PLANT RESEARCH INTERNATIONAL
WAGENINGEN UR

Embedded data scientist

Frits K. van Evert & Hugo A. Besemer

Report 552

PLANT RESEARCH INTERNATIONAL
WAGENINGEN UR

Embedded data scientist

Frits K. van Evert[1] & Hugo A. Besemer[2]

[1] Plant Research International
[2] Wageningen UR Library

Plant Research International, part of Wageningen UR
Business Unit Agrosystems Research
February 2014

Report 552

Copies of this report can be ordered from the (first) author. The costs are € 50 per copy (including handling and administration costs), for which an invoice will be included.

Plant Research International, part of Wageningen UR
Business Unit Agrosystems Research

Address : P.O. Box 616, 6700 AP Wageningen, The Netherlands
: Wageningen Campus, Droevendaalsesteeg 1, Wageningen, The Netherlands
Tel. : +31 317 48 05 73
Fax : +31 317 41 80 94
E-mail : info.pri@wur.nl
Internet : www.wageningenUR.nl/en/pri

Table of contents

1. Introduction

Digital research data offer unprecedented opportunities for reuse because they can easily be stored, moved, and used. At the same time, digital data are more easily lost than traditional (paper) data (unless a long-term storage and sharing scheme is designed and implemented). Consequently, there is a widely-felt idea that researchers should think carefully about how they are handling digital research data, both in terms of how to preserve them and of how to make use of them.

Some recent initiatives on agricultural research data include the Smarter Agriculture initiative at Purdue University[1], a workshop on 'Making Data Accessible to All' at University of Warwick[2], the iPlant Collaborative of the NSF[3], and the Global Open Data for Agriculture and Nutrition (GODAN) initiative[4]. Examples of data (and/or model) repositories are The Global Agricultural Trial Repository (AgTrials)[5], the Plant Production Systems Model Library[6] and facilities of the SEAMLESS Association[7]. The last few years have seen a lively debate on data in the opinion and letters pages of the scientific literature, with contributions in Science, Nature and many other journals, as well as in mainstream newspapers and magazines – the latter often focusing on data management as a means to reduce fraud.

The Library of Wageningen University and Research Centre considers that it has a role in digital research data. In 2011, the Library started a program 'Datasets' (7.710.031.624) with the goal of developing data curation services in close collaboration with researchers. With Wageningen Graduate Schools, a course 'Data management planning' and a template data management plan for Ph.D. students have been developed. The 'Datasets' program has also resulted in a search facility with which datasets registered in the national registry on research projects and research output (Metis / WaY) can be found.

The 'Datasets' program does not register datasets. The Library has no personnel with the specific role of 'data librarian', although, by now, the Library does have some experience with activities, such as adding metadata to datasets and transforming datasets to long-term readable formats, that are clearly an extension of the traditional tasks of a library. There is an ongoing discussion about the question whether data curation requires domain-specific specialists in addition to personnel with 'data librarian skills' (Swan and Brown 2008). But whatever the answer to this question may be, the Library cannot hope to hire data curation specialists for every one of the many disciplines studied at Wageningen UR.

The Library aims to develop its data curation services in collaboration with domain specialists from the Science groups of Wageningen UR. Thus, in 2013 a scientist (FKVE) from the Plant Sciences Group of Wageningen UR was contracted to work at the Library in the role of 'embedded scientist' (an 'embedded librarian' would be a librarian working with a Science group). The embedded scientist was tasked with the following:

- Systematically identify datasets for deposition with the Library, within a small number of Wageningen UR research groups.
- Document the method used to identify datasets, to facilitate similar work which is expected to take place with other research groups.
- Inform researchers about data deposition.
- Identify and document which services and facilities are needed to support data deposition.

---

1 https://ag.purdue.edu/arp/Pages/smarteragriculture.aspx
2 http://www2.warwick.ac.uk/fac/sci/lifesci/news/geworkshop/
3 http://www.iplantcollaborative.org/
4 http://www.godan.info/
5 http://agtrials.org/
6 http://models.pps.wur.nl/
7 http://www.seamlessassociation.org/

This report describes the outcome of the work of the embedded scientist. In §1.1 the methodology is described. The outcome of the interviews held with individual scientists is described in Chapter 2. Recommendations are given in Chapter 3.

1.1 Methods

The work described here focused on the business unit Agrosystems Analysis of Plant Research International (AGRO) and three Wageningen University departments: Plant Production Sciences of (PPS), Centre for Crop Systems Analysis (CSA) and Farming Systems Ecology (FSE). At each of these units, a presentation on data deposition was given during a lunch seminar (see Appendix I).

Interviews with individual researchers and assistants of the above named departments were held by FKVE. The aim of the interviews was twofold:

- Inform the researcher about facilities for data deposition with the Library
- Explore whether the researcher has datasets that he/she is willing to deposit with the Library

In principle, individual researchers were approached after the seminar was held, but this could not be maintained in all cases. Some people were interviewed before the seminar, some people could not attend the seminar, and some people upset the scheduling by attending the seminar with another department than their own.

When a dataset suitable for deposition was identified, a follow-up interview was held by Hugo Besemer and FKVE to gather details of that dataset. The detailed files were then sent to Anne-Marie Patist for the actual deposition. In some cases, the follow-up interview was not necessary and the data files were sent directly to Anne-Marie Patist.

2. Interviews

In principle all researchers at AGRO, PPS, CSA and FSE were considered for an interview. Not all researchers were interviewed, however, for several reasons, including (1) not willing to spend time, (2) not interested, (3) scheduling conflicts, (4) informal discussion yielded sufficient information.

A total of 31 interviews were held. Interviews were conducted in Dutch or English. The interviews were open format but a list of questions was used as a guide (Appendix II). Not all questions were explicitly answered in all interviews.

Interview transcripts are given in Appendix III. A summary of the main outcomes of the interviews is given here.

Every person interviewed was aware of the importance of avoiding data loss. So a striking outcome of the interviews is the variety of methods used to store data.

Probably the most common situation is that researchers store their data in a variety of files and folders, on their own computer or on a network drive. With a certain effort, the researcher is able to dig up data when asked to do so.

