialists do not notice problems in a dataset, who else can? As yet, few data archives and data journals have taken initiatives to implement systematic data reviews by users, but DANS is one of them (Grootveld & van Egmond, 2011; 2012).

Certification and Data Quality

The initiatives for certifying digital repositories usually concentrate on quality assurance procedures for long-term preservation by data storers and not on requirements for data providers.[5] The Data Seal of Approval (DSA)[6] is an exception, as its guidelines focus on three groups of actors: data producers, data repositories and data consumers. Data producers are held responsible for the quality of research data, repositories for storage and long-term access quality, and users for the correct use of the data.

According to the DSA, the quality of digital research data is determined by its intrinsic scientific quality, its data formats and supporting information, and its documentation. The first requirement of the DSA is that data producers deposit their data in a repository with sufficient information for users to assess the scientific and scholarly quality of the data, in compliance with disciplinary and ethical norms. Moreover, they should provide the research data in appropriate formats and with the metadata required by the data repository. Scientific quality criteria indicate to what degree the research data is of interest for scholarly use. The assessment by experts and colleagues in the field is the main decisive factor for the scientific quality of research data. Three questions must be answered to be able to provide an assessment:

- **Is the research data based on original work performed by the data producer and does the data producer have a solid reputation?** This question can be answered by providing information regarding the researcher and/or research group and by providing references to publications pertaining to these particular research data.

[5] For more information see the European Framework for Audit and Certification of Digital Repositories website: http://www.trusteddigitalrepository.eu

[6] Data Seal of Approval: http://www.datasealofapproval.org

- **Was the data creation carried out in accordance with prevailing criteria within the research discipline?** The answering of this question requires information on the methods and research techniques used, including those for data collection, digitization or other means of data creation.
- **Is the research data useful for certain types of research and suitable for reuse?** The answer requires information regarding the data format, content and structure. The data producer must provide sufficient information to enable fellow scientists to assess the research data.

Other Data Quality Guidelines

The OECD has published a set of thirteen principles and guidelines for access to research data from public funding, among which several relate to data quality and one is explicitly labelled "quality". This guideline says, among other things:

> "The origin of sources should be documented and specified in a verifiable way." (OECD, 2007).

The Strategic Committee on Information and Data (SCID) of the International Council for Science (ICSU) stresses the importance of data quality, although it does not state how this can be guaranteed (ICSU, 2008). A report commissioned by the Research Information Network (RIN) identifies three key purposes of quality assurance in the data creation process (Jubb, 2008). With regard to creating, publishing and sharing datasets, the RIN report states that:

1. Datasets must meet the purpose of fulfilling the goals of the data creators' original work;
2. Datasets must provide an appropriate record of the work that has been undertaken so that it can be checked and validated by other researchers;
3. Datasets should be discoverable, accessible and re-usable by others.

Fulfilling the first and second points implies a focus on scholarly method and content; the third implies an additional focus on the technical aspects of how data is created and curated. The scientific or scholarly value of datasets that are not accessible for reuse by others can obviously not be assessed by independent peers.

The RIN report distinguishes datasets created by machines (such as telescopes, spectrometers, gene sequencers) from those created in other ways (such as social surveys, databases created by manual input, source editions of texts, etc.). This distinction roughly, although by no means completely, coincides with the distinction between the sciences and the humanities. Machines that create data often have inbuilt data validation mechanisms. Manual checking is usually added, and in those disciplines where data are collected by other means, manual verification may involve very detailed work.

Although there is no information on how many datasets are checked by others than the researcher, in many cases it is taken for granted that when a paper is accepted for publication after peer review, the underlying data will pass the quality standard as

well. Peer review may sometimes involve checks of the supporting data. In some disciplines, reviewers do checks on data. In other cases, checking is superficial or absent because the data is too complex or voluminous to be judged satisfactorily. Most researchers take other researchers' outputs on trust in terms of data quality and integrity. Moreover, there are no apparent signs of dissatisfaction with this state of affairs.