A second common situation is that some researchers freely admit that they can no longer find their research data, even if those data have been used for publications, and even if the data is (in some cases) only a few years old.

A third common situation is that students are asked to provide their supervisors with a copy of their data on CD/DVD during the final months of writing their thesis. In principle this is a good method, but some supervisors recount that after the student has left, it is sometimes discovered that the DVD is empty (forgot to finalize the burning process?) or that the data on the DVD cannot be understood. In this project, several DVDs were examined and although they contained hundreds of files, these files could not be linked to the Ph.D. theses they were supposed to support.

Only a few of the interviewed researchers store their data in a planned manner. This may involve a shared network drive which can be accessed by several members of a team and which is backed up regularly. In some cases it also involves a relational database which implies that some thought has been given to the structure of the data.

The researchers interviewed were asked which would be reasons to deposit data and which would be reasons not to deposit data. Reasons to deposit data are listed as 'positive outcomes of data deposition' in Table 1; reasons not to deposit data are listed as 'obstacles to data deposition' in Table 2. The items listed in the Tables are not different from the ones mentioned in the literature.

A number of researchers were ready to deposit one or more datasets. A larger number were enthousiastic about depositing a dataset, but needed time to find the data or to transform it into a more understandable format. Interaction with these researchers will be finalized after this report has been written. Both kinds of datasets are listed in Table 3.

Table 1. Positive outcomes of data deposition mentioned in interviews (in decreasing order of importance).

No.	Outcome
1	Better science
2	Backup data; preserve data better than individuals or research groups can do
3	Receive more citations
4	Personal satisfaction from 'doing the right thing', and when data are used by others
5	Document professional output

Table 2. Obstacles to data deposition mentioned in interviews (in decreasing order of importance).

No.	Obstacle
1	Ownership, property rights
2	Effort (collecting, organizing, documenting)
3	Inappropriate use of data (inadvertent or wilful)
4	Need tools and protocols to prepare data for deposition
5	Confidentiality of data needs to be preserved (personal /commercial)
6	Intention to write papers precludes deposition now
7	Data may be of insufficient quality
8	Legal liability
9	Power shift (from data owners to data users; or more generally, from experimentalists to analysers)

Table 3. Datasets deposited or being deposited as a result of interviews

Dataset	Status
1998 dataset	Ongoing
Dataset on 100 plant species in several types of grassland, 1925-1960	Ongoing
Geneflow experiment	Ongoing
Review paper data	Ongoing
1) Grassland experiments 2) data retrieved from papers – 3 sandy soil locations 3) Ph.D. student's work on labour – papers are in review now	Ongoing
Ossekampen data	Ongoing
Data for Acta Hort paper	Ongoing
Report will be finished in January -> dataset on catch-crop, date of sowing, date of measurement, N yield, yield	Ongoing
http://library.wur.nl/WebQuery/wurpubs/444963 http://library.wur.nl/WebQuery/wurpubs/444962 http://library.wur.nl/WebQuery/wurpubs/444961 http://library.wur.nl/WebQuery/wurpubs/444960	Finished
Several Ph.D.s will finish in Feb 2014 – will need help depositing	Ongoing
http://library.wur.nl/WebQuery/wurpubs/443698	Finished
Yield trend of wheat cultivars	Ongoing

2.1 Procedure for identification and selection of datasets

The interviews were an efficient and enjoyable manner of connecting with the researchers. The interviews lasted generally around 30 minutes, although a few took almost an hour. During the interview, a brief introduction was given about the project and the purpose of the interview. In many cases, of course, the interviewee had already attended the seminar, which allowed the introduction to be brief. Even in those cases, one-on-one contact allowed for lingering questions to be answered and doubts to be allayed, thus smoothing the way for the remainder of the interview.

Interviewees in general appreciated the opportunity to talk about their work. Again, this was facilitated by the one-on-one format. It was noticeable that enthousiasm was greater when talking about recent datasets or about datasets that the researchers are passionate about. People did not enjoy the idea of having to look hard for an old dataset, no matter how valuable it would be.

Interviewees in general appreciated that they were talking to a scientist instead of a librarian. The interviews could be efficient in large part because no time was lost in having to explain domain-specific terminology.

3. Recommendations

The results of the interviews and of depositing the data sets discovered were used to formulate the following recommendations to further the main objectives of the study reported here, which are:

- Identify databases and datasets for deposition with the Library
- Develop services and facilities for domain-specific data curation

A number of recommendations relate directly to one or more of the obstacles in Table 2. No recommendations are made to address obstacles 5 (confidentiality of data) and 7 (data of insufficient quality).

Use a scientist to approach scientists

The 'embedded scientist' concept worked well in this project. The interviews held to inform scientists and identify datasets for deposition were efficient in large part because no time was lost in having to explain domain-specific terminology. It is recommended that this approach be used in future work.

Communicate that current methods of data storage don't work

The interviews provided convincing examples of data storage efforts that have utterly failed. Storing data on a DVD is a great idea, but if the DVD is empty because the burn process was not finished, or if the data on the DVD cannot be understood, then the data are lost. The interviews also provided examples of researchers who cannot find their own data any more – only a few years after the end of a project. The message that the current practice in many cases fails to preserve datasets should be communicated to students, researchers and managers alike.

Enable scientists to describe datasets (obstacles 2 and 4)

The ultimate aim of data deposition is that scientists use the datasets of other scientists. Typically, the dataset-user will not be able to communicate directly with the dataset-provider. Therefore, the dataset-provider must provide sufficient information for the dataset to be understood, and the dataset-user must be able to understand the documentation provided. Clearly, a common language is needed.

The scientific articles with which a dataset is linked will typically describe such things as the measurement methods, the quantities observed, and the units in which measurements are expressed. Entity-relationship modeling can be used to describe the internal structure of a dataset, such as the relationship between a measurement and the field plot on which it was taken (e.g. Carlis and Maguire, 2000). These descriptions are specific to a particular dataset.

In contrast to the above, Linked Open Data technologies allow the description of data using vocabularies that (a) can be shared and (b) are readily processed by machines (Allemang & Hendler, 2011; Berners-Lee *et al.*, 2001), thus allowing combinations of datasets to be made and to be analyzed. The use of Linked Open Data technology to describe datasets is highly recommended.

Scientists will need training and tools to describe their datasets using Linked Open Data technology. For example, Rijgersberg (2013) describes an Excel-plugin that allows scientists to annotate their data with a quantity and units picked from a list. This kind of tool allows the description of a dataset to be precise and to be done with a minimum of effort. It is recommended that such tools be further developed and made available to researchers. It is not readily apparent what the role of the library in this innovation can be.

Distinguish between storing, archiving and depositing datasets and adopt a mechanism appropriate to the goal

There is a need to inform scientists about three different goals with respect to handling datasets. Storing a dataset while it is being used occurs usually as part of the activities of a project. For the purposes of the project, it may be

necessary to make the dataset available via a shared folder or online repository, to back it up, to allow changes to be made and (possibly) to be remembered (versioning). Storing may involve the use of files, of a database, or both. Storing may also involve recording how the dataset is transformed into new data products (workflow management). When the project finishes, the continued existence of the dataset becomes uncertain. Archiving a dataset relates to a legal requirement to maintain documentation. Typically there is fixed duration during which an organization is required to maintain its records. At the termination of that period the archived documents are destroyed. Depositing (or publishing) a dataset has many similarities to the publishing of research findings: the data is physically preserved, it becomes immutable, it is described so it can be understood, it is given an identity, it's existence is advertised, and a mechanism is defined through which it can be obtained (White and Van Evert, 2008).

The above can be illustrated by some of the interviews. One interviewed researcher mentioned that the backup service implicitly provided by deposition is a good reason to deposit. However, this is wrong because deposition would create an identity for each (incremental) backup and preserve it indefinitely – which is not the intention of a backup. Another researcher mentioned that his group doesn't need deposition because they have already implemented good storage mechanisms. This is wrong because even if the data is stored well currently, they cannot easily be cited (no identity, not immutable) and long-term storage is not guaranteed.

It is recommended that the differences between storing, archiving and depositing be communicated to students, researchers and managers; and that they are encouraged to use appropriate tools to achieve each goal.

Develop a consistent policy for identifying and citing datasets

It is recommended that a consistent scheme be developed to cite datasets. In general such a policy should be based on the assumption that a unique identifier is used to refer to the dataset and that there is a commonly known resolver service that returns the current location of the object, i.e. the DOI scheme. DANS has plans to adopt the DOI scheme on top of the URN:NBN scheme, whereas 3TU Datacentrum represents the DOI consortium for datasets in the Netherlands.

The need for this is best illustrated by two examples. Van Oort and Timmermans (2012) is a record in WaY that points to http://library.wur.nl/WebQuery/wurpubs/443698. The bibliographic record at that location links to the physical object deposited with EASY at https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:55620. The physical object can also be reached via its persistent identifier urn:nbn:nl:ui:13-huy3-tl. Van Evert *et al.* (2011) is another record in WaY; it points to http://library.wur.nl/WebQuery/wurpubs/409283. The bibliographic record links to a composite edepot record at http://library.wur.nl/WebQuery/edepot/274717 which in turn links to the physical object (a zip-file with the data) at http://edepot.wur.nl/176769. In both examples, there are three different ways to refer to each dataset and it is not clear which one should be used for citing. In general, however, URLs (such as *.knaw.nl or *.wur.nl) will become invalid as soon as support for the corresponding domain is withdrawn and should not be used to identify datasets.

Implement licensing mechanism and educate about property rights (obstacles 1, 3, 8 and 9)

Property rights is the obstacle mentioned most frequently. This is a complex issue that cannot be remedied with a simple recommendation. Nevertheless, some recommendations can be made. It is recommended that a license agreement must be accepted before edepot data can be retrieved. This will avoid giving the impression that, for example, the data can be used without citation.
It is also recommended that a comprehensive overview be developed which clearly identifies the various stakeholders, their roles, and the issues, including power shifts between actors that occur as a result of collecting and depositing data (Whyte and Thompson, 2010). The overview could for example be presented in the form of several case studies.

Embargo policy tailored and clear (obstacle 6)

Depositing data under embargo is an important mechanism to encourage deposition, yet current embargo options seem limited. It is recommended that the need for more diverse embargo options be inventoried and guidelines for embargo use be developed.

4. Literature

Allemang, D. & J. Hendler, 2011.
Semantic Web for the working ontologist : effective modeling in RDFS and OWL Elsevier, Amsterdam [etc.].

Berners-Lee, T., J. Hendler & O. Lassila, 2001.
The semantic web. Available online at http://www.scientificamerican.com/article.cfm?id=the-semantic-web. Scientific American.

Carlis, J.V. & J.D. Maguire, 2000.
Mastering data modeling: a user-driven approach. Addison-Wesley, Boston.

Rijgersberg, H., 2013.
Semantic support for quantitative research Amsterdam: Vrije Universiteit.

Swan, A., & S. Brown, 2008.
The skills, role and career structure of data scientists and curators: an assessment of current practice and future needs.

Van Evert, F.K., D.A. van der Schans, W.C.A. van Geel, J.J. Slabbekoorn, R. Booij, J.N. Jukema, E.J.J. Meurs & D. Uenk, 2011.
Dataset - Droevendaal, Rolde and Colijnsplaat, 1996-2003. Available online at http://library.wur.nl/WebQuery/wurpubs/409283, Wageningen UR, Wageningen.

Van Oort, P.A.J. & B.G.H. Timmermans, 2012.
Key weather extremes affecting potato production in The Netherlands. Available online at http://library.wur.nl/WebQuery/wurpubs/443698, Wageningen UR, Wageningen.

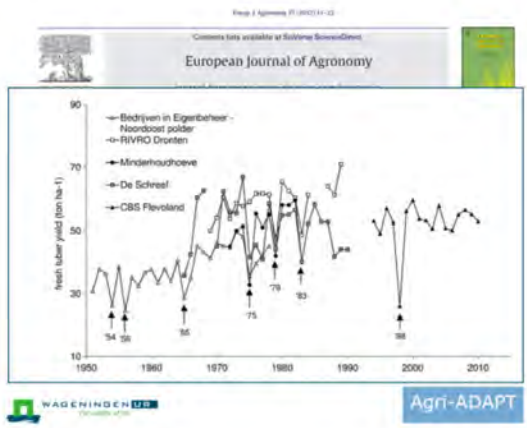
White, J.W. & F.K. van Evert, 2008.
Publishing agronomic data. Agronomy Journal 100:1396–1400.