The accessibility of data for reuse renders checking at a later time possible. Experimental, machine-produced data can, in principle, be re-created if the whole experiment is done again. In practice, such validity-checking procedures are rarely carried out. Nevertheless, new experiments with better measuring equipment, or the secondary analysis of survey data or text corpora, re-appraisal of digital scholarly editions and so on, may result in the discovery of earlier flaws, and in extreme cases in the exposure and shaming of mistakes or even fraud.

Although it is useful to distinguish between the scientific or scholarly content of data and the technical merits that facilitate reuse, it is questionable whether the two can be separately reviewed, as recommended in a report commissioned by the Arts and Humanities Research Council (2006). British researchers generally support the idea of instituting a formal process for assessing the quality of datasets, although they have concerns whether it will work effectively in practice. Among the potential problems is the difficulty of finding reviewers who are willing and who have the expertise to understand and appraise the data.[7] Another concern involves the costs (in terms of money and time) of a formalized data review process.

A Role for Funders

It is unlikely that the pressure to improve the quality assurance process for datasets will come from the researchers, although they generally seem to favour a more thorough assessment. Research funders, who are investing heavily in data creation and increasingly in the data infrastructure at large, are in a better position to take the initiative and introduce a formal assessment process. This also implies that data creation itself should be rewarded with scholarly merit and scientific credits. Funding agencies increasingly require a formal data management plan to be part of a project proposal.

Whilst datasets that are deposited at data centres must conform to certain quality standards, there is no such imperative yet for researchers who look after their own datasets. According to the RIN study, some researchers believe this to be outside the boundaries of their research function, while others lack the skills (and/or time) to publish their data in such a way that it can be discovered, accessed and reused by the scholarly community.

[7] In a study of researchers' attitudes to peer review, 40% of reviewers and 45% of journal editors said it was unrealistic to expect peer reviewers to review authors' data (Ware, 2008).

Data Cultures Across Scientific Domains

In 2010, DANS held a web survey among randomly selected researchers from the Dutch Research Database[8] for a quantitative picture of these researchers' opinions on data sharing, data storing and data usage. This section provides the main results of this survey. 556 researchers from random research disciplines in the Dutch sciences and humanities answered the questions of the web survey. The results were not far off from those of an international survey published in 2009 (Kuipers and van der Hoeven, 2009). A number of topics raised in the survey are summarized below.

Storage of Own Research Data

The researcher's own computer is still the most popular place for storing data from scientific research. More than 70% of researchers in the Netherlands are mainly using their own computer as their archive. Against this high proportion of 'home storage researchers' is a small minority (18%) mentioning external storage facilities for research data outside their own department or institute. External hard disks or backup media are mentioned by 46%, while 63% say they use a network disk of their own department or institute. As one would expect, the interviewees work with many different types of data. Spreadsheets (69%) and instrument data (61%) are most popular among researchers working in the exact sciences. Nevertheless, the respondents from the humanities and social sciences also often work with statistical data (68%) and spreadsheets (53%).[9]

Use of Other People's Data

The vast majority of researchers surveyed appear to use data collected by others, mostly on top of data they collected themselves. Just 30% of researchers surveyed work on the basis of their own data only. However, within this group, six out of every ten researchers admit they would occasionally need data from sources other than their own research. When asked if there are sufficient possibilities for making use of other people's data, 42% of the interviewees responded in the affirmative. Of the remaining 58%, equal numbers believe the possibilities are insufficient or they do not know. Even though four out of ten researchers believe the opportunities are sufficient, it still leaves the fact that 58% of all interviewees report obstacles and objections when trying to make use of other people's data. Almost three quarters of them report 'no access or limited access' as the main reason. Other reasons include problems with the quality of the required data (50%), privacy considerations (41%), missing metadata and insufficient screening of reliability (both 36%).

The Use of 'Own' Data by Others

While 70% of all interviewees say they make use of data from sources other than their own, only 61% say that others sometimes also make use of their data. About a third of the respondents say they have not experienced any data use by others. This may very well have to do with the objections some people have to the use of their data by

[8] The Dutch Research Database is part of the NARCIS Gateway to Scholarly Information in The Netherlands: http://www.narcis.nl

[9] More than one answer was permitted.

others. Although only 4% of the same group (n=417) say they have such objections in any event, a much larger group of 60% reports they object 'in some cases'. How often they objected was not asked. The differences among the various disciplines are notable. There is a marked difference in terms of the most often heard objection (possible misuse of data by others) among the exact sciences on the one hand and humanities and social sciences on the other hand (see Figure 1).