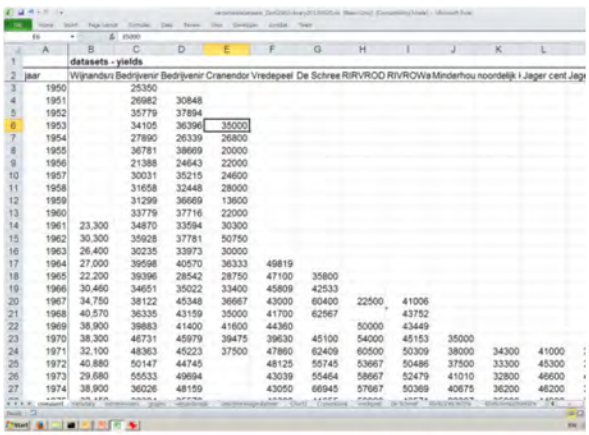
Whyte, K. & P. Thompson, 2010.
Ethical Considerations of Data Sharing: A Case Study From Animal Production. Available online at https://dl.sciencesocieties.org/publications/meetings/2010am/7791/61860.

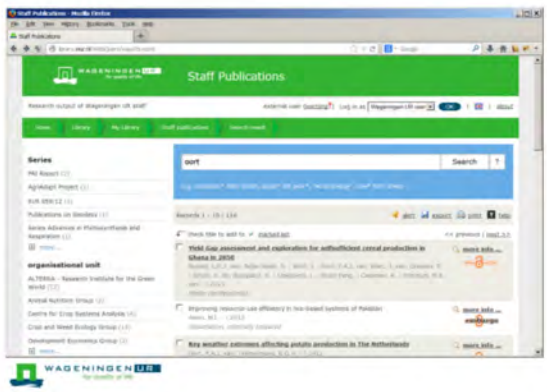
Appendix I.

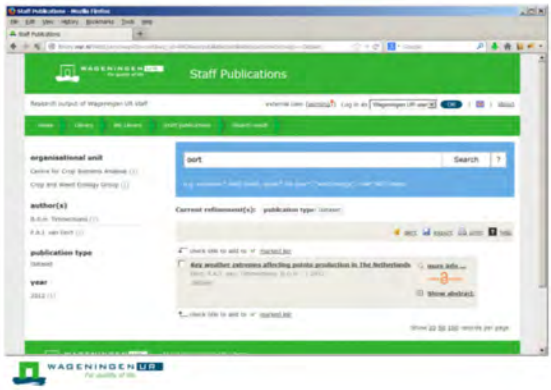
Presentation on data deposition

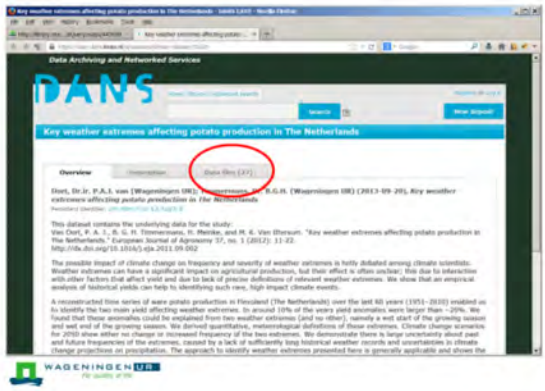
The presentation below was given on four occasions, namely at CSA on 31 October 2103, at AGRO on 13 November 2013, at PPS on 11 December 2013, and at FSE on 28 January 2014.

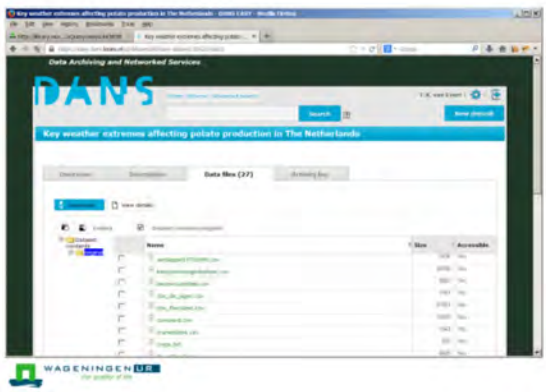












Good reasons to deposit data

- Science
  - Reproducible
  - New analysis, meta-analysis
  - Access to data after scientist becomes unavailable
- Advertise your work
- Make output visible
- Papers with data generate more citations
- Legal

Appendix II.
Interview questions

Questions used for the semi-structured interviews

1. What is your position and role in the organisation?
2. How do you store research data? Why in this way?
3. Do you ever search for research data (within WUR and/or outside WUR)? How do you search?
4. Which datasets are there that you know of? How many, in what form, what format?
5. What is a dataset (what is the delineation)?
6. Which datasets are you willing to deposit (share); why do you want to deposit these datasets?
7. Which datasets are you NOT willing to share; why not?
8. What can be offered to make people want to share?
9. If a dataset is going to be deposited, do you think raw data should be deposited; elaborated data should be shared; or both raw and elaborated data should be deposited?
10. What are your thought on depositing algorithms, computer code, computer models?

Appendix III.
Transcripts of interviews[8]

PER 1

PER 1 is a researcher at PPS. PER 1 is not involved in experiments, but does conduct surveys. PER 1 raises ethical questions on sharing surveys because subjects typically do not like it if they can be identified. Sharing data anonymously, however, may not in all cases provide sufficient anonymity and at the same time it makes it easy to hide fraud. PER 1 does see the value of sharing survey data and gives the example of a new Ph.D. student who is starting work in Flevoland – reading old surveys would enable the new student to familiarize himself with the area.

PER 2

PER 2 is assistant at CSA and is often in contact with Ph.D. students. PER 2 also was one of the organizers of a short CSA seminar on data archiving, and co-author of a discussion text on the same topic.
When his predecessor retired several years ago, PER 2 took charge of cardboard boxes full of CDs and DVDs, each of which is supposed to contain the data and models of a M.Sc. or Ph.D. student. PER 2 has continued Gon's practice of asking departing students for a DVD with a copy of the contents of their computers.
When after the interview some of the more recent DVDs were examined, it turned out that they were completely incomprehensible. The data of the Ph.D. student whose name was on the label could not be identified. Bits and pieces of the data from other students were found, but could not be understood.

PER 3

PER 3 is an assistant at PPS. He plays a central role in storing the Department's data and models, administrates the various websites of the Department, develops custom ICT- and GIS-based solutions for teaching and research.
Data and models are stored in the following ways:

1. Data that are meant to be shared with a restricted group of researchers within PPS are stored on a shared server disk.
2. The N2Africa project has many partners in Africa. Survey results from the N2Africa project are collected using an Excel template and then stored in a central MySQL database. Surveys are conducted in many African countries but all data is stored in Wageningen. From there it is distributed for analysis to the relevant project partners. Short instructional movies are made by professionals and stored at vimeo.com. The N2Africa project also has a Facebook page and a web site. Work is under way to create a metadata database on the web site; this will enable visitors to search for datasets and then write an email to request access.
3. Data from Ph.D. students are stored on CD/DVD and on hard disks. Students typically have many Excel files and many different versions of the same file. There is a protocol in use to describe metadata in Excel files.
4. Summarized data from the N2Africa project is stored at aware.com.
5. Data from the YieldGap and Seamless projects is stored on servers at Alterra.
6. Ph.D. students are encouraged to store their data at agtrials.org.
7. models.pps.wur.nl is a website for sharing models. Models can be freely downloaded after the user has agreed to a GPL-type license. Versioning of models is not implemented. Permanent availability of a model is not guaranteed.

PER 3 feels that storage of data and models at PPS is well-organized. Consequently, he sees no incentive to consider deposition with the Library.

PER 4

PER 4 is a researcher at CSA. At CSA there is no protocol or shared method of storing data.
PER 4 has recently deposited a dataset (data, models, input files) with DANS to support a recent publication.

[8] Please note that male nouns and pronouns have been used regardless of the gender of the person interviewed.

Searching data is not a common activity for PER 4, but one of his Ph.D. students is currently conducting a meta-analysis and scoures papers for data. The meta-analysis is made more difficult because the data are not reported in a uniform manner.
PER 4 is in favour of sharing datasets that were collected using public funding. He has several Ph.D. students who will finish in the next few months and therefore represent good cases to test data deposition. PER 4 does express the need for guidance in how to organize data for deposition.

PER 5

PER 5 is a post-doc at CSA and also works at Africa Rice.
PER 5 stores data on his own computer; the models eventually end up at models.pps.wur.nl. In addition, he explores the use of agtrials.org and expects it will be useful.
PER 5 uses data from scientific papers, either by copying tables and figures, or by contacting the author and asking for data files. Google is not useful.
A Ph.D. student may not finish his work, but the data collected may be useful; deposition may make it easier to find such data.
PER 5 would be willing to share data that could be used by others. PER 5 has recently published a review paper on time series of potato yield and is willing to deposit these data with the Library.
Disincentives to share data are the time and effort involved, possibly copyright/ownership issues, and lack of confidence in the quality of the data.
An incentive would be making your name more known. Data and models disappear because of lack of motivation to archive, not because of a lack of technical facilities. Professors should motivate their students to archive data.
When data are archived, metadata (incl. about persons involved in data collection) are very important. PER 5 sees no use in depositing data that are not described in a scientific paper.
Sharing model code is not sufficient, because additional documentation is usually necessary to understand the code.
There is no versioning of source code as CSA and PER 5 doesn't consider this a task for the library, but he would like to acquire the skills to implement his own versioning.

PER 6

PER 6 is a tenure track researcher and teacher at CSA. At CSA, there is no centralized mechanism to store data, this is handled by each Ph.D. student individually.
PER 6 does sometimes search data. The day before the interview, he copied some data from a pdf. Some other data were found in Ph.D. theses. This usually does not yield a complete dataset, rather just some indicative numbers.
PER 6 has 7 datasets that are good candidates for deposition. Half of those would have to be deposited with restricted access, because papers still need to be written.
A dataset is delineated as one experiment, e.g. field experiment, greenhouse experiment.
A dataset could probably also be the output of a stochastic simulation.
PER 6 considers that once papers have been published, a dataset could be shared without restrictions. The expectation of citations to the dataset is a strong incentive to share. Another incentive is new contacts.
If the shared data have been derived from raw data in some way or other, then the transformation must be clearly documented.
Scripts, models and model components are sometimes shared, but only after they have been published, under GPL.

PER 7

PER 7 is a researcher and team leader at PRI. He mentions three recent datasets. The first dataset is the result of a field experiment about geneflow. This data are kept by a colleague. The data was provided to a EU-funded project. Given that the data are being distributed, it would make sense to deposit them with the library. Before this can be done, permission will have to be obtained from the researchers involved and from the Netherlands Dept of Economic Affairs who funded the research. A second dataset resulted from confidential commercial research and cannot be shared. A third dataset is used to derive calculation rules for herbicide dosing which are licensed to commercial partners; this precludes sharing the dataset.
PER 7 is of the opinion that data resulting from publicly funded research should be shared but remarks that there may be political meaning in the moment of sharing. In high-profile dossiers (e.g. GMO, climate change) the policy

makers who commissioned the study may have specific wishes about the time and the manner in which data are shared: some think that their policies are best supported by early sharing, while others prefer more control.