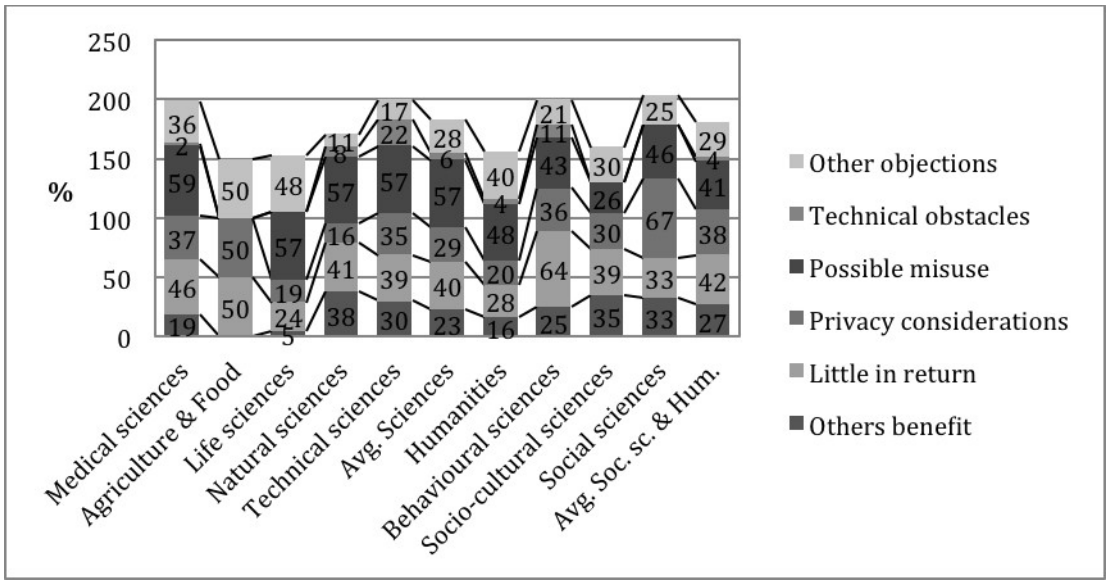

Figure 1. Percentages of researchers quoting objections to sharing their own data. More than one objection can be given (n=242).

Durability of Research Data

About two-thirds of the respondents state that until now they did not experience any difficulties related to the threat of research data getting lost. Software obsolescence is mentioned as the main problem by 28% of the respondents that experienced difficulties, followed by hardware problems and lacking metadata. But almost half of the respondents expect to encounter durability problems in the near future. Concerning the period data will be usable, a wide variety in opinions can be noticed. About 30% of the respondents expect that data will last for six to ten years.

The outcomes of the survey in 2010 revealed that there is a growing awareness of the importance of data management. The survey also makes clear which issues obstruct the further development of a transparent data infrastructure. These issues can be used for further actions to improve the situation.

Psychology: A Closed Data Culture

In 2010, a year before the data fraud described earlier in this paper occurred, DANS published a study on the attitudes of Dutch academic psychologists in relation to their research data (Voorbrood, van Luijn, 2010).

The first outcome of this study was the observation that psychologists hardly have a tradition of sharing data. Even though their opinions may differ widely, most psychologists would not even consider making their data available for verification or a second opinion by a colleague or outsider. Besides this, much of the data gets lost over the course of time because no clear rules exist for archiving data within the profession. 70% of the psychologists keep data on their own PC, on CD/DVD or "in a drawer". About one in five stores their data on a server run by the faculty that is usually not dedicated to long-term storage. Only very few deposit their data in an electronic archive, such as the online archiving system EASY,[10] run by DANS.

Secondly, the study uncovered substantial differences between sub-disciplines within psychology in terms of their view on data sharing and data archiving. For instance, psychologists who are concerned with particular populations, or those working with large datasets or on longitudinal studies, have a higher stake in (and therefore more attention for) making data available than their colleagues working on experimental studies. Some psychologists are familiar with making data available and/or reusing data. In most instances this seems to be the result of prior arrangements or individual requests, rather than a blanket arrangement for making data available for further scientific use or verification purposes. Mutual data sharing is far from being common practice among psychologists, but the situation is even worse in terms of sustainable archiving. Making backups of research data is generally accepted. However, in the absence of adequate data management and means for archiving, databases and backups often get lost over time or turn out to be unusable.