PER 8

PER 8 is a professor at CSA and is no longer involved in primary data collection. The Ph.D. students counseled by PER 8 are asked to hand over a CD/DVD containing all research data upon completion of their Ph.D. –usually it takes quite a bit of pressure to make sure that this actually happens during the last hectic months of the student's tenure. Ideally, it should become automatic for students to store their data.
PER 8 regularly works with datasets from other researchers. In some cases it is hard to understand how the data are coded. In a recent case, the other researcher was willing to share the data, but they had recently been thrown away. PER 8 typically works with large datasets for QTL-based research, but in many cases the data are not stored in such a way that they can be easily understood and manipulated. For example, in 2005 an important article on barley appeared. Later, an EU project yielded a dataset that could have been matched with the barley data, but it turned out that the EU dataset could not be understood.
PER 8 is proponent of a data management plan, specific to a field of study. Such a plan could define a preferred method to store data which would ensure that the data can be understood. Ideally there would also be standards for measurements, such as always reporting dry weights after drying at 70 C.
PER 8 is quite willing to share data after research papers have been published. Most of the data were collected using public funds. A mild restriction applies to data from research funded by STW[9]: for half a year after the end of the study, the preferred user has exclusive rights to the results.
PER 8 names liability as a disincentive to share data. There is a need to protect the party sharing data from legal action that might result if a third party incurs damages as a result of errors in your data.
Incentives to share data include reciprocity (researchers using data from other researchers will probably be willing to share their own data) and credit in the form of citations to a published dataset.
Models and algorithms should be treated similarly as data. PER 8 contributes to models.pps.wur.nl.

PER 9

PER 9 is an independent consultant who works at PPS and is involved in the N2Africa project.
PER 9 has in the past stored some data with agtrials.org. Starting in 2010, some PPS data are also stored there.
In the N2africa, among other things, maps are used to create a typology of environments. It has not been decided where this very useful dataset will be stored.
IP is an important issue to consider when sharing data. This is especially important for PER 9 because he works as an independent consultant. There are no fixed rules that define how to deal with IP, so every time a contract is written, this needs to be considered anew. It is clear, however, that it will be important to keep track of the provenance of data and the accompanying IP rights. As an aside: agtrials.org allows the specification of a separate license for each dataset; most people use the Creative Commons license.
PER 9 remarks that scientists are not the prime movers regarding either deposition of data or regulating IP. For this to happen, the administators have to be become involved.
Incentives for data sharing are increasing your visibility and documenting output (performance evaluation).
PER 9 published a paper based on a model at a time when the journal did not have a facility to attach model code to the paper. In this case, the code was published on a website – but inevitably the website disappeared so now the code is no longer accessible. It is clear that deposition with the Library would have been much better in this case!

PER 10

PER 10 is a researcher at PPS. The Department makes datasets and models available at models.pps.wur.nl. PER 10 was responsible for creating the license that users of that website must accept before downloading data and/or models.
All students of PPS are asked to prepare a CD/DVD with all data and models and hand this to PER 10. The students are asked to 'organize their data well', but there is really no definition of what this means or how to achieve this. And despite repeated requests, some students manage not to make a DVD, many DVDs cannot be understood, and

[9] Technologiestichting STW, www.stw.nl

sometimes the DVD turns out to be empty (did the student forget to press 'finalize' in the burn process?). A student's advisor can make a difference by emphasizing the importance of archiving the data.
There is an effort under way to store the data of Ph.D.s during the last 10 years retroactively at www.agtrials.org – even though many DVDs don't contain all that much.
PER 10 stresses the importance of some form of review for data and models. Models could for example be tested for time step convergence. Science's 'Policy Forum' has a topic 'Computational Science' where some discussion on peer-review of models may be found (e.g. DOI:10.1126/science.341.6143.237-a).

PER 11

PER 11 is a researcher and counselor of Ph.D. students at PPS. PER 11 requires students to share their data with him, but often the files are difficult to interpret, e.g. because units are not noted, because values are copied from one Excel-cell to another (instead of using formulas), and in general because some students are pretty chaotic workers. Storing raw data is important. PER 11 gives the example of a survey on labour requirements in Africa. 1 day of weeding is 10 hours, 1 day of land preparation is 5 hours; this may lead to confusion if some respondents quantify their workload in days and others in hours.
PER 11 is willing to share pretty much all data, because of publicly funded research. There may be restrictions placed on sharing when foreign universities of grantors are involved. Sharing only after papers have been published. Models fall under the same rules as data.

PER 12

PER 12 is a researcher at FSE. FSE has 3-4 Ph.D. students and 30-40 M.Sc. students per year. The data collected by these students is not stored in a systematic manner. In a large international project, DropBox is used to store data. This latter project is financed by a U.S. agency which demands that data are made available, but FSE is still working out how to do this.
At FSE a variety of models is used: balance models, dynamic models, landscape models. The inputs for these models is likewise varied (could include georeferenced data) and it might be good to store these in a systematic way.
PER 12 considers that probably all datasets generated at FSE could be shared. When data are shared, they must be well-documented.
Compass is a model toolkit consisting of 10 models and it is made available via a website.

PER 13

PER 13 is researcher at CSA. Main task is to advise students; also teaches a course in research methodology. Students organize storage of their own research data, typically on their own computer and using a folder/file structure. PER 13 encourages students to make paper printouts of the raw data. The type of files stored has changed in the last 20 years.
PER 13 sometimes attempts to find data that are mentioned in an article.
PER 13 is willing to deposit data with the library. A big advantage is that the data are safe-guarded from getting lost once they have been deposited. Deposition also gives the satisfying feeling that a job is completed. PER 13 would like to get credit for deposition, e.g. in the form of being cited.
PER 13 sees the time it takes to document the data as an obstacle. Documentation is needed for protocols in the experiment, about things that went wrong during the experiment, and to explain missing values. Other obstacles are that ownership of the data may not be clear and that all authors must agree with the deposition – contacting long-departed authors may be difficult. If deposited data are used by others, PER 13 would like contact with the user(s), to minimize the chance that the data are used in an inappropriate manner.
PER 13 thinks it is important to store raw data (plot level) because it is not uncommon to find errors that were made when converting, for example, from kg/plot to t/ha.

PER 14

PER 14 is researcher at PRI. Data storage is very ad-hoc. Many binders with old data (1970s, 1980s) have been thrown away during one of the several moves during the last decades. Slightly more recent data are stored on floppy disks that either have been thrown away or cannot be located anymore. Current data are stored in a

folders/file structure on several computers. PER 14 recounts that once the binders with data from a Ph.D. student were thrown away just two months before those data were requested.
PER 14 often uses internet search engines such as Google and indexing services such as Web of Science to find information. This usually doesn't yield datasets, but sometimes a M.Sc. thesis gives data in an appendix.
PER 14 is comfortable with the idea that others use his data, but would like some credit. This can be co-authorship or acknowledgement of the use of the data, depending on how large a contribution the data makes to the new paper. An incentive to share data comes from journals. PER 14 recently had a paper in New Phytologist (high impact; rejection rate 75%) but 'unpublished data' was not allowed, so PER 14 placed a mid-term report on a project website and cited that. An undesirable outcome was that now the mid-term report and the data in the appendix can be found with Google which is a little bit more visibility than the researchers want at this time. PER 14 thinks that deposition with the library (under a limited-time embargo) would have been better in this case.
PER 14 is willing to share data and sees it as a way of advertising your work. An additional incentives to share would derive from a facility which allows incremental deposition: whenever a new batch of data comes in, that batch is deposited and thereby safe-guarded from loss, from inadvertent changes, and from malicious tampering.
Disincentives to share data are restrictions imposed by clients, especially when working with commercial partners. Before plant material is shared with a commercial partner, both parties sign a Material Transfer Agreement. This illustrates the stringent restrictions thay also may apply to the use of data.

PER 15

PER 15 is a researcher at PPS. Currently works mostly with models. Data from earlier experimental work is stored ad-hoc in folders/files and can for the most part still be understood. PER 15 has much data on grassland that were originally stored on VAX machines.
For the YieldGap project, PER 15 needs data on crop growth for many countries. This problem is solved by seeking a local scientist in each country and having that person search for datasets.
PER 15 would be willing to share any dataset that is of scientific or general interest and would derive satisfaction from the fact that a dataset is reused by someone else. The effort to document a dataset before it can be shared is a reason not to share. Also, some data in a dataset may have to be removed before it can be shared, e.g. weather data obtained from KNMI[10] cannot be distributed further. It would help if the Library offers good support for depositing data. PER 15 is afraid that sharing data would lead to a deluge of questions about the data and that there would be no time to answer these questions.
PER 15 is of the opinion that a shared dataset should contain the raw data, e.g. kg harvested per plot and size of plot, to ensure that the resulting yield in kg/ha can be checked.
With regard to sharing models, PER 15's paper about a dairy model documented all equations in the model by linking to a pdf-file in the e-depot of the WUR Library.

PER 16

PER 16 is a team of researchers at PRI. They manage the data collected over many years from an experimental dairy farm and from commercial farms. The data are distributed over a number of computers: soil, crop and soil organic matter data are located at AGRO, animal and milk production data are located at ASG. Data are stored in a variety of formats, including some on an Oracle server. The members of the team work very closely together and they always know whom to approach in order to locate a particular piece of information. Processing the data has led to a situation in which a particular piece of information may appear in many files.
The team uses weather data and CBS[11] statistics data but rarely looks for data elsewhere.
The team recognizes that in an ideal world all data would be shared, but sees serious practical obstacles for sharing their own data. First they still plan to write papers, this may include papers that describe 20+ years timeseries data and thus precludes even sharing data from many years ago. Second, data from the commercial farms must remain anonymous at all times. Third, the team is worried about wilful or inadvertent abuse of the data should they be used by others. Fourth, funding by industry means that there is a lot of pressure to produce and this leaves little time to document the data for sharing or to assist others in understanding the data (after sharing).

[10] Koninklijk Nederlands Meteorologisch Instituut = Royal Netherlands Meteorological Institute, http://knmi.nl/
[11] Centraal Bureau voor de Statistiek = Statistics Netherlands, http://www.cbs.nl

The team also worries about researchers who refuse to engage in costly and laborious data collection, and use data collected by others instead. The prominence that is enjoyed now by 'data collectors' could shift to 'analysts' if data are shared on a large scale.

PER 17

PER 17 is a researcher at PRI and uses folders/files to store research data. PER 17 feels on top of his data, but admits that the African villages from which some data are collected appear with different spellings of their names and this makes it sometimes difficult to process the data.
PER 17 uses Google and Web of Science to search data and sometimes finds useful tables in report and articles.
Survey data must remain anonymous; this makes sharing less desirable.
PER 17 has published reports in which the source code of the model or algorithms is listed completely. In another the case, the model described in a paper was implemented in Excel and it was made available at models.pps.wur.nl.

PER 18

PER 18 is a researcher at AGRO and works mostly with models, is not involved in primary data collection.
PER 18 has searched data to support the modeling work. Some data on maize growth were found at agtrials.org but this was not extensive.
PER 18 sees a model and its inputs as a unit. The executable version of a model as used in a project is always saved, but it is not always clear which source code was used to compile the executable.
The source of a model is currently not shared: hardly anybody asks for it; PER 18 sees no benefit for himself in sharing; and source code is typically poorly documented and therefore couldn't be understood if it was shared.
PER 18 sees several advantages to depositing data with the Library: this will strengthen a paper because the supporting data can readily be found by readers; it may lead to further interactions with the readers of the paper and the users of the data; and the data is well-preserved and protected from accidental loss.
When data are deposited, is there a mechanism to guarantee against corruption of the data?
PER 18 mentions the example of study where some of the model output is presented in a paper, but where the bulk of the model output is not reported. It would certainly be useful to make the complete output available, but this is probably to bulky to be acceptable as supplemental information for the journal. Would it be appropriate to deposit this output with the library?

PER 19

PER 19 is a researcher at PPS. He is involved mostly with modeling work and not with primary data collection. For the Yield Gap project, data are collected from a variety of sources, including weather data from NOAA and country meteorological services, and soils data from ISRIC. Scientific literature can be searched using Web of Science to find reasonable values for LAI. Detailed crop growth data with which models can be calibrated are located through country agronomists (see also PER 15).
Data collected by PER 19 belong to others and can therefore not be shared.
Models used are published via models.pps.wur.nl but it is not clear how version management is handled: currently a paper cites a *model* at models.pps.wur.nl, but one would really want to cite a *version of a model*.
PER 19 makes the case that model output should be shared along with the model and the input data. In principle, model+input is sufficient to recreate the output; in practice, it is not always clear which version of a model was used (see above), which runtime environment was used (e.g. Windows / Linux), and which (version of which) compiler.