In spite of these differences between sub-disciplines, the discussion in general is dominated by arguments in favour of keeping research data for personal use only. Not only is this happening in the Netherlands, but also, for example, in the USA. This closed data culture does not match the code of conduct of the American Psychological Association (APA), which stipulates that data must be made available for verification if research is funded with public money.[11]

Six (specious) arguments why researchers are against data sharing, together with how to refute these arguments, are given below:[12]

- **No one else will be able to understand the complexity of my data.** Other fields of study have proven that this can be done if the circumstances of both the research and the data are properly documented and described.
- **Someone else analyzing my data may find a different answer that could discredit my findings.** In fact, falsification of statements forms the basis of the scientific method; by considering different perspectives on the same data set, we will come closer to the "right" answer.

[10] EASY: https://easy.dans.knaw.nl/

[11] APA Guidelines: http://www.apa.org/research/responsible/data/index.aspx

[12] These arguments are written by Stephen H. Koslow, former director of Neuroscience and Behavior of the National Institute of Mental Health (NIMH) in the United States. They have been freely adapted and tuned to a Dutch context.

- **Another person may find something in my data that I have overlooked.** Finding something new in an existing data set will increase the return on investment in the data collection.
- **I am still in the process of analyzing my data.** Any research is an ongoing process; a published paper suggests that the data have been substantially analyzed, thus sharing at this point would seem appropriate.
- **It is my data that I worked very hard to collect, and no one else has the right to it.** Publicly funded data should be publicly available; publication of a study already implies that its results and conclusions are to be shared. If these are to be evaluated in detail, reviewers and readers should have access to the primary data on which they are based.
- **I cannot trust or understand the data produced in another laboratory.** If this is not possible, whom can we trust in scientific literature? This is the mirror image of the first argument.

The DANS study shows that there is a remarkably closed data culture among psychologists. Such a culture is not without risk. The Stapel Affair revealed that in cases where data verification is nonexistent, manipulation of research data could easily go unnoticed. The findings of the Levelt Committee seem to prove this. A culture permeated by 'flawed science' surrounded Stapel. This is one of the reasons why his academic misconduct went undetected for so long.

Although Stapel is fully responsible for this extensive case of academic fraud, the Levelt Committee is also critical of the research culture in which this academic misconduct was allowed to go undetected. The Committee describes this as "a general culture of careless, selective and uncritical handling of research and data." It concludes, that "...from the bottom to the top there was a general neglect of fundamental scientific standards and methodological requirements." The Committee points the finger not only at Stapel's peers, but also at editors and reviewers of international journals.

The interesting question now is whether the discipline will draw its conclusions and will try to develop a more open data culture. In 2010 the psychologists themselves were convinced that in five years time data sharing and data archiving would be much higher on the agenda of their community. The first hesitating steps towards more transparency are indeed being taken both inside and outside the community of psychologists. However, it is still too early to determine whether this is the beginning of a real change.

Conclusion: Signs of Change

The data fraud affair and sloppy data practices that came to light during the past year, regrettable as they are, are also a wake-up call for psychology and other fields where the awareness of the importance of data management and data archiving is still deficient. However, let us remember that both the awareness and the practice of data management are themselves susceptible to change. In fact, there is evidence from various fields that the situation can change fairly quickly. A number of examples are given below.