PER 20

PER 20 is a researcher and team leader at AGRO. PER 20 manages the long-term grassland experiments at 'De Ossekampen', in which changes in the composition of grassland are being followed since 1957. Most of the data have been digitized and can be readily understood by the team of researchers and assistants in charge of the experiment.
During the last 20 years or so there have been at least 10 occasions on which a subset of the De Ossekampen data has been made available to outside researchers, mostly Ph.D. students. In some cases PER 20 or co-workers have been listed as co-author of the resulting papers, but not always. As a result, it is now fairly difficult to fully document the value of the experiment. PER 20 has no trouble seeing that if the data had been deposited with the library, then

all papers could have cited that dataset, and then the output from the experiment would have been well-documented. As a result of this interview, PER 20 will determine whether the effort of deposition will be worth the benefits.
PER 20 is supportive of sharing data resulting from publicly-funded research, but even so would prefer to have contact with potential users before the data are transferred. He is also involved in commercially funded research and those data cannot be shared.
PER 20 provides an anecdote that underlines the importance of depositing data: recently, he received a request for data from his own Ph.D. thesis and he was unable to find this data.

PER 21

PER 21 is a researcher at AGRO. PER 21 stores data in folders/files. He uses on purpose and consistently the same codes and units, e.g. DMY means dry matter yield in kg/ha, PY means phosphorus yield in kg P_2O_5 per ha. PER 21 mentions that ADAS[12] has a SOP (Standard Operating Procedure) for formatting data files.
PER 21 recently wrote a review paper summarizing the results of14 experiments. Five experiments were conducted by PER 21, the remaining 9 experiments were found by contacting his professional network.
PER 21 would be happy to share his data, but only after contact with the potential users.
Sharing models is a litte bit more problematic because in many cases DLO depends financially on producing model results for clients. From a scientific point of view it would be good to share the code for these models, but reality dictates otherwise.

PER 22

PER 22 is a researcher at AGRO. PER 22 finds that at the office there is not enough disk space to store all data. He therefore stores some data and makes backups on his home computer. As a result, data are spread over home and the office, not all data is easily found. Data from his Ph.D. thesis from less than 10 years ago are probably lost.
PER 22 uses Google to find datasets, but even if data are found, they cannot always be understood. Journal articles sometimes have useful data as 'supplemental information'. As an example, data about nutrient content of trees could be found when PER 22 searched for it.
PER 22 is willing to share data once one or more papers have been published about them. However, not all data has a level of quality that makes one comfortable sharing it – even if it has been published.
PER 22 is involved in a project where hundreds of answers to a series of questionnaires are collected. The advantage of a data deposition system is that the questionnaires, which run to many pages of text and of which there are many versions, could be deposited along with the data. This would help with the proper interpretation of the answers by future users of the dataset.
PER 22 thinks that one of the biggest advantages of data deposition is that it is an implicit backup system.
PER 22 notices that models are often used for commercial purposes, thus sharing the model code is not an option. On the other hand, there are also cases where the model code has been shared but where the commercial value is in the specific parameterization of the model.

PER 23

The interview with PER 23 quickly focused on datasets that he may be willing to deposit.

PER 24

The interview with PER 24 quickly focused on datasets that he may be willing to deposit.

PER 25

PER 25 will consider to deposit a dataset that supports a report to be published in January 2014. He needs some guidance on how to format the dataset so that it will be understandable.

PER 26

PER 26 is a post-doc at PPS. He is not involved in experiments but works with models.
PER 26 makes it a point to store model input files along with the models so that model output can be reproduced. It is often not possible to deposit the input data because the input data was provided by third parties. Typical sources

[12] http://www.adas.co.uk

of input data are FADN and Seamless. In addition, expert knowledge, surveys and interviews, and growers' manuals are used.
Models are typically stored at models.pps.wur.nl. Typically, the first version of a model is posted, but updates are not – models.pps.wur.nl does not provide versioning.

PER 27/PER 28

The interview with PER 27 and PER 28 did not follow the questionnaire but resulted from a contact initiated by PER 27 who is a researcher and coordinator of data for WU Animal Breeding and Genetics. PER 28 is a programmer at PRI-Bioscience.
PER 27 and PER 28 have students who write code (mosty R scripts). The logistics of reproducing a student's work, including determining exactly which version of the code was used to produce a certain result, has proven to be very hard. PER 27 and PER 28 propose using a Git repository. They would then base their discussions with the students on the code as found in the repository, thus ensuring that the code is archived as a part of day-to-day activities. Setting up and administering a Git repository is a burden they don't want to shoulder and they propose that the Library takes on this task.
PER 27 and PER 28 feel no direct need to store data (in the same Git repository, or in a separate database). This is only partly because the data used is already stored in large bioinformatics repositories.
There is some discussion about the difference between (1) workflow management and (2) deposition.
A versioning system such as Git is typically used to manage the process of writing software, including multi-author collaboration and tracking of changes. A versioning system is typically not guaranteed to persist in the same manner as documents and data stored with a library are expected to persist.
Deposition is meant to preserve and advertise a dataset or a model. Deposition is most useful when it is linked to a scientific paper.
There is some overlap between the two kinds of systems. A particular revision of the software in a versioning system could be cited in a scientific paper. Similarly, subsequent versions of data/model can be deposited, with links maintained between versions and with the relevant papers.
PER 28 argues that even if a paper were to be accompanied by a deposited model, he would still rather track down the versioning system used for that software and work with the latest version of that software – which presumable by that time has bugs fixed and features added.

PER 29

The interview with PER 29 quickly focused on a dataset on yield trends in wheat cultivars.

PER 30

PER 30 only recently joined the faculty of PPS and as such has no datasets that were collected in his present position and could be deposited. PER 30 mentions that priority should be with paper data that are about to get lost.

PER 31

PER 31 is a researcher at AGRO. A while ago, PER 31 received a request for his dataset on weed counts from a mathematician at another university. He deposited the dataset with the Library and then sent the link to the mathematician. Recently, he found a presentation by this mathematician that used the data but didn't mention the source. PER 31 sent an email, the mathematician apologized, and added a reference. PER 31 made the (tongue in cheek?) suggestion to provide along with the data, a photo of the research team, on their knees, in the field, dirty and tired, to illustrate the hard work that is often involved in collecting a dataset.