Archaeology

In archaeology, experience shows that things can change rapidly. Ten years ago, hardly any digital data of archaeological excavations were kept on record in the Netherlands. Following in the footsteps of the Archaeological Data Service in the UK and using the slogan 'Digital archaeology needs a digital memory', archaeologists in collaboration with DANS and the Cultural Heritage Agency have changed this situation within a short period of time. Since 2007, archaeologists are even compelled to deposit their data. The e-depot for Dutch archaeology now administers approximately 19,000 digital archaeological datasets and allows access for reuse. As an excavation can be carried out only once, archaeologists now generally recognise the need for a digital archive.[13]

Oral History and Qualitative Social Science

In the last couple of years in the Netherlands a number of interview collections were created that can be used in oral history research and qualitative social science research. Increasingly, these collections are accessible as Open Access datasets or have a licence that enables usage for research purposes. A collection of about 1,000 interviews with Dutch war veterans serves as a good example. These interviews were collected by the Netherlands Veterans Institute and made available via DANS for secondary analysis. The collection was used in a multi-disciplinary research project. An enhanced publication of the research outcomes contains links to the related interview data. In this way data transparency is optimized.[14]

Virology and Veterinary Medicine

The Italian virologist and veterinarian Ilaria Capua was the first to unravel the DNA-sequence of the African subtype of the bird flu-virus H5N1. She placed that sequence not in the then protected repository of the WHO, which was accessible worldwide by only fifteen labs (among which not one African institute), but in an open archive. After an initial storm of protests by colleagues, that even reached the front page of the Washington Post and an editorial in the New York Times, now the scientific practice has completely turned upside down and open access has become the norm (Capua, 2007).

Psychology

In response to the fraud case in Dutch psychology, several initiatives have been undertaken to come to a more open data culture. This year the School of Behavioural Sciences of Tilburg University undertook an initiative with DANS to publish a Journal of Open Psychology Data (JOPD)[15]. An international editorial advisory board assists the Dutch editor, Jelte Wicherts. The journal will feature peer reviewed data papers

[13] See: http://www.dans.knaw.nl/en/content/categorieen/projecten/edna-e-depot-dutch-archaeology and http://www.edna.nl/

[14] The "enhanced publication" containing the research outcomes can be found at the Dutch site: http://www.watveteranenvertellen.nl/. Background information on the interview collection and its usage can be found in Scagliola (2011).

[15] JOPD: http://openpsychologydata.metajnl.com/about/

describing psychology datasets with high reuse potential. The journal will focus on the potential replication of research, and will also pay attention to "negative results", i.e. research that did *not* confirm the initial hypotheses. The papers will contain concise descriptions of datasets, including where to find them.

Papers for JOPD will only be accepted in cases where authors agree to make related datasets freely available in a public repository. A data paper is a publication that is designed to make other researchers aware of data that is of potential use to them. The paper describes the methods used to create the dataset, its structure, its reuse potential, and a link to its location in a repository.

In September 2012, DANS sent a mailing to eight hundred Dutch psychologists drawing their attention to the online archiving system EASY. They received a fact sheet explaining the seven steps needed to deposit data in EASY. Unfortunately, this initiative has not led to an increase in the deposits of psychology data in EASY as yet. The only effect of the initiative seems to have been an increased interaction with the psychology discipline. Requests were received for presentations on data management and archiving and for cooperation.

Conclusion

Experience concerning data cultures shows that they are susceptible to change, but this change does not always come easily or rapidly. There are clearly forces to resist such changes. In the case of psychology, a continued dialogue is needed if we want to bring about a change with respect to the culture of data sharing.

In his foreword to the Schuyt Report, the president of the Academy of Arts and Sciences Hans Clevers states, that:

> "Maximum access to data promotes the pre-eminently scientific approach whereby researchers check one another's findings and build critically on one another's work" (Committee for Inquiry into Scientific Integrity, 2013)

The Academy supports the free movement of data and results, and endorses the recommendations of the Schuyt committee. Taking into account the variations across and within scientific disciplines, free availability of data should be the default. Moreover, it is not so much additional rules of conduct that are necessary, but the revitalisation of existing rules. Next, the examination of data management practices should become an integral part of official research evaluations. Both the committee and the Academy think that existing codes of conduct suffice: we do not need more rules but should focus our attention on making the existing rules better known.

Better access to research data contributes to the transparency of science. Archiving and publishing data, via trustworthy data repositories, adds to this. Funders should require that project proposals contain a data management plan, and that such plans contain a section on the accessibility of the data after publication of the results.

References

Arts and Humanities Research Council. (2006). *Peer review and evaluation of digital resources for the arts and humanities.* Swindon: Author. Retrieved from http://www.history.ac.uk/sites/history.ac.uk/files/Peer_review_report2006.pdf

Capua, I. (2007). *Ecology, epidemiology and human health implications of avian influenza viruses: Why do we need to share genetic information?* Paper presented at Berlin Open Access Conference 5. Padova, Italy. Retrieved from http://hdl.handle.net/10760/10882

de Moor T. & van Zanden, J.L. (2008). Do ut des (I give so that you give back): Collaboratories as a new method for scholarly communication and cooperation for global history. *Historical Methods, 41*(2) pp 67 – 80. doi:10.3200/HMTS.41.2.67-80

e-IRG. (2009). *Report on data management.* Report by the Data Management Task Force. Author. Retrieved from http://www.e-irg.eu/images/stories/e-irg_dmtf_report_final.pdf

Erasmus MC Follow-Up Investigation Committee. (2012). *Report on the 2012 follow-up investigation of possible breaches of academic integrity.* Retrieved from http://www.erasmusmc.nl/5663/135857/3675250/3706798/Integrity_report_2012-10.pdf?lang=en&lang=en

Committee for Inquiry into Scientific Integrity (2012). English translation of *Rapport Onderzoekscommissie Wetenschappelijke Integriteit*. Retrieved from http://www.eur.nl/fileadmin/ASSETS/press/2012/Juli/report_Committee_for_inquiry_prof._Smeesters.publicversion.28_6_2012.pdf

Grootveld, M. & van Egmond, J. (2011). *Data reviews: Peer reviewed research data.* DANS Studies in Digital Archiving 5. The Hague: DANS.

Grootveld, M. & van Egmond, J. (2012). Peer-reviewed open research data: Results of a pilot. *International Journal of Digital Curation, 7*(2), 81-91. doi:10.2218/ijdc.v7i2.231

Harnad, S. (1996). Implementing peer review on the net: Scientific quality control in scholarly electronic journals. In R. Peek & G. Newby (Eds.) *Scholarly Publication: The Electronic Frontier.* Cambridge MA: MIT Press.

ICSU. (2008). *Final Report to the ICSU Committee on Scientific Planning and Review.* ICSU, Ad hoc Strategic Committee on Information and Data.

Jubb, M. (2008). *To share or not to share: Publication and quality assurance of research data outputs.* London: Research Information Network. Retrieved from http://www.rin.ac.uk/our-work/data-management-and-curation/share-or-not-share-research-data-outputs

Kuipers, T. & van der Hoeven, J. (2009). *Insight into digital preservation of research output in Europe.* PARSE Insight Survey Report. Retrieved from http://www.parse-insight.eu

OECD. (2007). *OECD principles and guidelines for access to research data from public funding.* Author. Retrieved from http://www.oecd.org/dataoecd/9/61/38500813.pdf

Scagliola, S. (2011). *The Dutch veterans interview project: Recognition and attention in exchange for valuable information.* Paper presented at the Interuniversitary Seminar on Armed Forces and Society, Chicago. Retrieved from http://www.eur.nl/fileadmin/ASSETS/estudio/Bestanden/Dutch_Veterans.pdf

Committee on Scientific Research Data. (2013). *Responsible Research Data Management and the Prevention of Scientific Misconduct.* Advisory Report by the Committee on Scientific Research Data, April 2013. Royal Netherlands Academy of Arts and Sciences. Retrieved from http://www.knaw.nl/Content/Internet_KNAW/publicaties/pdf/20131009.pdf

Voorbrood, C. & van Luijn, H. (2010). *Data - Voer voor psychologen. Archivering, beschikbaarstelling en hergebruik van onderzoeksdata in de psychologie.* DANS Studies in Digital Archiving 4. Amsterdam: Aksant Academic Publishers. Retrieved from http://www.dans.knaw.nl/content/categorieen/publicaties/dans-studies-digital-archiving-4

Ware, M. (2008). *Peer review: Benefits, perceptions and alternatives.* London: Publishing Research Consortium. Retrieved from http://www.publishingresearch.net/documents/PRCPeerReviewSummaryReport-final-e-version.pdf

Wicherts, J.M. (2011). Psychology must learn a lesson from fraud case. *Nature 480*(7375), 7. doi:10.1038/480007a